Collection
Statistique
et probabilités
appliquées

Christian P. Robert

Le choix bayésien

Principes et pratique

$$\hat{\sigma}_{\bar{X}_1 - \bar{X}_2}$$

$$S_1^2 = \frac{\sum_{i=1}^{n_1}(X_i - \bar{X}_1)^2}{n_1 - 1}$$

$$S_1 = \sqrt{S_1^2}$$

$$\bar{X}_1 = \frac{\sum_{i=1}^{n_1} X_i}{n_1}$$

$$S_1 = \sqrt{S_1^2} \quad \text{et} \quad S_2 = \sqrt{S}$$

$$\bar{X}_1 = \frac{\sum_{i=1}^{n_1} X_i}{n_1} \quad \text{et} \quad \bar{X}_2 = \frac{\sum_{j=1}^{n_2}}{}$$

Springer

Le choix bayésien

Principes et pratique

Springer

Paris
Berlin
Heidelberg
New York
Hong Kong
London
Milan
Tokyo

Christian P. Robert

Le choix bayésien
Principes et pratique

Springer

Christian P. Robert
CEREMADE, Université Paris Dauphine et
CREST, INSEE, Paris

ISBN-10 : 2-287-25173-1 Springer Paris Berlin Heidelberg New York
ISBN-13 : 978-2-287-25173-3 Springer Paris Berlin Heidelberg New York

© Springer-Verlag France, Paris, 2006
Imprimé en France

Springer-Verlag France est membre du groupe Springer Science + Business Media

Cet ouvrage est soumis au copyright. Tous droits réservés, notamment la reproduction et la représentation la traduction, la réimpression, l'exposé, la reproduction des illustrations et des tableaux, la transmission par voie d'enregistrement sonore ou visuel, la reproduction par microfilm ou tout autre moyen ainsi que la conservation des banques de données. La loi française sur le copyright du 9 septembre 1965 dans la version en vigueur n'autorise une reproduction intégrale ou partielle que dans certains cas, et en principe moyennant le paiement de droits. Toute représentation, reproduction, contrefaçon ou conservation dans une banque de données par quelque procédé que ce soit est sanctionnée par la loi pénale sur le copyright.
L'utilisation dans cet ouvrage de désignations, dénominations commerciales, marques de fabrique, etc. même sans spécification ne signifie pas que ces termes soient libres de la législation sur les marques de fabrique et la protection des marques et qu'ils puissent être utilisés par chacun.
La maison d'édition décline toute responsabilité quant à l'exactitude des indications de dosage et des modes d'emploi. Dans chaque cas, il incombe à l'usager de vérifier les informations données par comparaison à la littérature existante.

SPIN : 11402 848

Maquette de couverture : Jean-François Montmarché
Dessin de couverture : détail d'un tableau de Michel Marin

À mon a priori de référence, Brigitte,
et à mes deux updates les plus importants,
Joachim et Rachel.

Collection
Statistiques et probabilités appliquées
dirigée par Yadolah Dodge

Professeur Honoraire
Université de Neuchâtel
2002 Neuchâtel
Suisse

Comité éditorial :

Christian Genest
Département de Mathématiques
et de statistique
Université de Laval
Québec GIK 7P4
Canada

Marc Hallin
Université libre de Bruxelles
Campus de la Plaine CP 210
1050 Bruxelles
Belgique

Ludovic Lebart
École Nationale Supérieure
des Télécommunications
46, rue Barrault
75634 Paris Cedex 13
France

Stephan Morgenthaler
École Polytechnique Fédérale
de Lausanne
Département des Mathématiques
1015 Lausanne
Suisse

Gilbert Saporta
Conservatoire national
des arts et métiers
292, rue Saint-Martin
75141 Paris Cedex 3
France

Dans la même collection :

– *Statistique. La théorie et ses applications*
Michel Lejeune, avril 2004

Préface

"The first lesson is what questions to ask."
Robert Jordan, *Knife of Dreams*.

Quinze ans plus tard...

La toute première version de ce livre a été publiée en 1992 chez Economica, sous le titre *L'Analyse Statistique Bayésienne*, comme premier titre d'une collection de Statistique dirigée par Paul Deheuvels. Le livre a ensuite été remanié à deux reprises pour donner les éditions de *The Bayesian Choice*, publiées chez Springer-Verlag (New York) en 1994 et 2001. Les changements par rapport à la première version française sont trop nombreux pour être décrits ici, d'autant que cette édition initiale est épuisée et n'est donc plus disponible qu'en bibliothèque. Si je me suis décidé à compléter le cercle et à entreprendre la retraduction du *Bayesian Choice*, c'est, d'une part, parce que la première édition française n'est plus disponible alors qu'il est toujours un peu délicat de suggérer un livre de référence rédigé en anglais en troisième (L3) et en quatrième (M1) années d'un cursus francophone... D'autre part, *The Bayesian Choice* ayant été distingué par la Société Internationale de Statistique Bayésienne (ISBA) en 2004 en obtenant le prix De Groot, il me semblait qu'une version en français pouvait présenter un intérêt pour les bayésien(ne)s francophones. Comme j'avais égaré le fichier TEX de la version française de 1992 (!) et que les modifications apportées dans les versions anglaises me semblaient globalement positives, je me suis fondé sur la seconde édition anglaise. (J'ai d'ailleurs choisi de garder les citations tirées de *The Wheel of Time* de

Robert Jordan, plutôt que de chercher de nouveau des citations en français ou, pire, de les traduire litéralement...)

Programmes de Cours

Sans très grande originalité, je suggère que, dans un premier cours d'analyse bayésienne (par exemple, en L3 ou en M1), les chapitres de base (Chapitres 1 à 6) devraient être traités presque entièrement à l'exception des Notes et des Sections 4.5 et 5.4, alors qu'un cours centré plutôt sur la Théorie de la Décision peut omettre quelques parties des Chapitres 1 à 3, et les Chapitres 4 et 6 entièrement, pour couvrir à la place les Chapitres 7 à 9.

Pour un programme d'études plus avancé concernant des étudiant(e)s déjà familiarisé(e)s avec la Statistique bayésienne (en M1 ou en M2), ma suggestion est de traiter d'abord l'impropriété abordée dans la Section 1.5, les lois a priori non informatives de la Section 3.5, les modèles dynamiques de la Section 4.5 et des Notes 4.7.3 et 4.7.4. Je passerais aussi du temps sur les tests abordés dans le Chapitre 5 (excepté éventuellement les Sections 5.3 et 5.4). Puis, après une présentation approfondie des méthodes de simulation à travers le Chapitre 6, je passerais au sujet plus controversé du choix de modèle dans le Chapitre 7, aux résultats récents d'admissibilité de la Section 8.2.5 et la Note 8.7.1, et à la modélisation hiérarchique et empirique du Chapitre 10.

Une alternative pour un cours de cinquième année (M2) d'un semestre est de couvrir ce livre et celui de *Méthodes de Monte Carlo par Chaînes de Markov* en simultané : on pourrait ainsi traiter les Chapitres 1 à 3 du présent ouvrage, disposant ainsi d'un matériel d'illustration suffisant pour l'introduction des méthodes de Monte Carlo et de Monte Carlo par chaînes de Markov. On peut ensuite revenir aux Chapitres 4, 5 et 7 du présent ouvrage, en éliminant bien entendu le Chapitre 6. La disponibilité des outils MCMC[1] permet ainsi de traiter des modèles beaucoup plus ambitieux et on peut s'appuyer en parallèle sur la dernière édition de *Monte Carlo Statistical Methods* pour les techniques les plus récentes dans ce domaine. (Il est en effet très vraisemblable que je n'entreprendrai pas une (re-)traduction de cet ouvrage en français !)

Remerciements

J'ai traduit ci-après la préface de l'édition de 2001 du *Bayesian Choice*, principalement à cause de sa section de remerciements, que je réitère ici. (Je n'ai pas voulu reprendre l'ensemble des trois préfaces pour ne pas surcharger l'introduction et surtout pour éviter les répétitions !) Je dois quand même rajouter quelques nouvelles "têtes" (et dettes) à ma liste de créditeurs. En particulier, travailler avec Jean-Michel Marin depuis son arrivée à l'Université Paris Dauphine m'a beaucoup apporté et, même si cette traduction n'a pas

[1] MCMC signifie *Markov chain Monte Carlo* ; il s'agit d'une méthode de simulation (re)découverte aux débuts des années 1990 par la communauté bayésienne.

intégré nos derniers travaux communs sur le choix de modèles et la sélection cohérente de lois a priori, cette perspective se retrouve dans *The Bayesian Core*, ouvrage que nous avons rédigé en commun à l'intention d'un public plus pragmatique (à l'origine, les étudiant(e)s du DESS MD de Dauphine), reprenant les fondements de l'analyse bayésienne dans un contexte d'études de cas et d'implémentation en langage R. Qui plus est, Jean-Michel est aussi à l'origine de la couverture de ce livre puisqu'elle a été réalisée par son père, Michel. Je les remercie vivement tous les deux.

Cette traduction n'aurait tout simplement pas été entamée sans un support financier initial de l'Université Paris Dauphine, obtenu grâce à l'insistance de Maria Esteban, directrice du CEREMADE, que je remercie très chaleureusement. La première partie du livre a été traduite avec brio par Claudia Lagos–Chopin, qui a su gérer son trilinguisme avec efficacité, et à qui j'exprime ma gratitude, ainsi qu'à son mari Nicolas, pour leur travail. Suite à l'arrivée de leur fille Alice et au bouleversement consécutif de leur emploi du temps, ils n'ont pas pu continuer cette traduction comme ils le désiraient et Loïs Rigouste de Télécom Paris a bien voulu reprendre le flambeau, accomplissant la traduction des quatre derniers chapitres avec efficacité et rapidité, tout en poursuivant sa thèse en parallèle. Je suis très reconnaissant à tous les trois de leur travail, les modifications apportées par mes soins étant simplement des actualisations de la seconde version anglaise. La relecture de parties du livre par Loïs Rigouste, Joachim et Rachel Robert, et Arafat Tayeb ont aussi permis de débusquer de nombreuses fautes de frappe qui m'avaient échappé[2].

Anne-Françoise Dutaud, secrétaire du laboratoire de Statistique du CREST, a également repris la traduction de la liste de référence en Bibtex (tout comme Manuella Delbois l'avait fait en son temps pour *Monte Carlo Statistical Methods*) et de l'actualisation des références dans le texte. Qu'elle ait pu s'acquiter de ce travail ingrat en quelques mois sans avoir fait de TEX auparavant est une mesure de son dévouement. (Comme toujours, la TEXpertise d'Olivier Cappé m'a été d'une aide précieuse.)

Des remerciements vont aussi à Nathalie Huilleret, de Springer-Verlag (Paris), qui a su gérer contretemps, gestion des droits et problèmes de production avec une grande efficacité.

Paris, France — Christian P. Robert
23 novembre 2005

[2]Il reste encore, avec une forte probabilité, des fautes de frappe que les lecteurs et lectrices sont invité(e)s à me signaler et qui seront affichées sur ma page web, à la rubrique *Books*. Merci.

Préface à la seconde édition de *The Bayesian Choice*

> "You can never know everything," Lan said quietly, "and part of what you know is always wrong. Perhaps even the most important part. A portion of wisdom lies in knowing that. A portion of courage lies in going on anyway."
>
> **Robert Jordan**, *Winter's Heart.*

Aperçu des Changements

Pourquoi une deuxième édition ? Quand on y réfléchit bien, il s'agit plutôt d'une troisième édition, car la version précédente, *The Bayesian Choice*, était en fait la traduction de la version française et incluait déjà des mises à jour et des corrections. Les raisons de cette nouvelle édition sont multiples. Depuis 1994, la communauté bayésienne a énormément évolué. La version précédente n'a pas seulement négligé d'importantes parties du domaine, mais elle a omis des avancées significatives survenues lors des sept dernières années.

Ainsi, la révolution MCMCa considérablement attisé les progrès de la modélisation bayésienne, avec des applications qui vont de la Statistique médicale au traitement du signal et à la Finance. Ces progrès n'étaient pas suffisamment soulignés dans l'édition de 1994. Par exemple, les méthodes MCMC n'y étaient présentées qu'à partir de l'avant-dernier chapitre.

Une autre avancée significative qui mérite notre attention est le développement de nouvelles approches pour les tests statistiques et, plus généralement, des outils de choix de modèles en connexion avec, et résultant des techniques MCMC, comme celle de saut réversible. D'autres avancées importantes incluent les modèles hiérarchiques et dynamiques dont le développement a commencé au début des années 1990.

Cette seconde édition est malgré tout loin d'être révolutionnaire par rapport à celle de 1994. Elle inclut cependant d'importantes avancées qui ont eu lieu depuis. Le seul chapitre véritablement nouveau traite du choix du modèle (Chap. 9), indépendamment de la théorie générale des tests (Chap. 5), parce que le choix de modèle se présente effectivement comme un problème différent et aussi parce qu'il exige des outils nouveaux, principalement informatiques. Pour cette raison, mais aussi pour souligner l'importance des techniques informatiques, le Chapitre 6, Chapitre 9 précédemment, a été placé plus haut dans le livre, après la présentation des fondements de la Statistique bayésienne. Le Chapitre 6 pourrait en fait être considéré comme un nouveau chapitre dans le sens où sa présentation a été profondément renouvelée à la lumière de dix ans de pratique des MCMC. Dans le Chapitre 3, la présentation des procédures non informatives a été élargie et inclut en particulier les a priori d'adéquation, puisque l'activité de recherche dans ce domaine a été assez intense ces dernières années. Le Chapitre 4 fait toujours référence aux problèmes généraux d'estimation mais j'ai ajouté une nouvelle section sur les modèles dynamiques, car ceux-ci font partie intégrante du développement de la Statistique bayésienne dans des domaines appliqués tels que le traitement du signal, la Finance et l'Économétrie.

Malgré une critique assez négative du Chapitre 11 par Mohan Delampady dans *The Mathematical Reviews*, j'ai décidé de maintenir ce chapitre de conclusion, car je considère qu'il offre un aperçu d'ensemble plus philosophique sur le sujet, le lecteur ayant très vraisemblablement déjà acquis une perspective suffisante pour comprendre de tels arguments. (En terme de programme de cours, ce chapitre peut être suggéré comme une lecture complémentaire, à l'instar des notes de fin de chapitre.)

Un autre changement notable, par comparaison avec l'édition précédente, est l'emphase moindre sur les principes de la Théorie de la Décision. Étant arrivé à la Statistique bayésienne par un chemin décisionnel, je crois toujours que les procédures statistiques doivent être fondées sur de tels principes. Cependant les développements des dix dernières années se sont principalement concentrés sur la méthodologie, y compris computationnelle, plus que sur la résolution plus large et plus ambitieuse des problèmes de décision (une fois de plus, méthodologie informatique comprise). Une partie du livre (qui comprend les Chapitres 6 et 7) est donc moins orientée vers la Théorie de la Décision, et, pour les Chapitres 8 à 10, a à peine changé.

En ce qui concerne la mise en page, des sous-sections et des séparations ont été introduites dans plusieurs sections afin d'améliorer la visibilité et la lecture. Un plus grand nombre de parties avancées ou incomplètes ont été déplacées en notes de fin de chapitre, suivant l'approche adoptée dans *Monte Carlo Statistical Methods*, écrit avec George Casella. La fin d'un exemple est associée au symbole ‖, tandis que la fin d'une démonstration est indiquée par le symbole □.

Plusieurs livres sur la Statistique bayésienne sont apparus entre-temps, parmi lesquels Bernardo et Smith (1994), Carlin et Louis (2001), Gelman

et al. (2003), O'Hagan (1994), O'Hagan et Forster (2002) et Schervish (1995). Cependant chacun de ces livres a soit mis l'accent sur l'approfondissement des aspects théoriques à un niveau mathématique très élevé (Bernardo et Smith, 1994, O'Hagan, 1994, O'Hagan et Forster, 2002, Schervish, 1995) et a ainsi visé une audience plus mûre que celle de ce livre, soit fait ressortir une vision différente de la pratique de la Statistique bayésienne (Carlin et Louis, 2001, Gelman *et al.*, 2003), en perdant par exemple le lien avec la Théorie de la Décision développée dans ce livre.

Remerciements

J'ai toujours éprouvé des sentiments mêlés sur le fait d'ajouter une section de remerciements dans un livre. En fait, cette section ne dira pas grand-chose à l'immense majorité des lecteurs, sauf à révéler certaines idiosyncrasies de l'auteur qui feraient sans doute mieux de rester cachées ! Elle pourrait aussi contrarier certaines personnes concernées parce qu'elles ne sont pas citées, ou parce qu'elles ne sont pas citées selon leurs attentes, ou même parce qu'elles le sont ! En revanche, une exigence éthique de base de tout travail intellectuel est de reconnaître ses sources. Cela s'étend à mon avis aux suggestions qui ont contribué à améliorer ce travail, à le rendre plus clair ou simplement différent. Il s'agit d'un petit témoignage de gratitude envers les personnes suivantes, pour le temps qu'ils et elles ont consacré aux versions successives de cette édition, pour que leurs efforts soient vus et connus de tous !

Bien que cette édition soit "juste" une révision, le temps passé sur cet ouvrage a été, en grande partie, volé aux soirs, aux matins (très tôt) et aux week-ends revenant normalement à Brigitte, Joachim et Rachel ! Je leur suis ainsi très reconnaissant pour avoir lu et joué (presque) sans faire de bruit pendant que je tapais furieusement sur mon clavier et cherchais désespérément dans des piles de papiers telle ou telle référence. Et aussi pour écouter Bartoli et Gudjónsson, plutôt que Manau ou Diana Krall ! Je ne peux pas promettre que cette expérience ne se répétera jamais, mais en attendant je m'engage à trouver plus de temps disponible pour lire les aventures de *Mister Bear to the Rescue*, assiéger le château Playmobil au complet, jouer aux échecs ou faire du vélo les dimanches après-midi !

Je suis reconnaissant à de nombreuses personnes pour les améliorations de cette édition. Pour commencer, j'ai eu un flot constant de retours et de suggestions de la part de ceux qui enseignent à partir de ce livre. Ce groupe inclut Ed Green, Tatsuya Kubokawa, et Marty Wells. En particulier, Judith Rousseau, cycliste radicale et Jordanienne autant que bayésienne, a contribué à la réorganisation du Chapitre 3. J'ai eu aussi beaucoup de commentaires utiles de plusieurs personnes, en particulier des deux "Cambridge Frenchies" Christophe Andrieu et Arnaud Doucet (sans compter un mémorable accueil pendant une semaine de retraite à Cambridge pour finir le Chapitre 6), ainsi que de Jim Berger (pour son soutien en général et pour m'avoir fourni des *preprints* sur le choix de modèle en particulier), d'Olivier Cappé (qui a aussi

installé Linux sur mon portable et par conséquent m'a apporté une immense liberté pour travailler sur le livre n'importe où, du bac à sable au métro et plus tard au CREST, d'où Unix est désormais banni !), de Maria De Iorio, de Jean-Louis Fouley, de Malay Ghosh (pour sa critique du livre très positive dans *JASA*), de Jim Hobert (qui m'a aidé à clarifier les Chapitres 6 et 10), d'Ana Justel, de Stephen Lauritzen (pour avoir signalé des erreurs sur les distributions de Wishart), d'Anne Philippe, de Walter Racugno (qui m'a donné l'opportunité de faire un cours concernant le choix des modèles à Cagliari l'automne dernier, cours qui constitue l'essentiel du Chapitre 7), d'Adrian Raftery, d'Anne Sullivan Rosen (pour le style de cette préface) et Jean-Michel Zakoian (pour ses conseils sur les nouvelles parties concernant les modèles dynamiques). Je profite aussi de cette occasion pour remercier d'autres ami(e)s et collègues comme George Casella, Jérôme Dupuis, Merrilee Hurn, Kerrie Mengersen, Eric Moulines, Alain Monfort, et Mike Titterington. Depuis que je travaille avec eux et avec elles, ils et elles m'ont donné une vision plus large du domaine, qui est, espérons-le, incluse dans cette version. En particulier, l'expérience de l'écriture de *Monte Carlo Statistical Methods* avec George Casella ces dernières années a laissé ses marques dans ce livre non seulement à travers le fichier de style et l'inclusion de notes en fin de chapitre, mais aussi pour un sens plus aigu de l'essentiel. Manuela Delbois m'a aidé très aimablement à transformer le texte de TEX à LATEX, puis à inclure les additions ultérieures et l'index. Et, *last but not least !*, John Kimmel et Jenny Wolkowicki de Springer-Verlag ont été très efficaces, en m'encourageant à écrire cette nouvelle édition pour le premier, en gardant le contrôle du calendrier et en faisant publier le livre à temps pour la seconde. Inutile de dire que l'avertissement d'usage s'applique : toute coquille, erreur, confusion, formulation obscure restante est de ma responsabilité et rien que de la mienne !

In Memoriam

Une pensée très émue pour deux personnes dont l'*absence* a marqué cette nouvelle édition. Durant l'été 1997, j'ai perdu mon ami Costas Goutis lors d'un accident de plongée à Seattle. Je ne suis pas, et de loin, le seul à regretter profondément son départ, mais sans aucun doute ce livre aurait bénéficié de sa vision des choses s'il avait été là... Deux étés plus tard, en 1999, Bernhard K. Flury est mort dans un accident de montagne dans les Dolomites. Bien que la critique de nos livres respectifs se soit toujours limitée aux couleurs de couverture, au point de s'envoyer l'un à l'autre une version piratée de nos livres avec les "bonnes" couleurs, le monde est moins drôle sans son sens de l'humour à nul autre pareil...

Paris, France — Christian P. Robert
Mars 2001

Table des matières

Préface VII

Préface à la seconde édition anglaise XI

1 Introduction 1
- 1.1 Problèmes statistiques et modèles statistiques 1
- 1.2 Le paradigme bayésien et le principe de dualité 9
- 1.3 Principes de vraisemblance et d'exhaustivité 15
  - 1.3.1 Exhaustivité 15
  - 1.3.2 Principe de vraisemblance 17
  - 1.3.3 Dérivation du principe de vraisemblance 20
  - 1.3.4 Mise en œuvre du principe de vraisemblance 21
  - 1.3.5 Estimation par maximum de vraisemblance 23
- 1.4 Distributions a priori et a posteriori 24
- 1.5 Distributions a priori impropres 30
- 1.6 Le choix bayésien 34
- 1.7 Exercices 35
- 1.8 Notes 50

2 Les bases de la Théorie de la Décision 55
- 2.1 Évaluation des estimateurs 55
- 2.2 La fonction d'utilité 58
- 2.3 Utilité et coût 66
- 2.4 Deux optimalités : minimaxité et admissibilité 71
  - 2.4.1 Estimateurs randomisés 71
  - 2.4.2 Minimaxité 73
  - 2.4.3 Existence d'une règle minimax et d'une stratégie maximin 76
  - 2.4.4 Admissibilité 81
- 2.5 Fonctions de coût usuelles 85
  - 2.5.1 Le coût quadratique 85

2.5.2 L'erreur de coût absolu 87
2.5.3 Le coût $0-1$ 89
2.5.4 Coûts intrinsèques 89
2.6 Critiques et alternatives 91
2.7 Exercices 94
2.8 Notes 105

3 Des informations a priori aux lois a priori 113
3.1 La difficulté du choix d'une loi a priori 113
3.2 Détermination subjective et approximations 115
3.2.1 Existence 115
3.2.2 Approximations de la loi a priori 117
3.2.3 Lois a priori d'entropie maximale 118
3.2.4 Approximations paramétriques 119
3.2.5 Autres techniques 122
3.3 Lois a priori conjuguées 122
3.3.1 Introduction 122
3.3.2 Justifications 124
3.3.3 Familles exponentielles 124
3.3.4 Lois conjuguées des familles exponentielles 130
3.4 Critiques et extensions 132
3.5 Lois a priori non informatives 137
3.5.1 Les lois a priori de Laplace 137
3.5.2 Lois invariantes 139
3.5.3 La loi a priori de Jeffreys 139
3.5.4 Lois de référence 143
3.5.5 Lois a priori coïncidantes 147
3.5.6 D'autres approches 151
3.6 Validation a posteriori et robustesse 152
3.7 Exercices 156
3.8 Notes 169

4 Estimation bayésienne ponctuelle 175
4.1 Inférence bayésienne 175
4.1.1 Introduction 175
4.1.2 Estimateur MAP 176
4.1.3 Principe de vraisemblance 177
4.1.4 Espace des paramètres restreint 179
4.1.5 Précision des estimateurs de Bayes 181
4.1.6 Prévision 182
4.1.7 Retour à la décision 184
4.2 Théorie bayésienne de la décision 184
4.2.1 Estimateurs de Bayes 184
4.2.2 Les lois a priori conjuguées 187
4.2.3 Estimation du coût 190

4.3 Modèles d'échantillonnage . . . 191
4.3.1 Règle de succession de Laplace . . . 192
4.3.2 Le problème du tramway . . . 193
4.3.3 Modèles de capture-recapture . . . 194
4.4 Le cas particulier du modèle normal . . . 198
4.4.1 Introduction . . . 198
4.4.2 Estimation de la variance . . . 199
4.4.3 Modèles linéaires et G-priors . . . 203
4.5 Modèles dynamiques . . . 206
4.5.1 Introduction . . . 206
4.5.2 Le modèle AR . . . 210
4.5.3 Le modèle MA . . . 211
4.5.4 Le modèle ARMA . . . 214
4.6 Exercices . . . 215
4.7 Notes . . . 230

5 Tests et régions de confiance . . . 237
5.1 Introduction . . . 237
5.2 Une première approche de la théorie des tests . . . 238
5.2.1 Tests décisionnels . . . 238
5.2.2 Le facteur de Bayes . . . 241
5.2.3 Modification de la loi a priori . . . 244
5.2.4 Hypothèses nulles ponctuelles . . . 245
5.2.5 Lois a priori impropres . . . 248
5.2.6 Pseudo-facteurs de Bayes . . . 251
5.3 Comparaisons avec l'approche classique . . . 258
5.3.1 Tests UPP et UPPS . . . 258
5.3.2 Lois a priori les moins favorables . . . 262
5.3.3 Critiques . . . 263
5.3.4 Les p-values . . . 266
5.3.5 Réponses bayésiennes moins favorables . . . 268
5.3.6 Le cas unilatéral . . . 271
5.4 Une deuxième approche décisionnelle . . . 273
5.5 Régions de confiance . . . 277
5.5.1 Intervalles de crédibilité . . . 278
5.5.2 Intervalles de confiance classiques . . . 280
5.5.3 Évaluation décisionnelle des ensembles de confiance . . . 283
5.6 Exercices . . . 285
5.7 Notes . . . 298

6 Méthodes de calcul bayésien . . . 305
6.1 Difficultés de mise en œuvre . . . 305
6.2 Méthodes classiques d'approximation . . . 313
6.2.1 Intégration numérique . . . 314
6.2.2 Les méthodes de Monte Carlo . . . 314

6.2.3 L'approximation analytique de Laplace 319
6.3 Méthodes de Monte Carlo par chaînes de Markov 322
6.3.1 Les MCMC en pratique 323
6.3.2 Algorithmes de Metropolis-Hastings 325
6.3.3 L'échantillonnage de Gibbs 329
6.3.4 Rao-Blackwellisation 332
6.3.5 L'échantillonnage de Gibbs général 334
6.3.6 L'échantillonnage par tranche 339
6.3.7 Impact sur la statistique bayésienne 341
6.4 Estimation bayésienne de mélanges 342
6.5 Exercices 344
6.6 Notes 360

7 Choix et comparaison de modèles 369
7.1 Motivations 369
7.1.1 Choix entre plusieurs modèles 371
7.1.2 Champs d'application 374
7.2 Comparaison bayésienne de modèles 375
7.2.1 Modélisation spécifique de l'a priori 375
7.2.2 Facteurs de Bayes 378
7.2.3 Le critère de Schwarz 380
7.2.4 Déviance bayésienne 382
7.3 Aspects numériques 384
7.3.1 Échantillonnage d'importance pour facteurs de Bayes . . 385
7.3.2 Échantillonnage par passerelle 387
7.3.3 Méthodes MCMC 388
7.3.4 MCMC à sauts réversibles 393
7.4 Moyenne de modèles 395
7.5 Projections de modèles 399
7.6 Adéquation à une famille de lois 405
7.7 Exercices 408
7.8 Notes 418

8 Admissibilité et classes complètes 423
8.1 Introduction 423
8.2 Admissibilité des estimateurs de Bayes 424
8.2.1 Caractérisations générales 424
8.2.2 Conditions aux limites 426
8.2.3 Estimateurs de Bayes généralisés inadmissibles 428
8.2.4 Représentations différentielles 429
8.2.5 Conditions de récurrence 431
8.3 Conditions nécessaires et suffisantes d'admissibilité 433
8.3.1 Risques continus 434
8.3.2 Condition suffisante de Blyth 436
8.3.3 Condition nécessaire et suffisante de Stein 441

8.3.4 Un autre théorème limite . . . 441
8.4 Classes complètes . . . 443
8.5 Conditions nécessaires d'admissibilité . . . 446
8.6 Exercices . . . 450
8.7 Notes . . . 460

9 Invariance, mesures de Haar et estimateurs équivariants . . . 463
9.1 Principes d'invariance . . . 463
9.2 Le cas particulier des paramètres de position . . . 465
9.3 Problèmes de décision invariants . . . 468
9.4 Distributions non informatives équivariantes . . . 473
9.5 Le théorème de Hunt-Stein . . . 479
9.6 L'invariance en Statistique bayésienne . . . 483
9.7 Exercices . . . 484
9.8 Notes . . . 492

10 Extensions hiérarchique et empirique . . . 495
10.1 Lois a priori incomplètes . . . 495
10.2 Analyse bayésienne hiérarchique . . . 498
10.2.1 Modèles hiérarchiques . . . 498
10.2.2 Justifications . . . 501
10.2.3 Décompositions conditionnelles . . . 504
10.2.4 Problèmes numériques . . . 507
10.2.5 Extensions hiérarchiques du modèle normal . . . 509
10.3 Optimalité des estimateurs bayésiens hiérarchiques . . . 514
10.4 L'alternative bayésienne empirique . . . 518
10.4.1 Le principe bayésien empirique non paramétrique . . . 519
10.4.2 Principe bayésien empirique paramétrique . . . 521
10.5 Justifications bayésiennes empiriques de l'effet Stein . . . 525
10.5.1 Estimation ponctuelle . . . 525
10.5.2 Évaluation de la variance . . . 528
10.5.3 Régions de confiance . . . 529
10.5.4 Commentaires . . . 531
10.6 Exercices . . . 532
10.7 Notes . . . 544

11 Une défense du choix bayésien . . . 549

A Distributions de probabilité . . . 563

B Notations . . . 567

Références . . . 571

Index des noms . . . 611

Index des matières 621

Liste des tableaux

2.1 Fonction d'utilité 75

3.1 Information a priori de capture et de survie 115
3.2 Loi a priori de capture et de survie 116
3.3 Étendue des valeurs des moments a posteriori 121
3.4 Lois a priori conjuguées naturelles 131
3.5 Lois a priori de référence coïncidantes 151
3.6 Approximation par mélange de lois conjuguées 172

4.1 Estimateurs de Bayes pour familles exponentielles 187
4.2 Probabilités de capture 195
4.3 Partition de la population de capture 195
4.4 Loi a posteriori de la population de cerfs 197
4.5 Espérance a posteriori de la population de cerfs 197
4.6 Population de cerfs estimée 198

5.1 Probabilités a posteriori de $p = 1/2$ 247
5.2 Probabilités a posteriori de $\theta = 0$ 247
5.3 Probabilités a posteriori de $\theta = 0$ 248
5.4 Probabilités a posteriori de $|\theta| < 1$ 249
5.5 Probabilités a posteriori de $\theta = 0$ 249
5.6 Probabilités a posteriori de $\theta = 0$ 250
5.7 Comparaison entre p-values et réponses bayésiennes 269
5.8 Comparaison entre p-values et réponses bayésiennes 270
5.9 Facteurs de Bayes et probabilités a posteriori 271
5.10 Comparaison entre p-values et probabilités a posteriori 272
5.11 Intervalles α-crédibles pour la loi $\mathscr{B}(n, p)$ 279

6.1 Paramètres de radiographies des poumons 311
6.2 Fréquences de passages de voitures 359

7.1 Circonférences d'orangers 373
7.2 Adéquation des modèles d'orangers 390
7.3 Paramètres pour divergences de Kullback-Leibler 402
7.4 Sous-modèles pour le cancer du sein 404
7.5 Nombre de femmes dans une file d'attente 416

10.1 Probabilités a posteriori et intervalles de confiance 509
10.2 Intentions d'achat de voiture par foyer 537
10.3 Achats de voitures et intentions 537

Table des figures

1.1 Taux de chômage mensuel et accidents 5
1.2 Histogramme d'une poitrine 5

2.1 Utilité moyenne 65
2.2 Comparaison des risques 75
2.3 Ensemble de risques de Bernoulli 78

3.1 Deux estimateurs de la moyenne 121
3.2 Densités $\mathscr{I}\mathscr{N}(\alpha, \mu, \tau)$ 128
3.3 Trois lois a priori de pile ou face 134
3.4 Lois a posteriori de pile ou face 134
3.5 Lois a posteriori pour cinquante observations 135

4.1 Évaluations de l'erreur bayésienne et fréquentiste 183
4.2 Cours de l'action IBM moyennée 209
4.3 Deux lois a priori sur ϱ 232
4.4 Échantillon du modèle de volatilité stochastique 234
4.5 Allocations pour le modèle de volatilité stochastique 235

5.1 Loi a priori intrinsèque pour test exponentiel 255

6.1 Variation des approximations de Monte Carlo 318
6.2 Chaîne de Markov pour modèle normal répulsif 328
6.3 Histogrammes de la loi bêta-binomiale 332

7.1 Histogramme des données galactiques 372
7.2 Simulations du nombre de composantes 396

8.1 Ensemble de risque et estimateurs admissibles 444

10.1 DAG pour le modèle HIV 498
10.2 Convergences pour l'expérience des rats 508

10.3 Échantillons de Gibbs pour l'expérience des rats 509

1

Introduction

> "Sometimes the Pattern has a randomness to it—to our eyes, at least—but what chance that you should meet a man who could guide you in this thing, and he one who could follow the guiding ?"
>
> **Robert Jordan**, *The Eye of the World.*

1.1 Problèmes statistiques et modèles statistiques

L'objet principal de la Statistique est de mener, grâce à l'observation d'un phénomène aléatoire, une *inférence* sur la distribution probabiliste à l'origine de ce phénomène, c'est-à-dire de fournir une analyse (ou une description) d'un phénomène passé, ou une prédiction d'un phénomène à venir de nature similaire[3]. Dans ce livre nous insistons sur les aspects *décisionnels* de l'inférence statistique parce que, tout d'abord, ces analyses et prédictions sont la plupart du temps motivées par un but objectif (une entreprise devrait-elle lancer un nouveau produit ? un bateau de course modifier sa trajectoire ? un nouveau médicament être mis sur le marché ou à la vente ? un individu vendre ses actions ? etc.) ayant des conséquences mesurables (résultats financiers, classement à la fin de la course, taux de guérison des patients, bénéfices, etc). Ensuite, parce que proposer des procédures inférentielles implique qu'on doit

[3]Comme la plupart des définitions formelles, cette vision de la Statistique laisse de côté quelques aspects supplémentaires de la Statistique appliquée tels que la collecte de données (sondages, plans d'expérience, etc). C'est le cas aussi de cet ouvrage, même si nous ne voulons pas mésestimer l'importance de ces sujets, non couverts ici.

être prêt à les défendre, c'est-à-dire à justifier le fait qu'elles soient préférables à d'autres. Il est donc nécessaire d'avoir un outil d'évaluation adapté à la comparaison de diverses procédures. Cette tâche est la raison d'être de la Théorie de la Décision.

Nous insistons également sur le fait que la Statistique doit être considérée comme l'*interprétation* d'un phénomène naturel, plutôt que son explication. En effet, l'inférence statistique s'accompagne d'une modélisation probabiliste du phénomène observé et implique nécessairement une étape de formalisation réductrice. Sans cette base probabiliste, aucune conclusion utile ne pourra être tirée.

Exemple 1.1. Les feux de forêt apparaissent généralement au hasard. Cependant, certains facteurs écologiques et atmosphériques favorisent leur déclenchement. Une détermination de la probabilité p d'apparition d'un feu comme fonction de ces divers facteurs devrait aider à la prévention des feux de forêt, même si une telle modélisation est évidemment incapable de conduire à l'éradication de ces feux et ne peut prendre en compte tous les facteurs impliqués. Une approche plus réductrice est d'imposer une forme paramétrique à la fonction p, prenant en compte des contraintes physiques sur les facteurs explicatifs. Par exemple, notant h le taux d'humidité, t la température, x le degré de gestion de la forêt, un *modèle logistique* peut être proposé, de la forme

$$p = \exp(\alpha_1 h + \alpha_2 t + \alpha_3 x)/\left[1 + \exp(\alpha_1 h + \alpha_2 t + \alpha_3 x)\right] ,$$

la phase statistique se chargeant de l'évaluation des paramètres $\alpha_1, \alpha_2, \alpha_3$. ‖

Apposer un modèle probabiliste sur un phénomène inexpliqué peut paraître dans certains cas trop réducteur, car il est possible que le phénomène observé soit entièrement déterministe, sans que la fonction régulatrice du processus soit connue ni qu'il soit possible de la reconstruire à partir des observations. C'est le cas par exemple des *phénomènes chaotiques* où, d'un point de vue statistique, une suite d'observations ne peut pas être distinguée d'une suite de variables aléatoires (voir Bergé *et al.*, 1984 et Gleick, 1987). Les générateurs pseudo-aléatoires sont en fait fondés sur cette propriété. Bien qu'ils reposent sur des algorithmes itératifs déterministes de la forme

$$a_{t+1} = f(a_t),$$

ils imitent–*simulent*–de façon satisfaisante le comportement d'une suite de variables aléatoires (voir Devroye, 1985, Gentle, 1998, Robert et Casella, 2004 pour une description des générateurs les plus courants).

Cependant, même si elle est valable d'un point de vue philosophique, cette critique de la modélisation probabiliste ne tient pas si nous considérons celle-ci sous l'angle de l'interprétation, évoquée ci-dessus. Ces modèles permettent d'incorporer simultanément les informations disponibles sur le phénomène (facteurs déterminants, fréquence, amplitude, etc.) et les incertitudes inhérentes à ces informations. Ils autorisent donc un discours qualitatif sur le

problème en fournissant, à travers la théorie des probabilités, un véritable *calcul de l'incertain* qui permet de dépasser le stade descriptif des modèles déterministes. C'est d'ailleurs la raison pour laquelle une interprétation probabiliste est nécessaire pour conduire une inférence statistique : elle donne un cadre qui permet de replacer le phénomène singulier observé dans la globalité d'un modèle et autorise ainsi les analyses et les généralisations. Loin de représenter un détournement des objectifs inférentiels, imposer une structure probabiliste qui n'est qu'une simple approximation de la réalité est essentiel pour que le traitement statistique qui en découle permette une compréhension plus profonde et plus proche du phénomène considéré.

Évidemment la modélisation probabiliste ne peut être défendue que si elle fournit une représentation suffisamment proche du phénomène observé. Une critique plus prosaïque de la modélisation probabiliste est qu'il est difficile de connaître exactement la distribution probabiliste sous-jacente de la génération des observations, c'est-à-dire savoir s'il s'agit de la loi normale, exponentielle, binomiale, etc., sauf dans des cas exceptionnels.

Exemple 1.2. On observe une substance radioactive de demi-vie H inconnue. Pour une particule donnée de cette substance, le temps passé avant désintégration suit exactement une loi exponentielle[4] de paramètre $\log(2)/H$. L'observation de plusieurs de ces particules permettra ainsi de mener une inférence sur H. ‖

Exemple 1.3. Pour déterminer le nombre N de bus dans une ville, on peut suivre la stratégie inférentielle suivante : observer les bus pendant toute une journée et noter leurs numéros. Ensuite on répète la même expérience le lendemain en relevant les numéros des bus déjà répertoriés la veille, n. Si vingt bus ont été observés la première journée et trente la deuxième, n suit une loi hypergéométrique,$\mathcal{H}(30, N, 20/N)$, et la connaissance des propriétés de cette distribution permet, par exemple, l'approximation de N par $20(30/n)$. Cette méthode dite de *capture-recapture*, a donné lieu à de nombreux développements moins anecdotiques en écologie et dynamique des populations (voir le Chapitre 4). ‖

Nous pourrions citer d'autres exemples où la distribution des observations est parfaitement connue, grâce à des considérations physiques, économiques ou autres. Cependant, dans la plupart des cas, la modélisation statistique est bien réductrice au sens où elle n'est qu'une approximation de la réalité, perdant une partie de sa richesse mais gagnant en efficacité.

Exemple 1.4. Les variations des prix et des salaires sont fortement reliées. Une façon de représenter cette dépendance est de supposer une relation linéaire

[4]Voir Appendice A pour une liste des distributions les plus courantes.

$$\Delta P = a + b\,\Delta S + \epsilon,$$

où ΔP et ΔS sont les variations de prix et de salaires, a et b les coefficients inconnus et ϵ le terme d'erreur. Une façon, drastique, de simplifier plus avant cette relation est de supposer que ϵ est normalement distribué. Bien que ϵ soit effectivement une variable aléatoire, de nombreux facteurs doivent être considérés dans la détermination des prix et des salaires et il est impossible d'établir la distribution de ϵ. Néanmoins, outre une justification par le *Théorème Central Limit* (soit l'effet additionnel d'une multitude de petits facteurs de même magnitude), cette modélisation avancée permet aussi une analyse statistique plus minutieuse, qui est valide même si la distribution de ϵ n'est pas exactement normale. (Voir aussi Exercice 1.3.) ‖

Exemple 1.5. Considérons le jeu de données de la Figure 1.1, qui représente le taux mensuel de chômage en fonction du nombre d'accidents (en milliers) dans le Michigan entre 1978 et 1987. Lenk (1999) soutient l'existence d'une relation entre ces deux variations : un taux plus élevé de chômage entraîne une diminution de la circulation sur les routes, et donc du nombre d'accidents. Une simplification supplémentaire est alors de postuler une structure paramétrique de dépendance, comme le modèle de *régression de Poisson*

$$N|\varrho \sim \mathcal{P}(\exp\{\beta_0 + \beta_1 \log(\varrho)\}), \tag{1.1}$$

où N représente le nombre d'accidents et ϱ le taux de chômage pour le même mois. La Figure 1.1 donne ainsi l'espérance estimée $\mathbb{E}[N|\varrho]$, qui a tendance à confirmer l'impact décroissant du chômage sur les accidents. Mais la validité de la modélisation (1.1) demande d'abord à être évaluée en utilisant des tests d'adéquation ou d'autres techniques de choix de modèles. (Voir le Chapitre 7.) ‖

Dans certains cas, l'effet réducteur est volontairement recherché pour ses conséquences positives de *lissage* des données. Il peut aussi enlever en partie les perturbations moins importantes d'un phénomène et souvent améliorer son analyse en mettant en évidence les facteurs essentiels comme dans l'exemple suivant.

Exemple 1.6. Les radiographies médicales peuvent être représentées comme une grille de $1\,000 \times 1\,200$ points fondamentaux appelés *pixels*, qui prennent un niveau de gris associé à un nombre compris entre 0 et 256. Par exemple, la Figure 1.2 donne l'histogramme des niveaux de gris pour une radiographie typique des poumons. Si nous considérons un pixel comme une variable aléatoire à valeurs dans $\{0, 1, \ldots, 256\}$, donc discrète, l'histogramme donne une approximation de la distribution de cette variable aléatoire. Comme le montre la figure, cette distribution est plutôt complexe, mais approximativement bimodale. Cette particularité a été observée dans la plupart des radiographies et suggère une modélisation de la distribution via une approximation continue par un *mélange de deux distributions normales* de densité

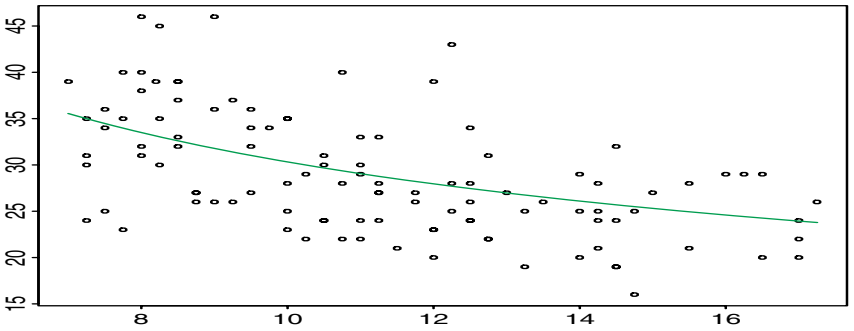

Fig. 1.1. Taux de chômage mensuel en fonction du nombre d'accidents (en milliers) dans le Michigan, de 1978 à 1987. (*Source :* Lenk, 1999.)

$$f(x) = \frac{p}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right] + \frac{1-p}{\sqrt{2\pi}\sigma_2} \exp\left[-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right] . \qquad (1.2)$$

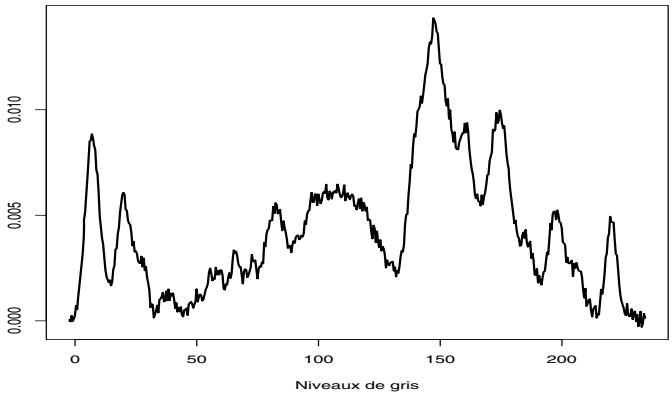

Fig. 1.2. Histogramme de niveau de gris d'une radiographie de la poitrine et sa modélisation par un mélange à deux composantes. (*Source :* Plessis, 1989.)

Évidemment cette modélisation a considérablement lissé l'histogramme (voir la Figure 1.2), mais permet aussi une description de l'image avec cinq paramètres, sans perte substantielle d'information. Il a été déterminé que les deux modes importants de la vraie distribution correspondent en fait aux deux

régions de la poitrine, les *poumons* et le *mediastinum*. Cette technique de lissage est utilisée dans un algorithme d'amélioration des radiographies appelé *Parametric Histogram Specification* (voir Plessis, 1989). Nous consacrerons la Section 6.4 à l'estimation bayésienne des mélanges. ‖

Face à cette réduction de la complexité du phénomène observé, deux approches statistiques s'opposent. La première approche suppose que l'inférence statistique doit prendre en compte cette complexité autant que possible et cherche donc à estimer la distribution sous-jacente du phénomène sous des hypothèses minimales, en ayant recours en général à l'estimation fonctionnelle (densité, fonction de régression, etc.). Cette approche est dite *non paramétrique*. Par opposition, l'approche *paramétrique* représente la distribution des observations par une fonction de densité $f(x|\theta)$, où seul le paramètre θ (de dimension finie) est inconnu.

Nous considérons cette seconde approche comme plus pragmatique dans la mesure où elle prend en compte le fait qu'un nombre fini d'observations ne peut estimer qu'un nombre fini de paramètres. De plus, la modélisation paramétrique permet une évaluation des outils inférentiels *pour une taille d'échantillon finie*, au contraire des méthodes non paramétriques, plus élaborées, dont la principale justification est asymptotique et qui ne peuvent donc s'appliquer que lorsque la taille de l'échantillon devient *infinie* (voir Field et Ronchetti, 1990, qui étudient l'applicabilité des résultats asymptotiques pour des échantillons à taille finie). Bien entendu, certaines approches non paramétriques, comme les tests (Hàjek et Sidàk, 1968), évacuent complètement l'aspect d'estimation et les problèmes de tailles d'échantillons infinies en construisant des statistiques de test indépendantes des distributions, mais leurs applications restent limitées.

Les deux approches ont leurs avantages respectifs et nous ne justifierons pas d'avantage le choix paramétrique. Naturellement, il existe aussi toute une littérature sur la construction de modèles. Voir par exemple Cox (1990) et Lehmann (1990) pour des références ainsi que pour des réflexions sur la notion même de modèle statistique. Nous verrons dans le Chapitre 7 quelques approches pour la comparaison de modèles qui peuvent être utilisées dans l'étape de modélisation, c'est-à-dire quand plusieurs modèles potentiels 's'affrontent".

Nous ne considérons dans ce livre que l'approche *paramétrique.* Nous supposons que les observations $x_1, \ldots, x_n$, sur lesquelles l'analyse statistique se fonde, proviennent de lois de probabilité paramétriques, donc que x_i $(1 \leq i \leq n)$ a une distribution de densité $f_i(x_i|\theta_i, x_1, \ldots, x_{i-1})$ sur $\mathbb{R}^p$, telle que le paramètre θ_i soit inconnu et la fonction f_i soit connue (voir l'Exercice 1.2 sur l'ambiguïté formelle de cette définition et la Note 1.8.2 pour des indications sur l'approche bayésienne de la statistique non paramétrique). Ce modèle peut être représenté plus succinctement par

$$x \sim f(x|\theta),$$

où x est le vecteur des observations et θ l'ensemble des paramètres, $\theta_1, \ldots, \theta_n$, éventuellement tous égaux. Cette représentation est unificatrice dans le sens où elle aborde de manière similaire une observation isolée, des observations dépendantes, et des observations distribuées de façon indépendante et identiquement distribuées (*iid*) $x_1, \ldots, x_n$ de même loi, $f(x_1|\theta)$. Dans le dernier cas, $x = (x_1, \ldots, x_n)$ et

$$f(x|\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

Notons que dans ce livre nous écrirons de manière identique les densités de variables aléatoires continues et discrètes, la mesure de référence étant fournie naturellement par le contexte. De plus, nous utiliserons la notation *"x est distribué selon f"* ou "$x \sim f$" au lieu de "*x est une observation de la distribution de densité f*" par souci de concision[1]. La plupart du temps, l'échantillon est réduit à une observation unique pour des raisons de simplification mais aussi parce que souvent nous avons affaire à des distributions pour lesquelles la taille de l'échantillon ne compte pas, car elles admettent une statistique exhaustive de dimension constante (voir la Section 1.3 et le Chapitre 3).

Définition 1.7. *Un modèle paramétrique statistique consiste en l'observation d'une variable aléatoire x distribuée selon $f(x|\theta)$, où seulement le paramètre θ est inconnu et appartient à un espace de dimension finie.*

Une fois le modèle statistique identifié, l'objectif principal de l'analyse statistique est de nous conduire à une *inférence* sur le paramètre θ. C'est à dire que nous utilisons l'observation de x pour améliorer notre connaissance du paramètre θ, afin de pouvoir prendre une décision concernant le paramètre, c'est à dire d'estimer une fonction de θ ou un futur événement dont la distribution dépend de θ. L'inférence peut concerner certaines composantes de θ, précisément ("*Quelle est la valeur de θ_1 ?*") ou non ("*θ_2 est-t-il plus grand que θ_3 ?*"). Une distinction est souvent faite entre *problèmes d'estimation* et *problèmes de tests*, suivant qu'on cherche la valeur exacte des paramètres (ou de certaines fonctions des paramètres) ou seulement la vérification d'une hypothèse sur ces paramètres. Par exemple, les deux livres

[1] Ce livre ne suit pas la convention probabiliste habituelle, qui note les variables aléatoires par des lettres majuscules, par exemple X, et leur *réalisation*, qui n'est autre que leur valeur observée, par la lettre minuscule correspondante, soit x, comme dans $P(X \leq x)$. Ceci s'explique par le fait que, d'un point de vue bayésien, nous conditionnons en la valeur réalisée x et considérons le paramètre θ comme une variable aléatoire. L'utilisation d'une majuscule grecque peut amener à une confusion extrême puisque Θ est plutôt, par convention, l'espace des paramètres. Cela rend aussi plus facile l'utilisation des expressions conditionnelles, nombreuses dans les calculs bayésiens. Dans les quelques cas où cette pratique prête à confusion, nous reviendrons à la convention usuelle.

de référence de la Statistique classique, Lehmann (1983) et Lehmann et Casella (1998), sont consacrés respectivement à chacun de ces sujets. D'autres auteurs ont proposé une distinction plus subtile entre *estimation* et *évaluation* des procédures d'estimation (voir, par exemple, Casella et Berger, 2001). Plus généralement l'inférence recouvre tout phénomène aléatoire lié à θ et inclut aussi la *prévision*, qui est l'évaluation de la distribution d'une future observation y dépendante de θ (et probablement de l'observation courante de x), $y \sim g(y|\theta, x)$. Nous verrons par la suite que ces divisions sont un peu artificielles, car tous les problèmes inférentiels peuvent se ramener à des problèmes d'estimation quand ils sont considérés dans une perspective de Théorie de la Décision.

Le choix du "tout paramétrique" fait dans ce livre est bien entendu critiquable, puisque nous ne pouvons pas toujours supposer que la distribution des observations est connue à un paramètre (de dimension finie) près. Cependant, outre le fait qu'un traitement rigoureux des méthodes bayésiennes non paramétriques demande un bagage théorique plus important, nous insistons sur le fait que cette réduction permet des développements plus profonds du processus inférentiel, même si cela peut paraître paradoxal. Les critiques sur le caractère réducteur de l'approche statistique et, a fortiori, du choix paramétrique, s'accompagnent en réalité d'autres critiques sur le choix des critères d'évaluation et de l'objectif même de la Théorie de la Décision, comme nous le verrons dans le Chapitre 2. Cependant, nous soutenons ces choix sur la base que ces étapes de plus en plus réductrices sont des exigences minimales pour qu'une approche statistique soit cohérente (c'est-à-dire fasse preuve de cohérence interne). Effectivement le but ultime de l'analyse statistique, dans l'énorme majorité des cas, est de défendre le choix d'une *décision* comme *optimale* (ou au moins raisonnable). Il est donc nécessaire de pouvoir comparer les différents processus inférentiels disponibles. Les sections qui suivent présentent les bases de l'analyse statistique bayésienne, laquelle nous paraît être l'approche la plus appropriée pour cette détermination des procédures optimales[2]. Il s'agit aussi de la méthode la plus cohérente, car elle construit ces procédures en partant des propriétés requises plutôt que l'inverse, c'est-à-dire en vérifiant le bon comportement de procédures choisies sans principe. Le choix bayésien, tel qu'il est présenté dans ce livre, peut apparaître comme une réduction inutile de la portée du cadre inférentiel, et a été souvent critiqué pour cette raison. Mais nous verrons dans les chapitres suivants que cette réduction est à la fois nécessaire et bénéfique. Le Chapitre 11 résume

[2]Comme le signalent Robins et Wasserman (2000), il existe plusieurs définitions formelles de la cohérence, de Savage (1954) à Heath et Sudderth (1989), lesquels sont arrivés à la conclusion qu'une procédure est cohérente si et seulement si elle est bayésienne.

plusieurs points de vue défendant le choix bayésien qui peuvent être lus en perspective avec les arguments précédents[3].

Notons qu'il existe aussi une approche bayésienne de la statistique non paramétrique. Elle met généralement en œuvre des distributions a priori sur des espaces fonctionnels comme les processus de Dirichlet. Voir Ferguson (1973, 1974), Escobar (1989), Escobar et West (1995), Dey *et al.* (1998), et la Note 1.8.2 pour des références sur ce domaine. L'Exemple 1.23 donne une illustration de l'intérêt de l'approche bayésienne dans ce cadre.

1.2 Le paradigme bayésien et le principe de dualité

Comparée[4] à la modélisation probabiliste, l'analyse statistique se ramène fondamentalement à une *inversion*, car elle doit déterminer les causes–réduites aux paramètres du mécanisme probabiliste générateur–à partir des effets–résumés par les observations[5]. En d'autres termes, quand nous observons un phénomène aléatoire contrôlé par le paramètre θ, une méthode statistique permet de déduire de ces observations une *inférence* (c'est-à-dire, en résumé, une caractérisation) sur θ, alors que la modélisation probabiliste caractérise le comportement des observations futures *conditionnellement* à θ. Ce caractère d'inversion propre à la Statistique apparaît de façon évidente dans la notion de fonction de *vraisemblance*, car, d'un point de vue formel, il s'agit simplement d'une densité réécrite dans le bon ordre,

$$\ell(\theta|x) = f(x|\theta), \tag{1.3}$$

soit donc comme fonction de θ, qui est *inconnu*, dépendant de la valeur observée x. Historiquement l'*approche fiduciaire* de Fisher (1956) se fonde aussi sur cette inversion (voir la Note 1.8.1).

Une description générale de l'inversion des probabilités est donnée par le *théorème de Bayes* : Si A et E sont des événements tels que $P(E) \neq 0$, $P(A|E)$ et $P(E|A)$ sont reliés par

$$\begin{aligned} P(A|E) &= \frac{P(E|A)P(A)}{P(E|A)P(A) + P(E|A^c)P(A^c)} \\ &= \frac{P(E|A)P(A)}{P(E)} \ . \end{aligned}$$

[3]Ce chapitre et le Chapitre 11 méritent d'être relus une fois qu'on a bien compris les points les plus techniques du processus inférentiel bayésien et les problèmes qui s'y rattachent.

[4]Le mot *paradigme*, qui est un terme grammatical, est utilisé ici abusivement comme synonyme de *modèle* ou *principes*.

[5]À l'époque de Bayes et de Laplace, c'est-à-dire à la fin du XVIIIème siècle, la Statistique était souvent appelée *Probabilités inverses*, à cause de cette perspective. (Voir Stigler, 1986, Chapitre 3.)

En particulier,

$$\frac{P(A|E)}{P(B|E)} = \frac{P(E|A)}{P(E|B)} \ , \tag{1.4}$$

quand $P(B) = P(A)$. Obtenir ce résultat à partir des axiomes de la Théorie des Probabilités est trivial. Il s'agit cependant de l'étape conceptuelle la plus importante dans l'histoire de la Statistique, constituant la première *inversion* des probabilités. L'équation (1.4) exprime le fait fondamental que, pour deux causes équiprobables, le rapport des probabilités pour un effet donné est égal au rapport des probabilités de ces deux causes. Ce théorème est aussi un principe d'actualisation, car il décrit la mise à jour de la vraisemblance de A de $P(A)$ vers $P(A|E)$, une fois que E a été observé. Bayes (1763) donne en réalité une version continue de ces résultats, à savoir, pour deux variables aléatoires x et y, de distributions conditionnelle[6] $f(x|y)$ et marginale $g(y)$, la distribution conditionnelle de y sachant x est

$$g(y|x) = \frac{f(x|y)g(y)}{\int f(x|y)g(y)\,dy} \ .$$

Bien que ce théorème d'inversion soit naturel d'un point de vue probabiliste, Bayes et Laplace sont allés plus loin et ont considéré que l'*incertitude* sur le paramètre θ d'un modèle peut être décrite par une distribution de *probabilité* π sur Θ, appelée *distribution a priori*. L'inférence est alors fondée sur la distribution de θ conditionnelle à x, $\pi(\theta|x)$, appelée *distribution a posteriori* et définie par

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)\,d\theta} \ . \tag{1.5}$$

Notons que $\pi(\theta|x)$ est ainsi proportionnelle à la distribution de x conditionnellement à θ, qui est aussi la vraisemblance, multipliée par la distribution a priori de θ. (Il semble que la généralité de (1.5) n'ait pas été perçue par Bayes, mais par Laplace, qui la développera plus avant.) La contribution principale apportée par un modèle statistique bayésien est donc de considérer en sus une distribution aléatoire pour les paramètres.

Définition 1.8. *Un modèle statistique bayésien est constitué d'un modèle statistique paramétrique, $f(x|\theta)$, et d'une distribution a priori pour les paramètres, $\pi(\theta)$.*

En termes statistiques, le théorème de Bayes actualise donc l'information sur θ en extrayant l'information contenue dans l'observation x. Son impact

[6]Souvent nous remplacerons *distribution* par *densité*, supposant que plus tard le concept sera mieux défini par rapport à la mesure naturelle dominante, comme la mesure de Lebesgue. C'est seulement dans un contexte plus avancé, comme pour la mesure de Haar dans le Chapitre 6, qu'une connaissance plus approfondie de la théorie de la mesure sera nécessaire.

provient de la décision audacieuse de mettre causes et effets sur le même niveau conceptuel, puisque les deux sont aléatoires. Du point de vue de la modélisation statistique, il y a donc peu de différences entre observations et paramètres, car les manipulations conditionnelles permettent l'échange de leurs rôles respectifs. Notons que, historiquement, cette idée que les paramètres sont aléatoires peut être perçue comme allant à l'encontre du déterminisme athée de Laplace[7], ainsi que des conceptions religieuses de Bayes, qui était un ecclésiastique non-conformiste. En imposant cette modification fondamentale de la perception du phénomène aléatoire, ces deux mathématiciens ont créé l'analyse statistique moderne et, plus particulièrement, l'analyse bayésienne.

En effet, le recours à une distribution a priori π pour les paramètres d'un modèle est vraiment révolutionnaire. Elle représente de fait une avancée majeure, passant de la notion de paramètre *inconnu* à celle de paramètre *aléatoire*; de nombreux statisticiens tracent une frontière hermétique entre ces deux concepts, bien qu'ils acceptent une modélisation probabiliste des observations. Ils défendent ce point de vue sur la base que, même si dans certains cadres, le paramètre est obtenu sous l'action simultanée de plusieurs facteurs et peut ainsi apparaître comme (partiellement) aléatoire, comme par exemple en physique quantique, dans la plupart des cas il ne peut être perçu comme le résultat d'une expérience aléatoire. Un cas typique est l'estimation de quantités physiques comme la vitesse de la lumière c. Une réponse dans ce cas particulier est que la précision limitée des instruments de mesure implique que la vraie valeur de c ne sera jamais connue et justifie le fait de considérer c comme uniformément distribué sur $[c_0 - \epsilon, c_0 + \epsilon]$, où ϵ est la précision maximale des instruments de mesure et c_0 la valeur obtenue.

Nous considérons dans le Chapitre 3 différentes approches au problème délicat de détermination de la distribution a priori. Cependant, et plus fondamentalement, nous voulons insister sur le fait que l'importance de la distribution a priori dans l'analyse statistique bayésienne ne réside en aucun cas dans le fait que le paramètre d'intérêt θ puisse (ou ne puisse pas) être perçu comme étant distribué selon π, ou même comme étant une variable aléatoire, mais plutôt que l'utilisation de la distribution a priori est la meilleure façon de résumer l'information disponible (et le manque d'information) sur ce paramètre ainsi que l'incertitude résiduelle, et qu'elle permet de cette façon l'incorporation de cette information inexacte dans le processus de décision. (Un raisonnement similaire a conduit Laplace à développer des modèles statistiques, malgré son déterminisme.) Un point plus technique est que le seul moyen de construire une approche mathématiquement justifiée opérant conditionnellement aux observations est d'introduire une distribution correspondante pour les paramètres. Voir aussi Lindley (1990) pour une justification axiomatique détaillée sur l'utilisation des distributions a priori.

[7] *"Nous devons envisager l'état présent de l'Univers comme un effet de l'état antérieur et comme la cause de l'état suivant."* – Laplace (1795).

Nous terminons cette section par des exemples historiques de Bayes et de Laplace.

Exemple 1.9. (Bayes, 1763) Une boule de billard W roule sur une ligne de longueur un, avec une probabilité uniforme de s'arrêter n'importe où. Supposons qu'elle s'arrête en p. Une deuxième boule O roule alors n fois dans les mêmes conditions, et on note X le nombre de fois que la boule O s'arrête à gauche de W. *Connaissant X, quelle inférence pouvons-nous mener sur p ?*

Dans la terminologie moderne, le problème est de déterminer la distribution a posteriori de p conditionnellement à X, quand la distribution a priori de p est uniforme sur $[0,1]$ et $X \sim \mathcal{B}(n,p)$, variable aléatoire binomiale (voir l'Appendice A). Comme

$$P(X = x|p) = \binom{n}{x} p^x (1-p)^{n-x},$$
$$P(a < p < b \text{ et } X = x) = \int_a^b \binom{n}{x} p^x (1-p)^{n-x} dp$$

et

$$P(X = x) = \int_0^1 \binom{n}{x} p^x (1-p)^{n-x}\, dp,$$

nous trouvons que

$$P(a < p < b|X = x) = \frac{\int_a^b \binom{n}{x} p^x (1-p)^{n-x}\, dp}{\int_0^1 \binom{n}{x} p^x (1-p)^{n-x}\, dp} = \frac{\int_a^b p^x (1-p)^{n-x}\, dp}{B(x+1, n-x+1)},$$

donc que la distribution de p conditionnellement à $X = x$ est une distribution bêta, $\mathcal{B}e(x+1, n-x+1)$ (voir l'Appendice A). ‖

Dans le même esprit, Laplace introduit une modélisation probabiliste de l'espace des paramètres. Mais ses exemples sont plus avancés que ceux de Bayes au sens où les distributions a priori qu'il prend en compte sont fondées sur un raisonnement abstrait, plutôt que sur une justification physique[8].

Exemple 1.10. (Laplace, 1773) Une urne contient un nombre n de cartes noires et blanches. Si la première carte sortie de l'urne est blanche, *quelle est la probabilité que la proportion p de cartes blanches soit p_0 ?* Pour résoudre ce problème, Laplace suppose que tous les nombres de 2 à $n-1$ sont

[8] On peut aussi imaginer un Bayes plus machiavélique qui choisit cet exemple particulier afin de passer outre les critiques potentielles sur ce choix d'a priori. Mais il semble que ce ne soit pas le cas et qu'en réalité Bayes ait étudié cet exemple pour son intérêt propre. Voir Stigler (1986) pour plus de détails.

équiprobables comme valeurs de pn, donc que p soit uniformément distribué sur $\{2/n, \ldots, (n-1)/n\}$. La distribution a posteriori de p peut être alors calculée en utilisant le théorème de Bayes,

$$P(p = p_0 | \text{ données }) = \frac{p_0 \times 1/(n-2)}{\sum_{p=2/n}^{(n-1)/n} p \times 1/(n-2)} = \frac{n\, p_0}{n(n-1)/2 - 1}.$$

‖

Évidemment le choix précédent de la distribution a priori peut être contesté comme étant partiellement arbitraire. Cependant, dans la perspective de la théorie des probabilités de Laplace, la plupart des événements peuvent être décomposés en événements *équiprobables* élémentaires et par conséquent, dans ce cas particulier, il semble raisonnable de considérer les événements $\{p = i/n\}$ $(2 \leq i \leq n-1)$ comme élémentaires. Un raisonnement similaire justifie l'exemple suivant.

Exemple 1.11. (Laplace, 1786) Considérant la proportion de naissances masculines à Paris, Laplace veut vérifier que la probabilité x d'une naissance masculine dépasse $1/2$. Observant 251 527 naissances masculines et 241 945 naissances féminines en 1785 et supposant que x a pour distribution a priori la loi uniforme sur $[0, 1]$, Laplace obtient[9]

$$P(x \leq 1/2 | (251\,527; 241\,945)) = 1.15 \times 10^{-42}.$$

(Voir Stigler, 1986, p. 134 et l'Exercice 1.8.) Il déduit alors que cette probabilité x est très vraisemblablement supérieure à 50%. Utilisant de nouveau une distribution a priori uniforme, il compare aussi les naissances masculines à Londres et à Paris et en déduit que la probabilité d'une naissance masculine est aussi significativement supérieure à 50% en Angleterre. ‖

L'exemple suivant résolu par Laplace est plus intéressant encore car, d'un point de vue pratique, il propose une méthode pour obtenir une procédure optimale, et d'un point de vue théorique, il s'agit de la première construction formelle d'un estimateur de Bayes (détaillée dans le Chapitre 2).

Exemple 1.12. En astronomie, il est fréquent d'obtenir plusieurs observations d'une quantité ξ. Ces mesures sont distribuées indépendamment selon une distribution supposée unimodale et symétrique autour de ξ. Si nous assignons une distribution a priori uniforme au paramètre ξ, il devrait s'agir d'une "distribution uniforme sur $(-\infty, +\infty)$", qui n'est pas définie en tant que distribution de probabilité. Cependant, si nous acceptons cette extension

[9] Les nombres décimaux sont indiqués dans ce livre en notation anglo-saxone et non française.

formelle (voir la Section 1.5 pour une justification), nous pouvons travailler plutôt avec la mesure de Lebesgue sur $(-\infty, +\infty)$.

En utilisant cette *distribution généralisée*, Laplace (1773) a établi que la *médiane a posteriori* de ξ, c'est-à-dire la médiane de la distribution de ξ conditionnellement aux observations, est un estimateur optimal au sens où il minimise l'erreur moyenne absolue

$$\mathbb{E}^{\xi}[\,|\xi - \delta|\,] \tag{1.6}$$

en δ, où $\mathbb{E}^{\xi}[\cdot]$ est l'espérance sous la distribution de ξ (voir l'Appendice B pour une liste des notations usuelles). Ce résultat justifie l'utilisation de la médiane a posteriori comme un estimateur de ξ, quelle que soit la distribution de l'observation. Bien qu'établi il y a plus de deux siècles, ce résultat est incroyablement moderne (généralité de la distribution et choix de la fonction de perte pour évaluer les estimateurs) et Laplace l'a étendu en 1810 en établissant un résultat similaire pour l'erreur quadratique.

Curieusement, Laplace était plutôt déçu par ce résultat, parce qu'il avait encore besoin de la distribution de l'erreur d'observation pour calculer l'estimateur résultant. En 1774, il considéra la distribution double exponentielle

$$\varphi_{\xi}(x) = \frac{\xi}{2} e^{-\xi|x|}, \qquad x \in \mathbb{R},\ \xi > 0, \tag{1.7}$$

appelée aussi *distribution de Laplace*, qui impliquait en théorie la résolution d'une équation du quinzième degré pour trois observations. (En réalité Laplace a fait une erreur et l'équation correcte est cubique, comme le montre Stigler, 1986.) Puis, en 1777, il considéra l'alternative plus compliquée encore

$$\varphi_{\xi}(x) = \frac{1}{2\xi} \log\left(\xi/|x|\right) \mathbb{I}_{|x|\leq\xi}, \qquad \xi > 0,$$

où $\mathbb{I}$ est la fonction indicatrice. Ce fut seulement en 1810, lorsque Legendre et Gauss exposèrent de façon indépendante l'importance de la *distribution normale*, que Laplace fut capable de calculer ses estimateurs de Bayes explicitement, désormais persuadé qu'il s'agissait de la distribution d'erreur idéale (ou "normale"). ‖

Nous considérerons de nouveau cet exemple, ainsi que d'autres résultats d'optimalité, dans le Chapitre 2, lorsque nous étudierons les différentes fonctions de perte pour évaluer les procédures d'estimation et les estimateurs de Bayes associés. Nous insistons ici sur le fait que la conséquence principale des travaux de Bayes et de Laplace a été d'introduire le concept de *perspective conditionnelle* en Statistique, c'est-à-dire de s'être rendu compte que paramètres et observations sont fondamentalement des objets identiques, même s'ils sont perçus de façon différente[10]. Construire en parallèle une distribution de probabilité sur l'espace des paramètres complète cette équivalence,

[10]Encore une fois, c'est la raison pour laquelle ce livre note indistinctement variables aléatoires, observations et paramètres en minuscules.

grâce au Théorème de Bayes, et permet un discours quantitatif sur les causes, c'est-à-dire, dans notre cadre paramétrique, une inférence sur les paramètres. Comme nous l'avons déjà évoqué auparavant, le choix de la distribution a priori est délicat, mais sa détermination devrait être incluse dans le processus statistique, en parallèle à la détermination de la distribution de l'observation. Une distribution a priori est effectivement la meilleure façon d'inclure de l'information résiduelle dans un modèle. De plus, l'analyse statistique bayésienne fournit des outils naturels pour prendre en compte l'incertitude associée à l'information résiduelle dans le modèle (éventuellement via la modélisation hiérarchique, voir le Chapitre 10). Pour finir, comme souligné par Lindley (1971), le paradigme bayésien est intrinsèquement logique : pour un ensemble donné de propriétés requises, représentées par la fonction de perte et la distribution a priori, l'approche bayésienne fournit les estimateurs qui satisfont ces propriétés, alors que d'autres approches évaluent les propriétés d'estimateurs construits indépendamment du processus inférentiel.

1.3 Principes de vraisemblance et d'exhaustivité

1.3.1 Exhaustivité

La Statistique classique peut être décrite comme étant guidée par des principes souvent justifiés par le "bon sens" ou par des axiomes supplémentaires. L'approche bayésienne permet d'incorporer naturellement une majorité de ces principes sans imposer de restrictions supplémentaires sur les procédures de décision, et d'en rejeter d'autres de façon tout aussi systématique, comme la notion d'*estimation sans biais*. Cette notion était à une époque la pierre angulaire de la Statistique classique et limitait le choix des estimateurs à ceux corrects en moyenne (voir Lehmann et Casella, 1998). Bien qu'intuitivement acceptable, l'estimation sans biais impose des conditions trop strictes sur le choix des procédures et mène souvent à des solutions peu performantes. (Voir, par exemple, le cas de l'effet Stein décrit dans la Note 2.8.2.) Plus importants encore, les problèmes qui peuvent être résolus à travers l'estimation sans biais représentent un pourcentage infime de l'ensemble des problèmes d'estimation (Exercice 1.17). Malgré ces inconvénients, une technique statistique récente appelée *bootstrap* (Efron, 1982, Hall, 1992) a été présentée pour réduire le biais (asymptotiquement).

Deux principes fondamentaux sont respectés par le paradigme bayésien : le principe de vraisemblance et le principe d'exhaustivité.

Définition 1.13. *Quand $x \sim f(x|\theta)$, une fonction T de x (aussi appelée statistique) est* exhaustive *si la distribution de x conditionnellement à $T(x)$ ne dépend pas de θ.*

Une statistique exhaustive $T(x)$ contient toute l'information apportée par x sur θ. Selon le *théorème de factorisation*, sous certaines conditions de régularité (voir Lehmann et Casella, 1998), la densité de x s'écrit alors

$$f(x|\theta) = g(T(x)|\theta)h(x|T(x)),$$

si g est la densité de $T(x)$. Nous verrons dans le Chapitre 2 que, quand un estimateur est évalué sous un coût convexe, la procédure optimale dépend uniquement de la statistique exhaustive (théorème de Rao-Blackwell). En particulier, quand le modèle admet une statistique *exhaustive minimale* (c'est-à-dire fonction de toute autre statistique exhaustive), nous devons ne considérer que les procédures dépendant de cette statistique ou, de façon équivalente, du modèle statistique restreint à cette statistique. Le concept d'exhaustivité a été développé par Fisher et conduit au principe suivant.

Principe d'exhaustivité *Deux observations x et y donnant la même valeur d'une statistique exhaustive T, c'est-à-dire telles que $T(x) = T(y)$, doivent conduire à la même inférence sur θ.*

Exemple 1.14. Soient $x_1, \ldots, x_n$ des observations indépendantes d'une distribution normale $\mathcal{N}(\mu, \sigma^2)$ (voir l'Appendice A). Le théorème de factorisation implique alors que le couple $T(x) = (\bar{x}, s^2)$, où

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{et} \quad s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2,$$

forme une statistique exhaustive pour le paramètre (μ, σ), de densité

$$g(T(x)|\theta) = \sqrt{\frac{n}{2\pi\sigma^2}}e^{-(\bar{x}-\theta)^2 n/2\sigma^2}\,\frac{(s^2)^{(n-3)/2}e^{-s^2/2\sigma^2}}{\sigma^n \Gamma(n-1/2)2^{n-1/2}}.$$

Par conséquent, suivant le principe d'exhaustivité, l'inférence sur μ ne devrait dépendre que de ce vecteur bidimensionnel, quelle que soit la taille de l'échantillon n. Nous verrons dans le Chapitre 3 que l'existence d'une statistique exhaustive de dimension constante est caractéristique des *familles exponentielles*[11]. ‖

Exemple 1.15. Soient $x_1 \sim \mathscr{B}(n_1, p)$, $x_2 \sim \mathscr{B}(n_2, p)$, et $x_3 \sim \mathscr{B}(n_3, p)$, trois observations binomiales indépendantes où les tailles des échantillons n_1, n_2 et n_3 sont connues. La fonction de vraisemblance est alors

[11] Pour les autres distributions, l'exhaustivité n'est pas un concept intéressant car la dimension de la statistique exhaustive est alors de l'ordre de la dimension de l'observation x (ou de l'échantillon correspondant), comme expliqué dans le Chapitre 3.

$$f(x_1, x_2, x_3|p) = \binom{n_1}{x_1}\binom{n_2}{x_2}\binom{n_3}{x_3} p^{x_1+x_2+x_3}(1-p)^{n_1+n_2+n_3-x_1-x_2-x_3}$$

et les statistiques

$$T_1(x_1, x_2, x_3) = x_1 + x_2 + x_3 \qquad \text{ou} \qquad T_2(x_1, x_2, x_3) = \frac{x_1 + x_2 + x_3}{n_1 + n_2 + n_3}$$

sont exhaustives, contrairement à, par exemple, $x_1/n_1 + x_2/n_2 + x_3/n_3$. ‖

Le principe d'exhaustivité est généralement accepté par la plupart des statisticiens, en particulier à cause du théorème de Rao-Blackwell, qui écarte tout estimateur ne dépendant pas uniquement de statistiques exhaustives. Dans un cadre de *choix de modèle*, ce principe est parfois critiqué, pour être trop réducteur. Soulignons cependant que le principe d'exhaustivité n'est légitime que lorsque les observations sont véritablement générées par le modèle statistique considéré. Toute incertitude sur la distribution des observations devrait être incorporée dans le modèle, une modification qui conduirait certainement à un changement des statistiques exhaustives. La même remarque s'applique d'ailleurs au principe de vraisemblance.

1.3.2 Principe de vraisemblance

Ce deuxième principe est en partie une conséquence du principe d'exhaustivité. Il peut être attribué à Fisher (1959) ou même à Barnard (1949), mais il a été formalisé par Birnbaum (1962). Il est fortement défendu par Berger et Wolpert (1988) qui ont fourni une étude approfondie du sujet. Dans la définition suivante, la notion d'*information* doit être considérée au sens large et non dans le sens mathématique d'information avancé par Fisher, définie au Chapitre 3. Elle désigne, de façon générale, l'ensemble des inférences possibles sur θ.

> **Principe de vraisemblance** *L'information apportée par une observation de x sur θ est entièrement contenue dans la fonction de vraisemblance $\ell(\theta|x)$. De plus, si x_1 et x_2 sont deux observations qui dépendent du même paramètre θ, et telles qu'il existe une constante c satisfaisant*
>
> $$\ell_1(\theta|x_1) = c\ell_2(\theta|x_2)$$
>
> *pour tout θ, elles apportent la même information sur θ et doivent conduire à la même inférence.*

Notons que le principe de vraisemblance n'est valide que lorsque

(i) l'inférence concerne le *même* paramètre θ ; et

(ii) θ prend en compte *tous* les facteurs inconnus du modèle.

L'exemple suivant donne une illustration devenue "classique" de ce principe.

Exemple 1.16. Soit l'étude de taux d'audience d'une émission de télévision, $0 \leq \theta \leq 1$ représentant la part d'audience. Un enquêteur a trouvé neuf téléspectateurs et trois personnes n'ayant pas vu l'émission. Si nous ne disposons pas de plus d'information, au moins deux modèles probabilistes peuvent être envisagés :

(1) l'enquêteur a interrogé 12 personnes, et a donc observé $x \sim \mathcal{B}(12, \theta)$ avec $x = 9$;
(2) l'enquêteur a interrogé N personnes jusqu'à obtenir 3 non téléspectateurs, avec $N \sim \mathcal{N}eg(3, 1-\theta)$ et $N = 12$.

En d'autres termes, la quantité aléatoire dans cette étude peut être soit 9, soit 12. (Notons qu'elles pourraient aussi être toutes deux aléatoires.) Le point à souligner est que, pour les deux modèles, la vraisemblance est proportionnelle à

$$\theta^3(1-\theta)^9.$$

Par conséquent, le principe de vraisemblance affirme que l'inférence sur θ devrait être la même pour les deux modèles. Comme on verra dans l'Exercice 1.29, ceci n'est pas le cas dans l'approche classique. ‖

Puisque l'approche bayésienne est entièrement fondée sur la distribution a posteriori

$$\pi(\theta|x) = \frac{\ell(\theta|x)\pi(\theta)}{\int \ell(\theta|x)\pi(\theta)d\theta}$$

(voir équation (1.5) et la Section 1.4), qui ne dépend de x qu'à travers $\ell(\theta|x)$, le principe de vraisemblance est automatiquement satisfait dans un cadre bayésien.

Au contraire, l'approche classique ou *fréquentiste*[12] est fondée sur des propriétés de comportement *moyen* des procédures et justifie donc l'utilisation d'un estimateur pour des raisons qui peuvent contredire le principe de vraisemblance. Cette perspective est particulièrement frappante en théorie des tests, traitée au Chapitre 5. Par exemple, si $x \sim \mathcal{N}(\theta, 1)$ et si nous cherchons à vérifier l'hypothèse $H_0 : \theta = 0$, la procédure de test classique de Neyman-Pearson au seuil 5% rejettera l'hypothèse si $x = 1.96$, sur la base que $P(|x - \theta| \geq 1.96) = 0.05$, donc conditionné par l'événement $|x| > 1.96$ plutôt que par $x = 1.96$ (ce qui est impossible pour la théorie fréquentiste). L'argument fréquentiste associé à cette procédure est alors que, dans 5% des cas où H_0 est vrai, l'hypothèse nulle est rejetée à tort. De tels arguments contredisent le principe de vraisemblance, car les comportements des queues

[12]La théorie avancée par Wald, Neymann et Pearson dans les années 50 est dite *fréquentiste*, car elle évalue les procédures par rapport à leurs performances sur le long terme, c'est-à-dire en moyenne (ou en *fréquence*) plutôt que de se concentrer sur la performance de la procédure pour l'observation obtenue, comme le ferait une approche conditionnelle. L'approche fréquentiste sera abordée en détail dans les Chapitres 2 et 5.

de distributions peuvent varier pour les mêmes vraisemblances (voir les Exercices 1.24 et 1.29). L'opposition entre paradigmes fréquentiste et bayésien est plus forte en théorie des tests que pour l'estimation ponctuelle, où l'approche fréquentiste apparaît souvent comme un cas limite de l'approche bayésienne (voir le Chapitre 5).

Exemple 1.17. Soient x_1, x_2 i.i.d. $\mathcal{N}(\theta, 1)$. La fonction de vraisemblance est alors

$$\ell(\theta|x_1, x_2) \propto \exp\{-(\bar{x} - \theta)^2\}$$

avec $\bar{x} = (x_1 + x_2)/2$. Soit maintenant la distribution alternative

$$g(x_1, x_2|\theta) = \pi^{-3/2} \frac{e^{-(x_1+x_2-2\theta)^2/4}}{1 + (x_1 - x_2)^2}.$$

Cette distribution donne une fonction de vraisemblance proportionnelle à $\ell(\theta|x_1, x_2)$ et par conséquent devrait conduire à la même inférence sur θ. Cependant, la distribution g est tout à fait différente de $f(x_1, x_2|\theta)$; par exemple, l'espérance de $(x_1 - x_2)$ n'est pas définie. Les estimateurs de θ auront donc des propriétés fréquentistes différentes s'ils ne dépendent pas que de $\bar{x}$. En particulier, les régions de confiance pour θ peuvent différer significativement, à cause des queues plus épaisses de g. ‖

Exemple 1.18. Une autre implication du principe de vraisemblance est le *principe des règles d'arrêt* en analyse séquentielle. Une *règle d'arrêt* τ peut être définie comme suit : si les expériences $\mathcal{E}_i$ produisent des observations $x_i \in \mathcal{X}_i$, avec $x_i \sim f(x_i|\theta)$, considérons la suite correspondante $\mathcal{A}_i \subset \mathcal{X}_1 \times \ldots \times \mathcal{X}_i$ telle que le critère τ prend la valeur n si $(x_1, \ldots, x_n) \in \mathcal{A}_n$, i.e., l'expérience s'arrête après la n-ième observation seulement si les n premières observations sont en $\mathcal{A}_n$. La vraisemblance de $(x_1, \ldots, x_n)$ est alors

$$\ell(\theta|x_1, \ldots, x_n) = f(x_1|\theta) f(x_2|x_1, \theta) \\ \ldots f(x_n|x_1, \ldots, x_{n-1}, \theta) \mathbb{I}_{\mathcal{A}_n}(x_1, \ldots, x_n),$$

et donc dépend seulement de τ via l'échantillon $x_1, \ldots, x_n$. Ceci implique le principe suivant.

> **Principe des règles d'arrêt** *Si une suite d'expériences, $\mathcal{E}_1, \mathcal{E}_2, \ldots$, admet une règle d'arrêt, τ, qui indique quand doivent s'arrêter les expériences, l'inférence sur θ ne doit dépendre de τ qu'à travers l'échantillon résultant.*

L'Exemple 1.16 illustre le cas de deux critères d'arrêt différents qui conduisent au même échantillon : ou bien on fixe la taille de l'échantillon à douze, ou bien l'expérience s'arrête quand on a obtenu neuf réponses positives. Un autre exemple frappant (même s'il est artificiel) de règle d'arrêt

consiste à observer des $x_i \sim \mathcal{N}(\theta, 1)$ et à prendre τ comme le premier entier n tel que

$$|\bar{x}_n| = \left| \sum_{i=1}^{n} x_i/n \right| > 1.96/\sqrt{n}.$$

Dans ce cas, la règle d'arrêt est évidemment incompatible avec la modélisation fréquentiste, parce que avec un tel échantillon on rejettera *toujours* l'hypothèse nulle $H_0 : \theta = 0$ au seuil de 5% (voir le Chapitre 5). En revanche, une approche bayésienne évite cette difficulté (voir Raiffa et Schlaifer, 1961 et Berger et Wolpert, 1988, p. 81). ‖

1.3.3 Dérivation du principe de vraisemblance

Une justification du principe de vraisemblance a été avancée par Birnbaum (1962) qui a établi que le principe de vraisemblance est une conséquence du principe d'exhaustivité, à condition d'accepter un second principe.

> **Principe de conditionnement** *Si deux expériences sur le paramètre θ, notées $\mathcal{E}_1$ et $\mathcal{E}_2$, sont possibles et si on choisit une de ces expériences avec probabilité p, l'inférence sur θ ne doit dépendre que de l'expérience choisie.*

Il semble difficile de refuser ce principe quand l'expérience choisie est connue, comme on peut le constater dans l'exemple (classique) suivant.

Exemple 1.19. (Cox, 1958) Dans un laboratoire de recherche, une quantité physique θ doit être mesurée par un appareil efficace, mais très souvent utilisé, qui donne une mesure $x_1 \sim \mathcal{N}(\theta, 0.1)$, avec une probabilité $p = 0.5$, ou grâce à un autre appareil, moins précis mais plus disponible, qui donne $x_2 \sim \mathcal{N}(\theta, 10)$. L'appareil a été choisi au hasard selon la disponibilité de l'appareil le plus précis. L'inférence sur θ ne devrait donc pas dépendre du fait que le second appareil *aurait pu être choisi*. Cependant, un intervalle de confiance classique au seuil 5% prenant en compte cette sélection, soit donc moyennant entre toutes les expériences possibles, est de demi-longueur 5.19, tandis que l'intervalle associé à $\mathcal{E}_1$ est de demi-longueur 0.62 (Exercice 1.26). ‖

Le résultat équivalent de Birnbaum (1962) est alors le suivant.

Théorème 1.20. *Le principe de vraisemblance est équivalent à la conjonction des principes d'exhaustivité et de conditionnement.*

Preuve. Définissons d'abord l'*évidence* associée à une expérience $\mathcal{E}$, $Ev(\mathcal{E}, x)$, comme l'ensemble des inférences possibles sur le paramètre θ pour cette expérience. Soit $\mathcal{E}^*$ l'expérience *mixte* correspondant à $\mathcal{E}_i$ avec probabilité 0.5

$(i = 1, 2)$, qui a donc comme résultat (i, x_i). Sous ces notations, le principe de conditionnement peut être énoncé ainsi : pour tout $j = 1, 2$,

$$Ev(\mathscr{E}^*, (j, x_j)) = Ev(\mathscr{E}_j, x_j)\,. \tag{1.8}$$

Soient x_1^0 et x_2^0 tels que

$$\ell(\cdot|x_1^0) = c\ell(\cdot|x_2^0). \tag{1.9}$$

Le principe de vraisemblance est alors équivalent à

$$Ev(\mathscr{E}_1, x_1^0) = Ev(\mathscr{E}_2, x_2^0)\,. \tag{1.10}$$

Supposons que (1.9) est vérifiée. Pour l'expérience mixte $\mathscr{E}^*$ construite à partir des deux expériences initiales, considérons la statistique

$$T(j, x_j) = \begin{cases} (1, x_1^0) & \text{si } j = 2,\ x_2 = x_2^0, \\ (j, x_j) & \text{sinon,} \end{cases}$$

qui prend la même valeur pour $(1, x_1^0)$ et pour $(2, x_2^0)$. Alors, cette statistique est exhaustive puisque, si $t \neq (1, x_1^0)$,

$$P_\theta(X^* = (j, x_j)|T = t) = \mathbb{I}_t(j, x_j)$$

et

$$P_\theta(X^* = (1, x_1^0)|T = (1, x_1^0)) = \frac{c}{1 + c},$$

de par la proportionnalité des fonctions de vraisemblance. Le principe d'exhaustivité implique alors que

$$Ev(\mathscr{E}^*, (1, x_1)) = Ev(\mathscr{E}^*, (2, x_2)) \tag{1.11}$$

et, combiné avec (1.8), donne (1.10), soit donc le principe de vraisemblance.

La réciproque de ce théorème se déduit du principe de vraisemblance, du fait que les fonctions de vraisemblance de (j, x_j) et de x_j sont proportionnelles et, pour le principe d'exhaustivité, du théorème de factorisation. □

Evans *et al.* (1986) ont démontré que le principe de vraisemblance peut être aussi obtenu comme une conséquence d'une version plus forte du principe de conditionnement.

1.3.4 Mise en œuvre du principe de vraisemblance

Il paraît donc tout à fait justifié de suivre le principe de vraisemblance, puisque celui-ci s'obtient à partir des principes irréfutables d'exhaustivité et de conditionnement. Cependant, ce principe est, somme toute, assez vague, puisqu'il ne mène pas à la sélection d'une procédure particulière pour un problème inférentiel donné. D'aucuns ont soutenu que le rôle du statisticien devrait s'arrêter à la détermination de la fonction de vraisemblance (Box et

Tiao, 1973) puisqu'elle suffit au client pour mener l'inférence, mais ce point de vue extrême n'est tenable que dans les cas les plus simples (ou d'un point de vue bayésien décisionnel, si le preneur de décision fournit aussi une distribution a priori et une fonction de perte). Pour de grandes dimensions (du paramètre), la fonction de vraisemblance est aussi difficile à manipuler à cause du manque d'outils de représentation adéquats.

Le caractère vague du principe de vraisemblance exige un renforcement des bases axiomatiques du processus inférentiel, c'est-à-dire l'ajout de structures dans la construction des procédures statistiques. Par exemple, une mise en œuvre efficace du principe de vraisemblance est l'estimateur du maximum de vraisemblance, comme décrit brièvement en Section 1.3.5. De façon similaire, le paradigme bayésien permet la mise en œuvre pratique du principe de vraisemblance, avec comme avantage supplémentaire la prise en compte des exigences décisionnelles du problème inférentiel, et même l'obtention de procédures optimales d'un point de vue fréquentiste (voir plus bas).

Si nous gardons à l'esprit l'aspect d'inversion de la Statistique présenté en Section 1.2, il est tentant de considérer la vraisemblance comme une densité généralisée en θ, dont le mode serait alors l'estimateur du maximum de vraisemblance, et de travailler avec cette densité comme une distribution ordinaire. Cette approche semble avoir été soutenue par Laplace qui proposait d'utiliser une distribution a priori uniforme lorsque aucune information n'était disponible sur θ (voir les Exemples 1.9-1.12). De même, Fisher introduisit l'approche fiduciaire (voir la Note 1.8.1) pour tenter de circonvenir la détermination de la distribution a priori lors de la mise en pratique du principe de vraisemblance, le choix de cette distribution étant subjectif (puisque ne dépendant que de la distribution des observations). Cependant, cette approche est surtout défendable quand θ est un paramètre de position (voir aussi l'Exemple 1.25), puisqu'il entraîne en général des paradoxes et des contradictions. L'exemple le plus frappant est le fait que $\ell(\theta|x)$ n'est pas nécessairement intégrable comme fonction de θ (Exercice 1.25). L'obtention de distributions a posteriori objectives exige en fait une théorie plus avancée des distributions non informatives (voir le Chapitre 3), qui montre que la fonction de vraisemblance ne peut pas toujours être considérée comme la distribution a posteriori la plus naturelle.

Beaucoup d'approches ont été proposées pour mettre en œuvre le principe de vraisemblance, comme par exemple la théorie de la *vraisemblance pénalisée* (Akaike, 1978, 1983) ou la théorie de la *complexité stochastique* (Rissanen, 1983, 1990). Voir aussi Bjørnstad (1990) pour une revue des méthodes non bayésiennes fondées sur le principe de vraisemblance dans le domaine de la prévision. La conclusion générale de cette section est que, malgré tout, mis à part le fait que plusieurs de ces théories ont une teneur bayésienne, une approche véritablement bayésienne est la plus adéquate pour tirer parti du principe de vraisemblance. (Voir Berger et Wolpert, 1988, Chapitre 5, pour une discussion approfondie sur ce point.)

1.3.5 Estimation par maximum de vraisemblance

Le principe de vraisemblance est en soi distinct de l'approche de *l'estimation par maximum de vraisemblance*, qui n'est qu'une façon parmi d'autres de mettre en œuvre ce principe. Puisque nous rencontrerons assez souvent cette technique dans les prochains chapitres, et qu'elle se situe à la lisière du paradigme bayésien, nous rappelons brièvement quelques faits élémentaires concernant le maximum de vraisemblance. Un traitement plus étendu peut être trouvé dans Lehmann et Casella (1998).

Lorsqu'on observe $x \sim f(x|\theta)$, l'approche par maximum de vraisemblance considère l'estimateur suivant de θ,

$$\hat{\theta} = \arg\sup_{\theta} \ell(\theta|x), \tag{1.12}$$

qui est donc la valeur de θ qui maximise la densité en x, $f(x|\theta)$, ou, exprimé de manière informelle, la probabilité d'observer la valeur donnée x. La maximisation (1.12) n'est pas toujours possible (voir, par exemple, le cas d'un mélange de deux distributions normales, détaillé au Chapitre 6), ou bien elle peut mener à plusieurs maxima globaux équivalents (voir notamment le cas d'une loi de Cauchy, $\mathscr{C}(0,1)$, avec deux observations bien séparées). Cependant, l'estimation par maximum de vraisemblance est largement utilisée, à cause d'une part de la motivation intuitive de maximiser la probabilité d'occurrence et d'autre part de ses propriétés asymptotiques fortes (*convergence* et *efficacité*). Une autre caractéristique intéressante de l'estimateur du maximum de vraisemblance est son invariance par reparamétrisation. En effet, pour toute fonction $h(\theta)$, l'estimateur de maximum de vraisemblance est $h(\hat{\theta})$ (même quand h n'est pas bijective). Cette propriété n'est partagée par aucune autre approche statistique (mis à part les estimateurs bayésiens dans le cas particulier des fonctions de coût intrinsèques, voir la Section 2.5.4.)

La méthode du maximum de vraisemblance a aussi ses défauts. Premièrement, la maximisation de $\ell(\theta|x)$ peut être assez complexe en pratique, particulièrement dans les cas multidimensionnels ou contraints. Prenons les exemples d'un mélange de distributions normales, d'une distribution de Weibull tronquée

$$\ell(\theta_1,\theta_2|x_1,\ldots,x_n) = (\theta_1\theta_2)^n (x_1\ldots x_n)^{\theta_1} \exp\left\{-\theta_2 \sum_{i=1}^{n} x_i^{\theta_1}\right\}$$

(voir l'Exercice 1.28), ou d'une table 10×10 où $x_{ij} \sim \mathscr{N}(\theta_{ij}, 1)$ et θ_{ij} croît en i et j (voir Robert et Hwang, 1996, et les Exercices 1.29 et 1.30). Certaines procédures numériques, comme l'algorithme EM de Dempster *et al.* (1977), pour des modèles à données manquantes, ou l'algorithme de Robertson *et al.* (1988) pour des espaces paramétriques restreints par ordre, ont été adaptées à cette approche, mais des problèmes non résolus demeurent (MacLachlan et Krishnan, 1997, Robert et Casella, 2004).

Deuxièmement, une technique de maximisation donne forcément des estimateurs peu lisses, par opposition à l'intégration par exemple. Cela est particulièrement vrai lorsque l'espace des paramètres est restreint. Par exemple Saxena et Alam (1982) montrent que, si $x \sim \chi^2_p(\lambda)$, loi du khi deux décentré à p degrés de liberté[13], l'estimateur du maximum de vraisemblance de λ est égal à 0 pour $x < p$. De même, les estimateurs du maximum de vraisemblance peuvent être numériquement instables, c'est-à-dire peuvent varier considérablement pour de petites variations des observations, du moins pour des tailles d'échantillon réduites (Exercice 1.31).

Un dernier défaut, mais non des moindres, de l'approche du maximum de vraisemblance est qu'elle n'admet pas de justifications probabiliste et décisionnelle. De fait, elle ne répond pas aux exigences d'une analyse décisionnelle et échoue ainsi à fournir des outils d'évaluation pour les estimateurs qu'elle propose. Par exemple, il n'est pas possible de faire des tests dans un contexte de maximum de vraisemblance pur : il est nécessaire de recourir à des justifications fréquentistes, même pour des tests du rapport de vraisemblance (voir la Section 5.3).

De même, les régions de confiance de la forme $C = \{\theta;\ \ell(\theta)/\ell(\hat{\theta}) \geq c\}$, qui sont les plus petites asymptotiquement, ne dépendront pas uniquement de la fonction de vraisemblance si la borne c doit être choisie de manière à obtenir un niveau de confiance α.

1.4 Distributions a priori et a posteriori

Supposons désormais que, en plus d'une distribution d'échantillonnage, $f(x|\theta)$, une distribution a priori sur θ, $\pi(\theta)$, soit disponible, c'est-à-dire que nous disposions d'un modèle complètement bayésien. Le Chapitre 3 traite du problème préliminaire d'obtention de cette distribution à partir de l'information a priori. Une fois données ces deux distributions, nous pouvons en construire plusieurs autres, à savoir :

(a) la *distribution jointe* de (θ, x),

$$\varphi(\theta, x) = f(x|\theta)\pi(\theta)\,;$$

(b) la *distribution marginale* de x,

$$\begin{aligned} m(x) &= \int \varphi(\theta, x)\, d\theta \\ &= \int f(x|\theta)\pi(\theta)\, d\theta\,; \end{aligned}$$

[13] Cet exemple montre aussi la limite de l'invariance mentionnée ci-dessus : lorsque $y \sim \mathcal{N}_p(\theta, I_p)$, l'estimateur maximum de vraisemblance de $\lambda = ||\theta||^2$ est $||y||^2 = x \sim \chi^2_p(\lambda)$, qui diffère de l'estimateur du maximum de vraisemblance fondé sur x (voir l'Exercice 3.56).

(c) la *distribution a posteriori* de θ, obtenue par la formule de Bayes,

$$\begin{aligned}\pi(\theta|x) &= \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)\,d\theta} \\ &= \frac{f(x|\theta)\pi(\theta)}{m(x)}\,;\end{aligned}$$

(d) la *distribution prédictive* de y, où $y \sim g(y|\theta, x)$, obtenue par

$$g(y|x) = \int g(y|\theta, x)\pi(\theta|x)d\theta\,.$$

Exemple 1.21. (Suite de l'Exemple 1.9) Si $x \sim \mathcal{B}(n, p)$ et $p \sim \mathcal{B}e(\alpha, \beta)$ (avec $\alpha = \beta = 1$ dans le cas particulier de Bayes),

$$\begin{aligned}f(x|p) &= \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, ..., n, \\ \pi(p) &= \frac{1}{B(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1}, \quad 0 \le p \le 1.\end{aligned}$$

La distribution jointe de (x, p) est alors

$$\varphi(x, p) = \frac{\binom{n}{x}}{B(\alpha, \beta)} p^{\alpha+x-1}(1-p)^{n-x+\beta-1}$$

et la distribution marginale de x est

$$\begin{aligned}m(x) &= \frac{\binom{n}{x}}{B(\alpha, \beta)} B(\alpha + x, n - x + \beta) \\ &= \binom{n}{x} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha + x)\Gamma(n - x + \beta)}{\Gamma(\alpha + \beta + n)},\end{aligned}$$

puisque la distribution a posteriori de p est

$$\pi(p|x) = \frac{p^{\alpha+x-1}(1-p)^{\beta+n-x-1}}{B(\alpha + x, \beta + n - x)},$$

qui est une loi bêta $\mathscr{B}e(\alpha + x, \beta + n - x)$. ‖

Parmi ces distributions, le concept fondamental du paradigme bayésien est la *distribution a posteriori*. En effet, cette distribution opère de façon conditionnelle sur les observations, et met donc en œuvre automatiquement l'*inversion* des probabilités définie dans la Section 1.2, tout en incluant les exigences du principe de vraisemblance. On évite ainsi de moyenner sur des valeurs de x non observées, ce qui est l'essence de l'approche fréquentiste. La distribution a posteriori représente l'actualisation de l'information disponible

sur θ, au vu de l'information contenue dans $\ell(\theta|x)$, tandis que $\pi(\theta)$ représente l'information disponible a priori, c'est-à-dire préalable à l'observation de x.

Notons que l'approche bayésienne jouit d'un type spécifique de cohérence (nous devrions en voir d'autres exemples dans les chapitres suivants) en ce que l'ordre suivant lequel des observations i.i.d. sont collectées n'a pas d'importance (il s'agit d'une conséquence du principe de vraisemblance), mais aussi que mettre à jour l'a priori une observation après l'autre, ou toutes les observations d'un coup, revient au même. En d'autres termes,

$$\begin{aligned}\pi(\theta|x_1,\ldots,x_n) &= \frac{f(x_n|\theta)\pi(\theta|x_1,\ldots,x_{n-1})}{\int f(x_n|\theta)\pi(\theta|x_1,\ldots,x_{n-1})d\theta}\\ &= \frac{f(x_n|\theta)f(x_{n-1}|\theta)\pi(\theta|x_1,\ldots,x_{n-2})}{\int f(x_n|\theta)f(x_{n-1}|\theta)\pi(\theta|x_1,\ldots,x_{n-2})d\theta}\\ &= \ldots\\ &= \frac{f(x_n|\theta)f(x_{n-1}|\theta)\ldots f(x_1|\theta)\pi(\theta)}{\int f(x_n|\theta)f(x_{n-1}|\theta)\ldots f(x_1|\theta)\pi(\theta)d\theta}\,.\end{aligned} \tag{1.13}$$

Il peut arriver que les observations ne modifient pas les distributions de certains paramètres. C'est évidemment le cas quand la loi de x ne dépend pas de ces paramètres, comme dans certains cas non identifiables.

Exemple 1.22. Considérons une observation x d'une distribution

$$\mathcal{N}\left(\frac{\theta_1+\theta_2}{2},1\right)$$

avec un a priori π sur (θ_1,θ_2) tel que $\pi(\theta_1,\theta_2)=\pi_1(\theta_1+\theta_2)\pi_2(\theta_1-\theta_2)$. Si nous réalisons le changement de variables

$$\xi_1=\frac{\theta_1+\theta_2}{2},\qquad \xi_2=\frac{\theta_1-\theta_2}{2}\,,$$

la distribution a posteriori de ξ_2 est alors

$$\begin{aligned}\pi(\xi_2) &\propto \int_{\mathbb{R}} \exp\left\{-(x-\xi_1)^2/2\right\} 2\pi_1(2\xi_1)2\pi_2(2\xi_2)d\xi_1\\ &\propto \pi_2(2\xi_2)\int_{\mathbb{R}} \exp\left\{-(x-\xi_1)^2/2\right\}\pi_1(2\xi_1)d\xi_1\\ &\propto \pi_2(2\xi_2)\end{aligned}$$

pour chaque observation x. L'observation n'apporte donc pas d'information sur ξ_2. ‖

Nous devons avertir le lecteur ou la lectrice[14] que tous les cas non identifiables ne mènent pas à cette conclusion simple : suivant le choix de la

[14]Dans la suite de l'ouvrage, le fait que le lectorat de cet ouvrage est mixte sera pris en compte de manière implicite par un pluriel neutre afin de ne pas surcharger le style.

distribution a priori et de la reparamétrisation du paramètre θ en (θ_1, θ_2), où la distribution de x ne dépend que de θ_1, la distribution marginale a posteriori de θ_2 peut dépendre ou non de x (Exercice 1.44). Un aspect important du paradigme bayésien dans un cadre non identifiable est cependant que la distribution a priori peut être utilisée comme un outil pour *identifier* les composantes du paramètre qui ne sont pas couvertes par la vraisemblance, même si un tel choix d'a priori peut avoir un impact sur la partie identifiable.

Cette invariance entre distributions a priori et distributions a posteriori peut aussi affecter certains paramètres quand le nombre de ceux-ci devient trop important par rapport à la taille de l'échantillon (Exercice 1.38).

Exemple 1.23. Une telle situation a lieu lorsque le nombre de paramètres est infini, par exemple quand l'inférence concerne une distribution entière. Dette et Studden (1997) considèrent n observations $x_1, \ldots, x_n$ provenant d'un *mélange de distributions géométriques*,

$$x \sim \int_0^1 \theta^x (1-\theta)\, dG(\theta),$$

x prenant ses valeurs dans $\mathbb{N}$ et la distribution probabiliste G étant inconnue. Dans ce cadre, G peut être représenté par la suite de ses moments non centrés $c_1, c_2, \ldots$ La fonction de vraisemblance est alors obtenue à partir de $P(X = k) = c_k - c_{k+1}$. Dette et Studden (1997) montrent (Exercice 1.45) que, bien que les c_i soient liés par un nombre infini d'inégalités (commençant par $c_1 > c_2 > c_1^2$), il est possible de construire de façon analytique des fonctions indépendantes entre elles des c_i, $p_1, p_2, \ldots$, prenant leurs valeurs dans $[0, 1]$ et telles que c_i ne dépende que de $(p_1, \ldots, p_i)$ (voir l'Exercice 1.45 pour les détails). Par conséquent, si la distribution a priori de $(p_1, p_2, \ldots)$ est

$$\pi(p_1, p_2, \ldots) = \prod_{i=1}^{+\infty} \pi_i(p_i)$$

et si la plus grande observation dans l'échantillon est k, la distribution a posteriori de $(p_{k+2}, p_{k+3}, \ldots)$ ne dépend pas des observations :

$$\pi(p_{k+2}, \ldots | x_1, \ldots, x_n) = \pi(p_{k+2}, \ldots) = \prod_{i=k+2}^{+\infty} \pi_i(p_i). \qquad \|$$

À l'inverse, la distribution marginale ne fait pas intervenir le paramètre d'intérêt θ. Il est donc rare de s'en servir directement, sauf dans l'*approche bayésienne empirique* (voir le Chapitre 10), car la distribution a posteriori est beaucoup mieux adaptée aux objectifs inférentiels. La distribution marginale peut cependant être utilisée pour construire la distribution a priori, si l'information disponible a été obtenue à partir de différentes expériences,

c'est-à-dire lorsqu'on traite différents θ dans une *méta analyse* (voir Mosteller et Chalmers, 1992, Mengersen et Tweedie, 1995, et Givens *et al.*, 1997).

Pour une distribution π sur θ donnée, la portée de l'approche bayésienne est bien plus étendue que celle de la perspective classique. Par exemple, non seulement la moyenne, la médiane ou le mode de $\pi(\theta|x)$ peuvent être calculés, mais en plus la performance de ces estimateurs (comme la variance et les moments d'ordres plus élevés) peut être évaluée. De plus, la connaissance de la distribution a posteriori permet la détermination des *régions de confiance* sous la forme de régions de plus forte densité a posteriori (*highest posterior density*, HPD), c'est-à-dire des régions de la forme

$$\{\theta; \pi(\theta|x) \geq k\},$$

dans le cas unidimensionnel comme dans le cas multidimensionnel. De la même manière, il est possible de calculer assez naturellement la probabilité d'une hypothèse H_0, en conditionnant sur les observations, soit $P^\pi(\theta \in H_0|x)$. Notons que l'approche bayésienne est la seule permettant ce type d'interprétation, car l'expression $P(\theta = \theta_0) = 0.95$ n'a aucun sens si θ n'est pas une variable aléatoire. D'un point de vue bayésien, cette expression signifie que nous sommes prêts à parier que θ est égal à θ_0 à 19 contre 1. Les Chapitres 4 et 5 sont consacrés à l'étude des techniques d'estimation qui incluent des exigences décisionnelles. Nous nous contentons ici d'illustrer la simplicité de cette approche en construisant un intervalle de confiance dans l'exemple suivant.

Exemple 1.24. Soient $x \sim \mathcal{N}(\theta, 1)$ et $\theta \sim \mathcal{N}(0, 10)$. Par conséquent, pour[15] x donné,

$$\begin{aligned}\pi(\theta|x) &\propto f(x|\theta)\pi(\theta) \propto \exp\left(-\frac{(x-\theta)^2}{2} - \frac{\theta^2}{20}\right)\\ &\propto \exp\left(-\frac{11\theta^2}{20} + \theta x\right)\\ &\propto \exp\left(-\frac{11}{20}\left\{\theta - (10x/11)\right\}^2\right)\end{aligned}$$

et donc $\theta|x \sim \mathcal{N}(\frac{10}{11}x, \frac{10}{11})$. Une région de confiance naturelle est alors

[15] Le symbole de proportionnalité s'entend en termes de fonctions de θ (et non de x). Tout en restant tout à fait rigoureux, les calculs qui reposent sur des relations proportionnelles permettent en général une plus grande efficacité dans l'obtention de la distribution a posteriori. En effet, les densités de probabilité sont uniquement déterminées par leur forme fonctionnelle et la constante de normalisation peut être retrouvée, si nécessaire, à la fin du calcul. Cette technique sera donc utilisée abondamment dans cet ouvrage. Évidemment, elle n'est pas toujours appropriée, par exemple quand la constante de proportionnalité est nulle ou infinie, comme on le verra dans la Section 1.5.

$$C = \{\theta; \pi(\theta|x) > k\}$$
$$= \left\{\theta; \left|\theta - \frac{10}{11}x\right| > k'\right\} .$$

Nous pouvons aussi associer un *niveau de confiance* α à cette région, dans le sens où, si $z_{\alpha/2}$ est le fractile $\alpha/2$ de $\mathcal{N}(0,1)$,

$$C_\alpha = \left[\frac{10}{11}x - z_{\alpha/2}\sqrt{\frac{10}{11}},\ \frac{10}{11}x + z_{\alpha/2}\sqrt{\frac{10}{11}}\right]$$

a une probabilité a posteriori $(1-\alpha)$ de contenir θ. ‖

Nous verrons dans le Chapitre 10 que la distribution a posteriori peut parfois être décomposée en plusieurs niveaux selon une structure hiérarchique, les paramètres des premiers niveaux étant traités comme des variables aléatoires, suivant des distributions a priori supplémentaires. Mais cette décomposition est purement utilitaire et ne modifie pas la structure fondamentale du modèle bayésien.

Un problème que nous n'avons pas évoqué ci-dessus est le fait que, bien que toutes les quantités a posteriori soient définies automatiquement d'un point de vue conceptuel comme intégrales par rapport à la distribution a posteriori, il est assez difficile dans la pratique de fournir une valeur numérique. En particulier, une forme explicite de la distribution a posteriori ne peut pas toujours être obtenue. En fait, la complexité de la distribution a posteriori augmente quand les paramètres sont continus et la dimension de Θ est importante.

Ces difficultés de calcul sont étudiées dans le Chapitre 6, où nous fournissons quelques solutions générales. Cependant, elles ne doivent pas être considérées comme un inconvénient majeur de l'approche bayésienne. En effet, la Statistique numérique[16] est actuellement en train de subir un développement très rapide et elle nous permet de rejeter la notion de distribution a priori choisie pour la simplicité des calculs, même si nous pouvons toujours compter sur ces distributions particulières pour présenter les exemples de façon claire et simple dans cet ouvrage. Au contraire, il est encourageant de voir que nous nous approchons de l'objectif de fournir un outil statistique plus performant et plus efficace grâce à ces nouvelles techniques de calcul qui permettent l'utilisation de distributions a priori plus complexes et aussi plus représentatives de l'information a priori.

[16] Nous avons préféré traduire *computational* en *numérique*, plutôt que d'employer le néologisme *computationnel*, assez lourd, même si *comput* et *computer* ont existé en ancien français... En particulier, avant la Renaissance, *comput* était employé à la place de *mathématique* en tant que matière scolaire.

1.5 Distributions a priori impropres

Lorsque le paramètre θ peut être traité comme une variable aléatoire avec une distribution de probabilité π connue, nous avons vu dans la section ci-dessus que le théorème de Bayes est la base de l'inférence bayésienne, car il donne la distribution a posteriori. Cependant, dans de nombreux cas, la distribution a priori est déterminée par des critères subjectifs ou théoriques qui conduisent à une mesure σ-finie sur l'espace des paramètres Θ plutôt qu'à une mesure de probabilité, c'est-à-dire une mesure π telle que

$$\int_{\Theta} \pi(\theta)\, d\theta = +\infty.$$

Dans de tels cas, on dit que la distribution a priori est *impropre* (ou *généralisée*). (Une définition alternative des estimateurs de Bayes généralisés est considérée dans le Chapitre 2.)

Lorsque cette distribution découle de raisons subjectives, le décideur évaluant par exemple la vraisemblance relative des différentes parties de l'espace des paramètres Θ (voir le Chapitre 3), il est logique que, pour de grands espaces de paramètres, par exemple lorsque Θ n'est pas dénombrable, la somme des poids, c'est-à-dire la mesure de Θ, soit infinie.

Exemple 1.25. Soit une distribution $f(x-\theta)$ telle que le *paramètre de position* θ appartient à $\mathbb{R}$ sans restriction. Si aucune information a priori n'est disponible sur le paramètre θ, il est assez raisonnable de considérer que la vraisemblance d'un intervalle $[a, b]$ doit être proportionnelle à sa longueur $b-a$: l'a priori est donc proportionnel à la *mesure de Lebesgue* sur $\mathbb{R}$. C'est aussi la distribution choisie par Laplace (voir l'Exemple 1.12). ‖

Quand une telle loi a priori impropre a été obtenue par des méthodes automatiques, à partir de la densité $f(x|\theta)$ (voir le Chapitre 3), elle paraît plus susceptible aux critiques, mais soulignons les points suivants.

(1) Ces approches automatiques sont souvent la seule façon d'obtenir une distribution a priori dans un cadre non informatif. Dans certains cas, l'unique information disponible (ou retenue) est la connaissance de la distribution d'échantillon $f(x|\theta)$. Cette généralisation du paradigme bayésien rend ainsi possible une extension supplémentaire de l'applicabilité des techniques bayésiennes.

(2) Les performances des estimateurs obtenus à partir de ces distributions généralisées sont en général suffisamment bonnes pour justifier leur utilisation. De plus, elles permettent souvent l'obtention d'estimateurs classiques comme l'estimateur du maximum de vraisemblance, et garantissent donc une fermeture du champ inférentiel en proposant une approche alternative située aux frontières du paradigme bayésien.

(3) Les lois a priori généralisées se situent souvent à la limite des distributions propres (suivant plusieurs topologies). Elles peuvent être donc interprétées comme un cas extrême où la précision de l'information a priori a complètement disparu et elles semblent donner une réponse plus *robuste* (ou plus *objective*) en termes d'une possible *erreur de spécification* de la loi a priori (interprétation erronée de la faible information a priori).

(4) Ce type de distributions est généralement plus acceptable par les non-bayésiens, en partie pour les raisons évoquées aux points (2) et (3), mais aussi parce qu'elles peuvent avoir des justifications fréquentistes, comme :
   (i) *la minimaxité*, qui conduit habituellement aux *distributions les moins favorables* définies dans le Chapitre 2 ;
   (ii) *l'admissibilité*, les lois propres et certaines lois impropres engendrant des estimateurs admissibles et, réciproquement, les estimateurs de Bayes étant parfois les seuls estimateurs admissibles (voir le Chapitre 8) ; et
   (iii) *l'invariance*, le meilleur estimateur équivariant étant un estimateur de Bayes pour la *mesure de Haar* (généralement impropre) définie pour le groupe de transformations correspondant (voir le Chapitre 9).

(5) Une perspective récente (voir par exemple Berger, 2000) est que les lois a priori impropres devraient être privilégiées par rapport aux lois a priori propres vagues, comme une distribution $\mathcal{N}(0, 100^2)$, car ces dernières donnent une fausse impression de sécurité due à leur caractère propre tout en manquant de robustesse en termes d'influence sur les résultats d'inférence.

Ces raisons ne convainquent pas tous les bayésiens (voir, par exemple, Lindley, 1965), mais l'introduction de distributions impropres dans le schéma bayésien permet une fermeture du cadre inférentiel au sens topologique.

D'un point de vue plus pratique, le fait que la distribution a priori soit impropre affaiblit la symétrie entre observations et paramètres, mais *tant que la distribution a posteriori est définie*, les méthodes bayésiennes restent applicables. En fait, la notion de mesure conditionnelle n'est pas clairement définie en théorie de la mesure, bien que Hartigan (1983) l'ait préconisée comme une extension. Cependant, la convention est de considérer la distribution a posteriori $\pi(\theta|x)$ définie par la formule de Bayes

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_\Theta f(x|\theta)\pi(\theta)\,d\theta},$$

pourvu que la pseudo-distribution marginale $\int_\Theta f(x|\theta)\pi(\theta)\,d\theta$ soit correctement définie. C'est une condition impérative pour utiliser les lois a priori impropres, qui est (presque) toujours vérifiée par les lois a priori propres (Exercice 1.46).

Exemple 1.26. (Suite de l'Exemple 1.25) Si $f(x-\theta)$ est la densité de la distribution normale $\mathcal{N}(\theta, 1)$ et $\pi(\theta) = \varpi$, une constante arbitraire, la

pseudo-distribution marginale est la mesure

$$m(x) = \varpi \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-(x-\theta)^2/2\right\} d\theta = \varpi$$

et, par la formule de Bayes, la distribution a posteriori de θ est

$$\pi(\theta \mid x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x-\theta)^2}{2}\right\},$$

c'est-à-dire qu'elle correspond à $\mathcal{N}(x, 1)$. Notons que la constante ϖ ne joue pas un rôle dans la distribution a posteriori, et que cette dernière est en fait la fonction de vraisemblance. Par conséquent, même si les lois a priori impropres ne peuvent pas être normalisées, ceci n'a pas d'importance, car la constante n'a pas d'intérêt pour l'inférence statistique (cependant, voir le Chapitre 5 pour une exception importante). ‖

Dans la version bayésienne du principe de vraisemblance, seule importe la distribution a posteriori. La généralisation à des distributions a priori impropres ne devrait donc pas poser de problèmes, au sens où une distribution a posteriori correspondant à une loi (a priori) impropre peut être utilisée de la même façon qu'une distribution a posteriori normale *quand elles sont définies*. Évidemment, l'interprétation de la loi a priori est plus délicate. Par exemple, dans l'Exemple 1.25, le poids a priori relatif de tout intervalle est nul, mais cela ne veut pas dire qu'un intervalle est invraisemblable a priori. En réalité, traiter des lois a priori impropres comme des lois a priori standard peut conduire à des difficultés comme les *paradoxes de marginalisation* (voir le Chapitre 3), car le calcul habituel des probabilités conditionnelles ne peut pas s'appliquer dans ce cadre. Comme l'affirme Lindley (1990), *l'erreur est de les interpréter [les lois a priori non informatives] comme des représentations d'une complète ignorance.*

Il peut arriver que pour certaines observations x, la distribution a posteriori ne soit pas définie (Exercices 1.48-1.51). La solution la plus habituelle est de déterminer la réponse impropre comme une limite définie à partir de lois a priori propres (tout en s'assurant que la distribution impropre obtenue est justifiée).

Exemple 1.27. Soit une observation binomiale, $x \sim \mathcal{B}(n, p)$, comme dans l'exemple originel de Bayes. Quelques auteurs (voir Novick et Hall, 1965, et Villegas, 1977) contestent le choix de Laplace de la loi uniforme sur $[0, 1]$ comme distribution a priori automatique, car celle-ci apparaît comme étant biaisée contre les valeurs extrêmes 0 et 1. Ils proposent de considérer plutôt l'a priori de Haldane (1931)

$$\pi^*(p) \propto [p(1-p)]^{-1}.$$

Dans ce cas, la loi marginale,

$$m(x) = \int_0^1 [p(1-p)]^{-1} \binom{n}{x} p^x (1-p)^{n-x} dp$$
$$= B(x, n-x),$$

n'est définie que pour $x \neq 0, n$. En conséquence, $\pi(p|x)$ n'existe pas pour ces deux valeurs extrêmes de x, car le produit $\pi^*(p)p^x(1-p)^{n-x}$ ne peut pas être normalisé pour ces deux valeurs. Pour les autres valeurs de x, la distribution a posteriori est $\mathscr{B}e(x, n-x)$, avec une moyenne a posteriori x/n, qui est aussi l'estimateur du maximum de vraisemblance.

La difficulté en 0 et n peut être résolue de la façon suivante ; la mesure a priori π^* apparaît comme une limite de lois bêta dénormalisées,

$$\pi_{\alpha,\beta}(p) = p^{\alpha-1}(1-p)^{\beta-1},$$

lorsque α et β tendent vers 0. Ces distributions $\pi_{\alpha,\beta}$ donnent comme lois a posteriori $\mathscr{B}e(\alpha + x, \beta + n - x)$, malgré l'absence de facteur normalisant, puisque le choix de cette constante n'a pas d'impact. La distribution a posteriori $\pi_{\alpha,\beta}(p|x)$ a pour espérance

$$\delta^{\pi}_{\alpha,\beta}(x) = \frac{x+\alpha}{\alpha+\beta+n},$$

qui tend vers x/n quand α et β tendent vers 0. Si la moyenne a posteriori est la quantité d'intérêt, nous pouvons alors étendre la procédure inférentielle aux cas $x = 0$ et $x = n$ en considérant également x/n comme un estimateur bayésien (uniquement) formel. ‖

Exemple 1.28. Soit $x \sim \mathscr{N}(0, \sigma^2)$. Il découle de considérations d'invariance qu'une distribution a priori intéressante pour σ est la mesure $\pi(\sigma) = 1/\sigma$ (voir le Chapitre 6). Ceci donne comme loi a posteriori

$$\pi(\sigma^2|x) \propto \frac{e^{-x^2/2\sigma^2}}{\sigma^2},$$

qui n'est pas définie pour $x = 0$. Cependant, de par la continuité de la variable aléatoire x, cette difficulté a beaucoup moins d'importance que dans l'Exemple 1.27. ‖

Évidemment, ces arguments limites sur mesure ne sont pas toujours justifiés, en particulier parce que l'estimateur résultant peut dépendre du choix de la suite convergente. Un exemple de ce phénomène est fourni par Richard (1973) (voir aussi Bauwens, 1991) dans le cas d'une distribution normale $\mathscr{N}(\theta, \sigma^2)$, lorsque $\pi(\theta)$ est la mesure de Lebesgue et σ^{-2} est distribué selon une loi gamma $\mathscr{G}(\alpha, s_0^2)$, c'est-à-dire quand

$$\pi(\theta, \sigma^2) \propto \frac{1}{\sigma^{2(\alpha+1)}} e^{-s_0^2/2\sigma^2} ;$$

l'estimateur de θ dépend alors du comportement du rapport $s_0^2/(\alpha-1)$ quand numérateur et dénominateur tendent simultanément vers 0.

De plus, lorsqu'on estime une fonction *discontinue* de θ, l'estimateur pour la loi limite peut différer de la limite des estimateurs. C'est le cas par exemple, en théorie des tests, pour le *paradoxe de Jeffreys-Lindley* (voir le Chapitre 5). Enfin, dans certains cadres, la distribution a priori impropre ne peut pas être utilisée si facilement, comme dans l'estimation des modèles de mélange (voir l'Exercice 1.56 et le Chapitre 6) ou en théorie des tests lorsqu'on teste des hypothèses bilatérales (voir les Exercices 1.60-1.62 et le Chapitre 5).

Il est donc important de prendre plus de précautions quand on a affaire à des lois impropres, afin d'éviter les distributions mal définies. Dans cet ouvrage, les lois impropres seront toujours utilisées en supposant implicitement que la distribution a posteriori correspondante existe, même s'il existe des situations où cette condition peut être relâchée (voir la Note 1.8.3).

La difficulté pratique est de vérifier la condition d'intégrabilité

$$\int f(x|\theta)\pi(\theta)\,d\theta < \infty$$

dans des situations complexes, comme les modèles hiérarchiques (voir l'Exercice 1.66 et le Chapitre 10), où l'utilisation de lois a priori impropres au niveau supérieur de la hiérarchie est assez commune. Le problème y est même plus aigu parce que les nouveaux outils de calcul comme les algorithmes MCMC (Chapitre 6) ne nécessitent pas dans la pratique de vérifier cette condition. (Voir Note 1.8.3 et Hobert et Casella, 1996, 1998.)

Nous voudrions insister de nouveau sur le fait que la principale justification des distributions a priori impropres est de vouloir clore le champ inférentiel bayésien pour des raisons subjectives, axiomatiques (liées aux *résultats sur les classes complètes*, voir le Chapitre 8) et pratiques. Cette extension ne modifie pas la complexité de l'inférence, cependant, puisque la distribution a posteriori est bien une distribution de probabilité.

1.6 Le choix bayésien

Pour clore cette introduction, nous voudrions attirer l'attention des lecteurs sur le fait qu'il existe un choix bayésien. Il est donc toujours possible d'adhérer ou non à ce choix. Bien que nous le défendions avec vigueur, ce n'est pas une excuse pour devenir trop véhément. La plupart des théories statistiques, comme celles présentées par Lehmann et Casella (1998), ont un niveau raisonnable de cohérence et donnent le plus souvent des résultats similaires lorsque le nombre d'observations devient grand en regard du nombre de paramètres (voir la Note 1.8.4).

Si nous ne présentons pas ces autres théories dans ce livre, c'est pour des raisons à la fois philosophiques et pratiques (exposées dans le Chapitre 11), et aussi par souci de présenter un discours unifié sur la Statistique,

tel que toute procédure soit une conséquence logique d'un ensemble donné d'axiomes. Tel est sans doute pour nous l'argument premier pour adhérer au choix bayésien, à savoir la cohérence fondamentale des axiomes de l'inférence statistique bayésienne. En modélisant des paramètres inconnus de la distribution d'échantillonnage à travers une structure probabiliste, donc en probabilisant l'inconnu, l'approche bayésienne autorise un discours quantitatif sur ces paramètres. Elle permet aussi l'incorporation de l'information a priori *et* de l'imprécision de cette information dans la procédure inférentielle. En outre, à part des arguments subjectifs et axiomatiques en faveur de l'approche bayésienne, qui reste le seul système permettant de conditionner sur les observations (et donc de mettre en œuvre le principe de vraisemblance), il faut prendre en compte le fait que les estimateurs de Bayes sont aussi essentiels pour les notions d'optimalité fréquentiste en Théorie de la Décision. De fait, ils peuvent fournir des outils essentiels même pour les statisticiens qui refusent l'élicitation a priori et l'interprétation bayésienne de la réalité.

1.7 Exercices

Section[17] **1.1**

1.1 *(Kelker, 1970) Un vecteur $x \in \mathbb{R}^p$ est distribué selon une *distribution à symétrie sphérique* si $e.x$ a la même distribution que x pour toute transformation orthogonale e.

a. Montrer que, lorsqu'une distribution à symétrie sphérique admet une densité, celle-ci est fonction de $x^t x$ uniquement.

b. Montrer que, si la densité de x est $\varphi(x^t x)$, la densité de $r = ||x||$ est proportionnelle à

$$r^{p-1}\,\varphi(r^2),$$

et donner le coefficient de proportionnalité.

c. Montrer que, si $x = (x_1', x_2')'$ avec $x_1 \in \mathbb{R}^q$ et $x_2 \in \mathbb{R}^{p-q}$, et $||x||^2 = ||x_1||^2 + ||x_2||^2$, la densité de $(r_1, r_2) = (||x_1||, ||x_2||)$ est proportionnelle à

$$r_1^{q-1}\, r_2^{p-q-1}\, \varphi\left(r_1^2 + r_2^2\right).$$

d. En déduire que

$$U = \frac{||x_1||^2}{||x_1||^2 + ||x_2||^2}$$

est distribuée selon une distribution bêta $\mathcal{B}e(q/2,\ (p-q)/2)$.

e. Conclure que

$$\frac{p-q}{q}\,\frac{||x_1||^2}{||x_2||^2}$$

[17] Les exercices signalés par une étoile sont plus avancés, mais ils offrent une vision plus générale des points traités dans chaque chapitre. Ils peuvent être pris comme des compléments utiles, ou, pour la plupart des lecteurs, comme une lecture guidée des articles pertinents.

est distribuée selon la distribution $\mathcal{F}_{p-q,q}$ indépendamment de la distribution à symétrie sphérique de x. En déduire que le rapport de F est *robuste* au sens où sa distribution est constante sur l'ensemble des distributions à symétrie sphérique.

1.2 ⋆(Gouriéroux et Monfort, 1996) Cet exercice illustre le fait que la frontière entre modèles paramétriques et non paramétriques est relativement difficile à déterminer. Cependant, le paramètre ne peut pas être identifié dans le second cas.

a. Montrer qu'une fonction de répartition se caractérise par les valeurs qu'elle prend en les nombres rationnels.

b. En déduire que la collection des fonctions de répartition sur $\mathbb{R}$ a la puissance du continu (cardinal de l'ensemble des parties de $\mathbb{N}$, ensemble des entiers naturels) et donc que toutes les distributions de probabilité sur $\mathbb{R}$ peuvent être indexées par un paramètre réel.

1.3 Montrer que, si $x_1, \ldots, x_n$ sont des variables explicatives et $y_1, \ldots, y_n$ sont distribués selon $\mathbb{E}[y_i] = bx_i$, l'*estimateur des moindres carrés* de b, solution de

$$\min_b \sum_{i=1}^{n} (y_i - bx_i)^2,$$

est aussi estimateur du maximum de vraisemblance sous l'hypothèse de normalité.

1.4 Dans l'Exemple 1.3, donner l'espérance de n. Est-ce que cela signifie que $20 \times 30/n$ est un estimateur sans biais de N ?

1.5 Dans l'Exemple 1.6, montrer que les moments de $x \sim f(x)$ peuvent s'écrire $\mathbb{E}[x^k] = p\mathbb{E}[x_1^k] + (1-p)\mathbb{E}[x_2^k]$. En déduire un estimateur des moments de $(p, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$. [*Note :* Historiquement, il s'agit de l'estimateur de Pearson, 1894.]

Section 1.2

1.6 Calculer les probabilités de l'Exemple 1.11 pour l'approximation

$$\Phi(-x) \simeq \frac{1}{\sqrt{2\pi}x}\, e^{-x^2/2},$$

qui est valide lorsque x est grand.

1.7 Un examen comporte quinze questions, chacune ayant trois réponses possibles. Supposons que 70% des étudiants passant l'examen sont bien préparés et répondent correctement à chaque question avec une probabilité de 0.8 ; les 30% restants répondent au hasard.

a. Caractériser la distribution de S, la note de chaque étudiant, si un point est accordé à chaque bonne réponse.

b. Il faut huit bonnes réponses pour réussir l'examen. Conditionnellement au fait qu'un étudiant réussisse un examen, quelle est la probabilité qu'il était bien préparé ?

1.8 Démontrer les versions discrètes et continues du théorème de Bayes.

1.9 ⋆(Romano et Siegel, 1986) Le *paradoxe de Simpson* fournit une illustration de la nécessité d'une approche conditionnelle en Statistique. Soient deux traitements médicaux, T_1 et T_2, T_1 étant appliqué à cinquante patients et T_2 à cinquante

autres. Le résultat de cette expérience donne les pourcentages de survie suivants : 40% pour le traitement T_1, 32% pour le traitement T_2. Donc le traitement T_1 semble meilleur puisqu'il entraîne un taux de survie plus élevé. Cependant, si on prend l'âge en compte, et l'on sépare les patients entre juniors (50) et seniors (50), les taux de succès sont donnés dans la table suivante :

	T_1	T_2
junior	40	50
senior	10	35

et T_1 est moins bon que T_2 dans les deux cas. Expliquer ce paradoxe en utilisant le théorème de Bayes.

1.10 Montrer que la quantité δ qui minimise (1.6) est la médiane de la distribution de ξ. Donner la quantité δ qui minimise le coût quadratique moyen $\mathbb{E}^{\xi}[(\xi-\delta)^2]$.

1.11 Calculer la médiane de la distribution a posteriori associée à la distribution d'échantillonnage (1.7) et l'a priori plat $\pi(\xi) = 1$ sur ξ. [*Note :* Voir Stigler, 1986, pour une solution.]

Section 1.3

1.12 Montrer que, pour un échantillon normal $\mathcal{N}(\theta, \sigma^2)$, il n'existe pas d'estimateur sans biais de σ, mais seulement de puissances entières de σ^2.

1.13 Soit $x \sim P(\lambda)$. Montrer que $\delta(x) = \mathbb{I}_0(x)$ est un estimateur sans biais de $e^{-\lambda}$ qui est nul avec probabilité $1 - e^{-\lambda}$.

1.14 *Une statistique S est dite *libre* si sa distribution ne dépend pas du paramètre θ et *complète* si $\mathbb{E}_\theta[g(S)] = 0$ pour tout θ implique $g(s) \equiv 0$. Montrer que, si S est complète et exhaustive minimale, elle est indépendante de toute statistique libre. [*Note :* Ce résultat est appelé *théorème de Basu*. La réciproque est fausse.]

1.15 Soit un échantillon $x_1, \ldots, x_n$ de variables i.i.d. de fonction de répartition F.
a. Donner la densité de la statistique d'ordre.
b. Montrer que $O = (X_{(1)}, ..., X_{(n)})$ est exhaustive. Quelle est la distribution conditionnelle de $(X_1, ..., X_n)$ sachant O ?
c. Soient $X_1, ..., X_n$ i.i.d. de densité complètement inconnue. Montrer que O est alors complète.

1.16 Montrer qu'une statistique T est exhaustive si et seulement si

$$\ell(\theta|x) \propto \ell(\theta|T(x)).$$

1.17 (Berger et Wolpert, 1988, p. 21) Soit x de support $\{1, 2, 3\}$ et de distribution $f(\cdot \mid 0)$ ou $f(\cdot \mid 1)$, avec

	x		
	1	2	3
$f(x\|0)$	0.9	0.05	0.05
$f(x\|1)$	0.1	0.05	0.85

Montrer que la procédure qui rejette l'hypothèse $H_0 : \theta = 0$ (pour accepter $H_1 : \theta = 1$) est correcte avec une probabilité de 0.9 lorsque $x = 2, 3$ (sous H_0 et H_1). Quelle est l'implication du principe de vraisemblance quand $x = 2$?

1.18 Montrer que le principe de la règle d'arrêt exposé dans l'Exemple 1.18 est une conséquence du principe de vraisemblance dans le cas discret. [*Note :* Voir Berger et Wolpert, 1988, pour une généralisation au cas continu.]

1.19 Pour l'Exemple 1.18, montrer que la règle d'arrêt τ est finie avec probabilité 1. (*Indication :* Utiliser la *loi du logarithme itéré*. Voir Billingsley, 1995.)

1.20 (Berger et Wolpert, 1988) Montrer que, si $z \sim f(z|\theta)$ et si $x = t(z)$, x est une statistique exhaustive si et seulement si pour tout a priori π sur θ, $\pi(\theta|x) = \pi(\theta|z)$.

1.21 Soient $x_1, \ldots, x_n$ distribués selon $\mathcal{E}xp(\lambda)$. Ces données sont *censurées* au sens où il existe n variables aléatoires $y_1, \ldots, y_n$ distribuées selon $f(y)$, indépendamment de λ, et $z_1 = x_1 \wedge y_1, \ldots, z_n = x_n \wedge y_n$ sont les variables réellement observées.

a. Montrer que, selon le principe de vraisemblance, l'estimation de λ ne devrait pas dépendre de f.

b. Étendre ce résultat à d'autres types de censures.

1.22 (Berger, 1985b) Dans le cadre de l'Exemple 1.16, montrer que, pour le test UMPU $H_0 : p = 1/2$, l'hypothèse nulle sera acceptée ou rejetée au niveau 5%, selon la distribution considérée. En déduire que la théorie fréquentiste des tests n'est pas compatible avec le principe de vraisemblance. (*Indication :* Voir Chapitre 5 pour des définitions.)

1.23 Montrer que la densité $g(x_1, x_2|\theta)$ donnée dans l'Exemple 1.17 est effectivement une densité de probabilité.

1.24 Cet exercice a pour but de généraliser les Exemples 1.16 et 1.17 au cas continu, en démontrant qu'il peut y avoir aussi incompatibilité entre l'approche fréquentiste et le principe de vraisemblance dans ce cas.

a. Si $f(x|\theta)$ est une densité telle que x soit une statistique complète, montrer qu'il n'existe pas d'autre densité $g(x|\theta)$ telle que les deux fonctions de vraisemblance $\ell_f(\theta|x) = f(x|\theta)$ et $\ell_g(\theta|x) = g(x|\theta)$ sont proportionnelles (en tant que fonctions de θ) pour tout x.

b. Soit maintenant un échantillon $x_1, \ldots, x_n$ distribué selon $f(x|\theta)$. Nous supposons qu'il existe une statistique exhaustive complète $T(x_1, \ldots, x_n)$ de dimension 1 et une statistique libre $S(x_1, \ldots, x_n)$ telle que le couple (T, S) soit une fonction bijective de $(x_1, \ldots, x_n)$. Montrer que, s'il existe une autre densité $g(x_1, \ldots, x_n|\theta)$ telle que les deux fonctions de vraisemblance soient proportionnelles,

$$\ell_g(\theta|x_1, \ldots, x_n) = \omega(x_1, \ldots, x_n)\ell_f(\theta|x_1, \ldots, x_n),$$

le facteur de proportionnalité ω ne dépend que de $S(x_1, \ldots, x_n)$.

c. Dans le cas particulier où $f(x|\theta)$ est la densité exponentielle, $f(x|\theta) = \theta e^{-\theta x}$, donner un exemple d'une densité $g(x_1, \ldots, x_n|\theta)$ telle que les deux fonctions de vraisemblance soient proportionnelles. (*Indication :* Trouver une statistique libre S et construire une fonction $h(x_1, \ldots, x_n)$ ne dépendant que de $S(x_1, \ldots, x_n)$ telle que $\mathbb{E}_\theta[h(x_1, \ldots, x_n)] = 1$.)

– Comparer les longueurs des intervalles de confiance au seuil 10% dans le cadre de l'Exemple 1.19.

– Montrer que les intervalles de confiance de l'Exemple 1.19 sont corrects : pour l'expérience mixte, $x \sim 0.5\mathcal{N}(\theta, 0.1) + 0.5\mathcal{N}(\theta, 10)$ et $P(\theta \in [x - 5.19, x +$

5.19]) = 0.95, tandis que pour l'expérience $\mathscr{E}_1$, $x \sim \mathscr{N}(\theta, 0.1)$ et $P(\theta \in [x - 0.62, x + 0.62]) = 0.95$.

Les exercices suivants (1.25 à 1.35) présentent quelques aspects supplémentaires de l'estimation par maximum de vraisemblance.

1.25 Montrer que, si la fonction de vraisemblance $\ell(\theta|x)$ est utilisée comme une densité en θ, l'inférence résultante n'obéit pas au principe de vraisemblance. (*Indication :* Montrer que la distribution a priori de $h(\theta)$, lorsque h est une transformation bijective, n'est pas la transformée de $\ell(\theta|x)$ selon la règle du jacobien.)

1.26 Soit une variable aléatoire de Bernoulli $y \sim \mathscr{B}([1 + e^{\theta}]^{-1})$.

a. Si $y = 1$, montrer qu'il n'existe pas d'estimateur du maximum de vraisemblance de θ.

b. Montrer qu'on a le même problème lorsque $y_1, y_2 \sim \mathscr{B}([1 + e^{\theta}]^{-1})$ et $y_1 = y_2 = 0$ ou $y_1 = y_2 = 1$. Donner l'estimateur du maximum de vraisemblance dans les autres cas.

1.27 Soient x_1, x_2 deux observations indépendantes de $\mathscr{C}(\theta, 1)$. Montrer que, lorsque $|x_1 - x_2| > 2$, la fonction de vraisemblance est bimodale. Trouver des exemples de x_1, x_2, x_3 i.i.d. $\mathscr{C}(\theta, 1)$ tels que la fonction de vraisemblance ait trois modes.

1.28 La *loi de Weibull* $\mathscr{W}e(\alpha, c)$ est très utilisée en ingénierie et en fiabilité. Sa densité est donnée par

$$f(x|\alpha, c) = c\alpha^{-1}(x/\alpha)^{c-1}e^{-(x/\alpha)^c}.$$

a. Montrer que, lorsque c est connu, ce modèle est équivalent à un modèle gamma.

b. Donner les équations de vraisemblance en α et c et montrer qu'elles n'admettent pas de solutions explicites.

c. Soit un échantillon i.i.d. $x_1, \ldots, x_n$ de $\mathscr{W}e(\alpha, c)$ censuré à droite en y_0. Donner la fonction de vraisemblance correspondante lorsque α et c sont inconnus et montrer qu'il n'existe pas d'estimateur du maximum de vraisemblance explicite dans ce cas.

1.29 *(Robertson *et al.*, 1988) Pour un échantillon $x_1, \ldots, x_n$, et une fonction f sur $\mathscr{X}$, la régression isotonique de f avec les poids ω_i est la solution de la minimisation en g de

$$\sum_{i=1}^{n} \omega_i (g(x_i) - f(x_i))^2,$$

sous la contrainte $g(x_1) \leq \cdots \leq g(x_n)$.

a. Montrer que la solution à ce problème est obtenue par l'algorithme d'agrégation des mauvais classements :

ALGORITHME 1.1. Si f n'est pas isotonique,

(i) trouver i tel que $f(x_{i-1}) > f(x_i)$;

(ii) remplacer $f(x_{i-1})$ et $f(x_i)$ par

$$f^*(x_i) = f^*(x_{i-1}) = \frac{\omega_i f(x_i) + \omega_{i-1} f(x_{i-1})}{\omega_i + \omega_{i-1}},$$

et répéter (i)-(ii) jusqu'à ce que la contrainte soit satisfaite.
Prendre $g = f^*$.

b. Appliquer au cas $n = 4$, $f(x_1) = 23$, $f(x_2) = 27$, $f(x_3) = 25$, $f(x_4) = 28$, avec des poids tous égaux.

1.30 *(**Suite de l'Exercice 1.29**) Le *classement par arbre* simple est obtenu en comparant les effets d'un traitement à un état test. La régression isotonique est alors obtenue sous la contrainte $g(x_i) \geq g(x_1)$ pour $i = 2, \ldots, n$.
a. Montrer que l'algorithme suivant fournit la régression isotonique g^* :

ALGORITHME 1.2. Si f n'est pas isotonique,

(i) classer les $f(x_i)$ par ordre croissant $(i \geq 2)$;

(ii) trouver le plus petit j tel que

$$A_j = \frac{\omega_1 f(x_1) + \ldots + \omega_j f(x_j)}{\omega_1 + \ldots \omega_j} < f(x_{j+1})$$

(iii) poser $g^*(x_1) = A_j = g^*(x_2) = \ldots = g^*(x_j)$, $g^*(x_{j+1}) = f(x_{j+1})$, ...

b. Appliquer au cas $n = 5$, $f(x_1) = 18$, $f(x_2) = 17$, $f(x_3) = 12$, $f(x_4) = 21$ et $f(x_5) = 16$, avec $\omega_1 = \omega_2 = \omega_5 = 1$ et $\omega_3 = \omega_4 = 3$.

1.31 (Olkin *et al.*, 1981) Soient n observations $x_1, \ldots, x_n$ de $\mathscr{B}(k, p)$, k et p étant inconnus.
a. Montrer que l'estimateur du maximum de vraisemblance de k, $\hat{k}$, est tel que

$$(\hat{k}(1-\hat{p}))^n \geq \prod_{i=1}^{n}(\hat{k} - x_i) \quad \text{et} \quad ((\hat{k}+1)(1-\hat{p}))^n < \prod_{i=1}^{n}(\hat{k} + 1 - x_i),$$

où $\hat{p}$ est l'estimateur du maximum de vraisemblance de p.
b. Si l'échantillon est $16, 18, 22, 25, 27$, montrer que $\hat{k} = 99$.
c. Si l'échantillon est $16, 18, 22, 25, 28$, montrer que $\hat{k} = 190$ et conclure sur la stabilité de l'estimateur du maximum de vraisemblance.

1.32 Donner l'estimateur du maximum de vraisemblance de p pour l'Exemple 1.6 lorsque les autres paramètres sont connus et deux observations sont disponibles. Comparer avec la moyenne a posteriori lorsque $p \sim \mathcal{U}_{[0,1]}$.

1.33 (Basu, 1988) Une urne contient 1 000 tickets ; 20 sont marqués θ et 980 sont marqués 10θ. Un ticket est tiré au hasard, et est marqué x.
a. Donner l'estimateur du maximum de vraisemblance de θ, $\delta(x)$, et montrer que $P(\delta(x) = \theta) = 0.98$.
b. Supposons maintenant que 20 tickets soient marqués θ et 980 soient marqués $a_i\theta$ $(i \leq 980)$, avec $a_i \in [10, 10.1]$ et $a_i \neq a_j$ $(i \neq j)$. Donner le nouvel estimateur du maximum de vraisemblance, δ', et montrer que $P(\delta'(x) < 10\theta) = 0.02$. Conclure sur l'attrait de l'estimateur du maximum de vraisemblance dans ce cas.

1.34 (Romano et Siegel, 1986) Pour

$$f(x) = \frac{1}{x} \exp\left[-50\left(\frac{1}{x} - 1\right)^2\right] \qquad (x > 0),$$

montrer que f est intégrable et qu'il existe $a, b > 0$ tels que

$$\int_0^b af(x)dx = 1 \quad \text{et} \quad \int_1^b af(x)dx = 0.99.$$

Pour la distribution de densité

$$p(y|\theta) = a\theta^{-1} f(y\theta^{-1}) \mathbb{I}_{[0,b\theta]}(y),$$

donner l'estimateur du maximum de vraisemblance, $\delta(y)$, et montrer que $P(\delta(y) > 10\theta) = 0.99$.

1.35 (Romano et Siegel, 1986) Soient x_1, x_2, x_3 i.i.d. $\mathcal{N}(\theta, \sigma^2)$.

a. Donner l'estimateur du maximum de vraisemblance de σ^2 pour $(x_1, x_2, x_3) = (9, 10, 11)$ et pour $(x_1, x_2, x_3) = (29, 30, 31)$.

b. Pour trois observations supplémentaires x_4, x_5, x_6, donner l'estimateur du maximum de vraisemblance lorsque $(x_1, \ldots, x_6) = (9, 10, 11, 29, 30, 31)$. Ce résultat contredit-il le principe de vraisemblance ?

Section 1.4

1.36 Si $x \sim \mathcal{N}(\theta, \sigma^2)$, $y \sim \mathcal{N}(\varrho x, \sigma^2)$, comme dans un modèle autorégressif, avec ϱ connu, et $\pi(\theta, \sigma^2) = 1/\sigma^2$, calculer la distribution prédictive de y sachant x.

1.37 Si $y \sim \mathcal{B}(n, \theta)$, $x \sim \mathcal{B}(m, \theta)$, et $\theta \sim \mathcal{B}e(\alpha, \beta)$, donner la distribution prédictive de y sachant x.

1.38 Pour une distribution a priori propre $\pi(\theta)$ et une distribution d'échantillonnage $f(x|\theta)$, montrer que $\pi(\theta|x)$ et $\pi(\theta)$ sont identiques si et seulement si $f(x|\theta)$ ne dépend pas de θ.

1.39 Considérons une distribution a priori π positive sur Θ et $x \sim f(x|\theta)$. Supposons que la vraisemblance $\ell(\theta|x)$ est bornée, continue et admet un maximum unique $\hat{\theta}(x)$.

a. Montrer que, pour un échantillon artificiel $x_n = (x, \ldots, x)$ fait de n réplications de l'observation initiale x, la distribution a posteriori $\pi(\theta|x_n)$ converge vers une masse de Dirac en $\hat{\theta}(x)$.

b. Construire un algorithme bayésien pour calculer les estimateurs du maximum de vraisemblance.

1.40 *Pour un couple (x, y) de variables aléatoires, les distributions marginales $f(x)$ et $f(y)$ ne suffisent pas à caractériser la distribution jointe de (x, y).

a. Donner un exemple de deux distributions bivariées différentes admettant les mêmes distributions marginales. (*Indication :* Prendre des distributions uniformes $\mathcal{U}([0, 1])$ pour les marginales et trouver une fonction de $[0, 1]^2$ dans $[0, 1]^2$ croissante dans les deux dimensions.)

b. Montrer que, à l'inverse, lorsque les deux distributions conditionnelles $f(x|y)$ et $f(y|x)$ sont connues, la distribution du couple (x, y) est définie de manière unique.

c. Étendre b. à un vecteur $(x_1, \ldots, x_n)$ tel que les conditionnelles complètes $f_i(x_i|x_j, j \neq i)$ soient connues. [*Note :* Ce résultat est le théorème de *Hammersley-Clifford*, voir Robert et Casella, 2004.]

d. Montrer que la propriété b. n'est pas forcément vérifiée lorsque $f(x|y)$ et $f(x)$ sont connus, donc que plusieurs distributions $f(y)$ peuvent relier $f(x)$ et $f(x|y)$. (*Indication :* Trouver un contre-exemple.)

e. Donner des conditions suffisantes sur $f(x|y)$ pour que la propriété ci-dessus soit vraie. (*Indication :* Relier ce problème à la théorie des statistiques complètes.)

1.41 Soient $x_1, \ldots, x_n$ i.i.d. $\mathscr{P}(\lambda)$. Montrer que $\sum_{i=1}^n x_i$ est une statistique exhaustive et donner une région de confiance comme dans l'Exemple 1.24 lorsque $\pi(\lambda)$ est une distribution $\mathscr{G}(\alpha, \beta)$. Pour un seuil α donné, comparer sa longueur avec une région de confiance symétrique.

1.42 Donner les distributions marginales et a posteriori dans les cas suivants :
(i) $x|\sigma \sim \mathcal{N}(0, \sigma^2)$, $1/\sigma^2 \sim \mathscr{G}(1, 2)$;
(ii) $x|\lambda \sim \mathcal{P}(\lambda)$, $\lambda \sim \mathscr{G}(2, 1)$;
(iii) $x|p \sim \mathcal{N}eg(10, p)$, $p \sim \mathcal{B}e(1/2, 1/2)$.

1.43 Montrer que, pour un échantillon $x_1, \ldots, x_n$ d'une distribution de densité conditionnelle $f(x_i|\theta, x_{i-1})$, la décomposition d'actualisation (1.13) s'applique aussi. [*Note :* La suite x_i est alors une chaîne de Markov.]

1.44 Montrer que, dans le cadre de l'Exemple 1.22, la distribution a posteriori marginale de ξ_2 est différente de la distribution a priori marginale lorsque $\pi(\xi_1, \xi_2)$ ne se factorise pas en $\pi_1(\xi_1)\pi_2(\xi_2)$.

1.45 *(Dette et Studden, 1997) Dans le cadre de l'Exemple 1.23, nous définissons les *moments canoniques* d'une distribution et montrons qu'ils peuvent être utilisés comme une représentation de cette distribution.

a. Montrer que les deux premiers moments c_1 et c_2 sont reliés par les inégalités suivantes :

$$c_1^2 \leq c_2 \leq c_1$$

et que la suite (c_k) est monotone décroissante vers 0.

b. Soit un polynôme de degré k

$$P_k(x) = \sum_{i=0}^{k} a_i x^i.$$

Déduire de

$$\int_0^1 P_k^2(x) g(x)\, dx \geq 0 \tag{1.14}$$

que

$$a^t C_k a \geq 0, \qquad \forall a \in \mathbb{R}^{k+1}, \tag{1.15}$$

où

$$C_k = \begin{pmatrix} 1 & c_1 & c_2 & \ldots & c_k \\ c_1 & c_2 & c_3 & \ldots & c_{k+1} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ c_k & c_{k+1} & & \ldots & c_{2k} \end{pmatrix}$$

et $a^t = (a_0, a_1, \ldots, a_k)$.

c. Montrer que pour toute distribution g, les moments c_k satisfont

$$\begin{vmatrix} 1 & c_1 & c_2 & \ldots & c_k \\ c_1 & c_2 & c_3 & \ldots & c_{k+1} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ c_k & c_{k+1} & & \ldots & c_{2k} \end{vmatrix} > 0. \tag{1.16}$$

(*Indication :* Interpréter (1.15) comme une propriété de C_k.)

d. En utilisant des inégalités semblables à (1.14) pour les polynômes $t(1-t)P_k^2(t)$, $tP_k^2(t)$, et $(1-t)P_k^2(t)$, prouver les inégalités suivantes sur les moments de g :

$$\begin{vmatrix} c_1 - c_2 & c_2 - c_3 & \dots & c_{k-1} - c_k \\ c_2 - c_3 & c_3 - c_4 & \dots & c_k - c_{k+1} \\ \dots & \dots & \dots & \dots \\ c_{k-1} - c_k & \dots & \dots & c_{2k-1} - c_{2k} \end{vmatrix} > 0, \tag{1.17}$$

$$\begin{vmatrix} c_1 & c_2 & \dots & c_k \\ c_2 & c_3 & \dots & c_{k+1} \\ \dots & \dots & \dots & \dots \\ c_k & c_{k+1} & \dots & c_{2k-1} \end{vmatrix} > 0, \tag{1.18}$$

$$\begin{vmatrix} 1 - c_1 & c_1 - c_2 & \dots & c_{k-1} - c_k \\ c_1 - c_2 & c_2 - c_3 & \dots & c_k - c_{k+1} \\ \dots & \dots & \dots & \dots \\ c_{k-1} - c_k & \dots & \dots & c_{2k-2} - c_{2k-1} \end{vmatrix} > 0. \tag{1.19}$$

e. Montrer que (1.16) (resp. (1.17)) permet de majorer (resp. de minorer) par $\underline{c}_{2k}$ (resp. $\bar{c}_{2k}$) c_{2k} et que (1.18) (resp. (1.19)) permet de majorer (resp. de minorer) par $\underline{c}_{2k-1}$ (resp. $\bar{c}_{2k-1}$) c_{2k-1}.

f. Définissant p_k par

$$p_k = \frac{c_k - \underline{c}_k}{\bar{c}_k - \underline{c}_k},$$

montrer que la relation entre $(p_1, ..., p_n)$ et $(c_1, ..., c_n)$ est bijective pour tout n et que les p_i sont indépendants.

g. Montrer que la transformation inverse est donnée par les formules récursives suivantes. Soit

$$q_i = 1 - p_i, \quad \zeta_1 = p_1, \quad \zeta_i = p_i q_{i-1} \qquad (i \geq 2).$$

Alors

$$\begin{cases} S_{1,k} = \zeta_1 \ + \ldots + \zeta_k \quad (k \geq 1), \\ S_{j,k} = \sum_{i=1}^{k-j+1} \zeta_i S_{j-1,i+j-1} \quad (j \geq 2), \\ c_n = S_{n,n}. \end{cases}$$

Section 1.5

1.46 La difficulté avec les lois a priori impropres, à savoir la non-existence éventuelle de l'intégrale $\int_\Theta f(x|\theta)\pi(\theta)\,d\theta$, ne concerne pas les a priori propres.

a. Rappeler le théorème de Fubini et l'appliquer au couple de fonctions $(f(x|\theta), \pi(\theta))$.

b. En déduire que, si π est une mesure positive finie,

$$\int_\Theta f(x|\theta)\pi(\theta)\,d\theta < \infty \tag{1.20}$$

presque partout.

c. Montrer que, si π est impropre et $f(x|\theta)$ a un support fini, alors $\pi(\theta|x)$ est défini si et seulement si (1.20) est fini pour tout x dans le support de $f(x|\theta)$.

1.47 Montrer que, si π est une mesure positive sur Θ, l'intégrale (1.20) est positive presque partout.

1.48 (Fernandez et Steel, 1999) Soient n observations i.i.d. $x_1, \ldots, x_n$ d'un mélange

$$p\mathcal{N}(\mu_0, \sigma_0^2) + (1-p)\mathcal{N}(\mu_0, \sigma_1^2)\,,$$

où p, μ_0 et σ_0 sont connues. L'a priori sur σ_1 est une distribution bêta $\mathcal{B}e(\alpha, \beta)$. Montrer que, si $r \geq 1$ observations sont égales à μ_0, la distribution a posteriori n'est définie que pour $\alpha > r$. [*Note :* D'un point de vue de théorie de la mesure, l'ensemble des x_i égaux à μ_0 est de mesure nulle. Si une ou plusieurs observations valent exactement μ_0, cela signifie que ce modèle de mélange continu n'est pas approprié.]

1.49 (Suite de l'Exercice 1.48) Soit une observation x d'une loi normale $\mathcal{N}(0, \sigma^2)$.

a. Si la loi a priori sur σ est une distribution exponentielle $\mathcal{E}xp(\lambda)$, montrer que la loi a posteriori n'est pas définie pour $x = 0$.

b. Si la loi a priori sur σ est la distribution impropre $\pi(\sigma) = \sigma^{-1}\exp(-\alpha\sigma^{-2})$, avec $\alpha > 0$, montrer que la loi a posteriori est toujours définie.

1.50 (Suite de l'Exercice 1.49) Soit une observation y telle que $y = x - \lambda$, où x suit la distribution de Laplace ,

$$f(x|\theta) = \theta^{-1}\exp(-|x|/\theta)\,,$$

et λ est distribué selon

$$\pi(\lambda) = |\lambda|^{-1/2}\mathbb{I}_{[-1/2,1/2]}(\lambda)\,.$$

Si θ suit une loi gamma $\mathcal{G}(1/2, a)$ $(a > 0)$, montrer que, si $y = 0$, la distribution a posteriori n'est pas définie.

1.51 (Musio et Racugno, 1999) Soit le modèle de Poisson $\mathcal{P}(\theta)$

$$P_\theta(X = x) = \frac{\theta^x}{x!}e^{-\theta}, \quad x = 0, 1, \ldots, \quad \theta > 0,$$

et la distribution a priori $\pi(\theta) = 1/\theta$. Montrer que pour $x = 0$, la distribution a posteriori n'est pas définie.

1.52 (Raiffa et Schlaifer, 1961) Soit une loi a priori $\mathcal{B}e(\alpha m, (1-m)\alpha)$ sur $p \in [0, 1]$. Montrer que, si m est fixe et α tend vers 0, la loi a priori converge vers une distribution concentrée en deux points, de poids m pour $p = 1$ et $(1 - m)$ pour $p = 0$. Commenter les inconvénients d'une telle approche.

1.53 (Bauwens, 1991) Soient $x_1, \ldots, x_n$ i.i.d. $\mathcal{N}(\theta, \sigma^2)$ et

$$\pi(\theta, \sigma^2) = \sigma^{-2(\alpha+1)}\exp(-s_0^2/2\sigma^2).$$

a. Calculer la distribution a posteriori $\pi(\theta, \sigma^2|x_1, \ldots, x_n)$ et montrer qu'elle ne dépend que de $\bar{x}$ et $s^2 = \sum_{i=1}^n (x_i - \bar{x})^2$.

b. Calculer l'espérance a posteriori $\mathbb{E}^\pi[\theta|x_1, \ldots, x_n]$ et montrer que son comportement lorsque α et s_0 convergent simultanément vers 0 dépend de la limite du rapport $s_0^2/\alpha - 1$.

1.54 Montrer que si l'a priori $\pi(\theta)$ est impropre et l'espace d'échantillonnage $\mathcal{X}$ est fini, la distribution a posteriori $\pi(\theta|x)$ n'est pas définie pour certaines valeurs de x.

1.55 Soient $x_1, \ldots, x_n$ distribués selon $\mathcal{N}(\theta_j, 1)$, avec $\theta_j \sim \mathcal{N}(\mu, \sigma^2)$ $(1 \leq j \leq n)$ et $\pi(\mu, \sigma^2) = \sigma^{-2}$. Montrer que la distribution a posteriori $\pi(\mu, \sigma^2 | x_1, \ldots, x_n)$ n'est pas définie.

1.56 Dans le cadre de l'Exemple 1.6, c'est-à-dire pour un mélange de distributions normales,

a. Montrer que l'estimateur du maximum de vraisemblance n'est pas défini quand tous les paramètres sont inconnus.

b. De même, montrer qu'il n'est pas possible d'utiliser un a priori impropre de la forme

$$\pi_1(\mu_1, \sigma_1)\pi_2(\mu_2, \sigma_2)\pi_3(p)$$

pour estimer ces paramètres. (*Indication :* Écrire la vraisemblance comme une somme de $n + 1$ termes, dépendant du nombre d'observations allouées à la première composante.)

[*Note :* Mengersen et Robert, 1996, montrent qu'il est possible d'utiliser certaines lois a priori impropres en introduisant une dépendance a priori entre les composantes.]

1.57 *__(Suite de l'Exercice 1.56)__ Pour un mélange de deux distributions normales (1.2), si la distribution a priori sur les paramètres est de la forme

$$\pi_1(\mu_1, \sigma_1)\pi_1(\mu_2, \sigma_2)\pi_3(p)$$

et $\pi_3(p) = \pi_3(1 - p)$, montrer que la distribution a posteriori marginale de (μ_1, σ_1) est identique à la distribution a posteriori marginale de (μ_2, σ_2), quel que soit l'échantillon des observations. En déduire que l'espérance a posteriori de (μ_1, σ_1) est égale à l'espérance a posteriori de (μ_2, σ_2) et que ce n'est donc pas un estimateur pertinent. [*Note :* Ce problème est une conséquence de la non-identifiabilité des indices des composants dans un mélange. Il peut être résolu par des contraintes d'identification, comme l'ordonnancement $\mu_1 \leq \mu_2$, ou par l'utilisation de fonctions de perte invariantes par permutation des indices de composantes. Voir Celeux *et al.*, 2000.]

1.58 Construire un argument limite comme dans l'Exemple 1.27 afin de résoudre l'indétermination de l'Exemple 1.28. Calculer l'espérance a posteriori.

1.59 Montrer que, si la distribution a priori est impropre, la pseudo-distribution marginale est aussi impropre.

1.60 *(Hobert et Casella, 1998) Soit un modèle à effets aléatoires.

$$y_{ij} = \beta + u_i + \varepsilon_{ij}, \qquad i = 1, \ldots, I,\ j = 1, \ldots, J,$$

où $u_i \sim \mathcal{N}(0, \sigma^2)$ et $\varepsilon_{ij} \sim \mathcal{N}(0, \tau^2)$. Pour l'a priori

$$\pi(\beta, \sigma^2, \tau^2) = \frac{1}{\sigma^2 \tau^2}\ ,$$

l'a posteriori n'existe pas.

a. En intégrant sur les effets aléatoires (non observables) u_i, montrer que la distribution a posteriori jointe de $(\beta, \sigma^2, \tau^2)$ est

$$\pi(\beta, \sigma^2, \tau^2 | y) \propto \sigma^{-2-I}\ \tau^{-2-IJ}\ \exp\left\{-\tfrac{1}{2\tau^2}\ \textstyle\sum_{i,j}\ (y_{ij} - \bar{y}_i)^2\right\} \times \exp\left\{-\frac{J\ \sum_i (\bar{y}_i - \beta)^2}{2(\tau^2 + J\sigma^2)}\right\}\ (J\tau^{-2} + \sigma^{-2})^{-I/2}\ .$$

b. Intégrer sur β pour obtenir la densité marginale a posteriori

$$\pi(\sigma^2,\tau^2|y) \propto \frac{\sigma^{-2-I}\ \tau^{-2-IJ}}{(J\tau^{-2}+\sigma^{-2})^{I/2}}\ (\tau^2+J\sigma^2)^{1/2} \times \exp\left\{-\frac{1}{2\tau^2}\ \sum_{i,j}\ (y_{ij}-\bar{y}_i)^2 - \frac{J}{2(\tau^2+J\sigma^2)}\ \sum_i\ (\bar{y}_i-\bar{y})^2\right\}.$$

c. Montrer que la densité a posteriori jointe n'est pas intégrable. (*Indication :* Pour $\tau \neq 0$, $\pi(\sigma^2,\tau^2|y)$ se comporte comme σ^{-2} au voisinage de 0.)

d. Montrer que les distributions conditionnelles

$$\begin{aligned} U_i|y,\beta,\sigma^2,\tau^2 &\sim \mathscr{N}\left(\frac{J(\bar{y}_i-\beta)}{J+\tau^2\sigma^{-2}}, (J\tau^{-2}+\sigma^{-2})^{-1}\right)\ , \\ \beta|u,y,\sigma^2,\tau^2 &\sim \mathscr{N}(\bar{y}-\bar{u},\tau^2/JI)\ , \\ \sigma^2|u,\beta,y,\tau^2 &\sim \mathcal{IG}\left(I/2,(1/2)\sum_i\ u_i^2\right)\ , \\ \tau^2|u,\beta,y,\sigma^2 &\sim \mathcal{IG}\left(IJ/2,(1/2)\sum_{i,j}\ (y_{ij}-u_i-\beta)^2\right)\ , \end{aligned}$$

sont bien définies. [*Note :* Les conséquences de cette définition de la densité a posteriori jointe seront clarifiées dans le Chapitre 6.]

1.61 *Soit un modèle probit dichotomique, où $(1 \leq i \leq n)$

$$P(d_i=1) = 1 - P(d_i=0) = P(z_i \geq 0)\,, \tag{1.21}$$

avec $z_i \sim \mathcal{N}(r_i\beta,\sigma^2)$, $\beta \in \mathbb{R}$, r_i étant une variable explicative. (Noter que les z_i ne sont *pas* observés.)

a. Montrer que le paramètre (β,σ) n'est pas identifiable.

b. Pour la distribution a priori $\pi(\beta,\sigma) = 1/\sigma$, montrer que la distribution a posteriori n'est pas définie.

c. Pour la distribution a priori

$$\sigma^{-2} \sim \mathcal{G}a(1.5,1.5)\ , \qquad \beta|\sigma \sim \mathcal{N}(0,10^2)\ ,$$

montrer que la distribution a posteriori est bien définie.

d. Une contrainte d'identification possible est $\sigma = 1$. Donner des conditions suffisantes sur les observations (d_i,r_i) pour que la distribution a posteriori sur β soit définie si $\pi(\beta) = 1$.

e. Même question que d. lorsque la distribution normale sur les z_i est remplacée par la fonction logistique, c'est-à-dire

$$P(d_i=1) = 1 - P(d_i=0) = \frac{\exp(r_i\beta)}{1+\exp(r_i\beta)}\,,$$

ce qui donne le modèle logit dichotomique.

1.62 *(Kubokawa et Robert, 1994) Dans les *modèles de calibration linéaire*, on s'intéresse à la détermination des valeurs du régresseur x, partant des valeurs observées y, à l'inverse de la régression linéaire standard. Une version simplifiée de ce problème peut s'inscrire dans le cadre de l'observation de variables aléatoires indépendantes

$$y \sim \mathcal{N}_p(\beta, \sigma^2 I_p),\ z \sim \mathcal{N}_p(x_0\beta, \sigma^2 I_p),\ s \sim \sigma^2\chi^2_q\,, \tag{1.22}$$

avec $x_0 \in \mathbb{R}$, $\beta \in \mathbb{R}^p$. Le paramètre d'intérêt est x_0.

a. Une distribution a priori de référence sur (x_0, β, σ) donne la distribution a posteriori jointe

$$\pi(x_0, \beta, \sigma^2|y, z, s) \propto \sigma^{-(3p+q)-\frac{1}{2}}\ \exp\{-(s + \|y-\beta\|^2 + \|z - x_0\beta\|^2)/2\sigma^2\}\ (1+x_0^2)^{-1/2}\,.$$

Montrer que cet a posteriori est compatible avec la distribution d'échantillonnage (1.22).

b. Montrer que la distribution marginale a posteriori de x_0 est

$$\pi(x_0|y, z, s) \propto \frac{(1+x_0^2)^{(p+q-1)/2}}{\left\{\left(x_0 - \dfrac{y^t z}{s+\|y\|^2}\right)^2 + \dfrac{\|z\|^2+s}{\|y\|^2+s} - \dfrac{(y^t z)^2}{(s+\|y\|^2)^2}\right\}^{(2p+q)/2}}\,.$$

c. En déduire que la distribution a posteriori de x_0 est bien définie.

[*Note :* Voir Osborne, 1991, pour une introduction aux problèmes de calibration. Le modèle (1.22) est aussi équivalent au problème de Fieller, 1954. Voir, notamment, Lehmann et Casella, 1998.]

Note 1.8.2

1.63 *(Diaconis et Kemperman, 1996) Montrer que la définition du processus de Dirichlet $\mathscr{D}(F_0, \alpha)$ donnée en Section 1.8.2 est compatible avec la définition suivante : pour une suite de x_i i.i.d. tirés de F_0 et une suite de poids ω_i telles que

$$\omega_1 \sim \mathscr{B}e(1, \alpha), \quad \omega_1 + \omega_2 \sim \mathscr{B}e(1, \alpha)\mathbb{I}_{[\omega_1, 1]}, \ldots$$

la distribution aléatoire

$$F = \sum_{i=1}^{\infty} \omega_i \delta_{x_i}$$

suit $\mathscr{D}(F_0, \alpha)$.

1.64 ***(Suite de l'Exercice 1.63)** Si $F \sim \mathscr{D}(F_0, \alpha)$, la quantité $X = \int xF(dx)$ est une variable aléatoire.

a. Si $\alpha = 1$ et F_0 est une distribution de Cauchy, montrer que X suit aussi une distribution de Cauchy. [*Note :* Ceci est relié à la propriété caractéristique des distributions de Cauchy, qui est que la moyenne de variables aléatoires de Cauchy est aussi une variable de Cauchy, avec les mêmes paramètres.]

b. Si $\alpha = 1$ et $F_0 = \varrho\delta_0 + (1-\varrho)\delta_1$, montrer que X suit une loi bêta $\mathscr{B}e(\varrho, 1-\varrho)$.

c. Montrer que, si $\alpha = 1$ et F_0 est $\mathscr{U}_{[0,1]}$, X a pour densité

$$\frac{e}{\pi}\,\frac{\sin(\pi y)}{(1-y)^{(1-y)}y^y}\,.$$

[*Note :* Voir Diaconis et Kemperman, 1996 pour une formule générale reliant F_0 à la densité de X.]

1.65 *(Diaconis et Kemperman, 1996) Le processus a priori de Dirichlet $\mathscr{D}(F_0, \alpha)$ peut aussi se décrire via le processus dit du *restaurant chinois*. Soit un restaurant ayant beaucoup de grandes tables et assignons à chaque table j une réalisation y_j de F_0. Puis traitons les arrivées comme suit : la première personne qui arrive s'assoit à la première table. La $(n+1)$-ième personne s'assoit à une nouvelle table avec probabilité $\alpha/(\alpha+n)$, ou à la droite d'une personne déjà assise avec probabilité $n/(\alpha+n)$.

a. Si x_i est le numéro z_j de la table à laquelle la personne i est assise, montrer que la suite $x_1, x_2, \ldots$ est échangeable (c'est-à-dire que la distribution est invariante sous toute permutation d'indices).

b. Montrer que $x_1, x_2, \ldots$ peut être considérée comme une suite de réplications i.i.d. tirées de F, où F est distribuée selon $\mathscr{D}(F_0, \alpha)$, en utilisant la distribution conditionnelle donnée en Note 1.8.2.

c. Montrer que cette définition est aussi compatible avec celle de l'Exercice 1.63.

Note 1.8.3

1.66 *(Hadjicostas et Berry, 1999) Soient des observations indépendantes x_i $(i = 1, \ldots, n)$ de distributions de Poisson $\mathscr{P}(\lambda_i t_i)$, où les durées t_i sont connues. Les λ_i suivent indépendamment la distribution a priori gamma $\mathscr{G}(\alpha, \beta)$. Ce modèle est *hiérarchique*, car on suppose que les paramètres (α, β) suivent une distribution a priori $\pi(\alpha, \beta)$ telle que

$$\pi(\alpha,\beta) \propto \alpha^{k_1}(\alpha+s_1)^{k_2}\beta^{k_3}(\beta+s_2)^{k_4}\,, \tag{1.23}$$

où les valeurs k_i sont $s_j > 0$ connues $(i = 1, \ldots, 4,\ j = 1, 2)$.

a. Montrer que la distribution a priori (1.23) est propre, si et seulement si,

$$k_1 + k_2 + 1 < 0, \quad k_1 + 1 > 0, \quad k_3 + k_4 + 1 < 0, \quad \text{et} \quad k_3 + 1 > 0.$$

b. En intégrant sur les λ_i la distribution jointe des λ_i's et de (α, β), calculer la distribution (marginale) a posteriori de (α, β).

c. Montrer que la distribution (marginale) a posteriori de (α, β) est définie (propre) si et seulement si

$$k_1 + y + 1 > 0, \quad k_3 + r + 1 > 0, \quad k_3 > k_1 + k_2$$

et, de plus, soit $k_3 + k_4 + 1 < 0$, soit $k_3 + k_4 + 1 = 0$ et $k_1 + y > 0$, avec

$$y = \sum_{i=1}^{n} \mathbb{I}_0(x_i), \qquad r = \sum_{i=1}^{n} x_i\,.$$

d. Vérifier que les conditions de a. impliquent les conditions de b. (comme convenu).

e. Montrer que les conditions de b. sont satisfaites lorsque $(k_1, \ldots, k_4) = (-8, 0, -5, 0)$ et $(y, r) = (10, 337)$, et que les conditions de a. ne le sont pas dans ce cas.

f. Montrer que les conditions de b. ne sont pas satisfaites lorsque $(k_1, \ldots, k_4) = (-12, 0, 1, 1)$ et $(y, r) = (10, 337)$.

Note 1.8.4

1.67 *(Robins et Ritov, 1997) Soient des observations i.i.d. (x_i, y_i) dans $(0,1)^k \times \mathbb{R}$ tirées du modèle suivant : $x \sim f(x)$, $y|x \sim \mathcal{N}(\theta(x), 1)$, où la fonction moyenne θ est bornée uniformément sur $(0,1)^k$ et la densité f est telle que $c < f(x) < 1/c$ uniformément sur $(0,1)^k$, où $c < 1$ est une constante fixée. Supposons que la quantité d'intérêt est

$$\varphi = \int_{(0,1)^k} \theta(x) dx \,.$$

a. Montrer que l'espace Θ des fonctions moyennes θ est de dimension infinie.

b. Donner la vraisemblance $\ell(\theta, f)$ et montrer qu'elle se factorise en une fonction de f multipliée par une fonction de θ.

c. Lorsque f est connue, montrer que $(x_1, \ldots, x_n)$ est une statistique libre.

d. Lorsque f est inconnue, montrer que $(x_1, \ldots, x_n)$ est θ-libre, au sens où la vraisemblance conditionnelle en $(x_1, \ldots, x_n)$ est fonction de θ uniquement, la distribution marginale de $(x_1, \ldots, x_n)$ est fonction de f uniquement, et l'espace des paramètres est un espace produit. (Voir Cox et Hinkley, 1987, et Robins et Wasserman, 2000, pour plus de détails sur cette notion.)

e. Lorsque f est connue, montrer que

$$\frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{f(x_i)}$$

est un estimateur convergent de φ. (En fait, il s'agit d'un estimateur uniformément convergent en $\sqrt{n}$.)

f. Lorsque f est inconnue, Robins et Ritov (1997) ont démontré qu'il n'existe pas d'estimateur uniformément convergent de φ. En déduire que, si la distribution a posteriori sur (θ, f) se factorise en $\pi_1(\theta)\pi_2(f)$, l'inférence bayésienne sur θ (et donc sur φ) est la même quelle que soit la valeur de f.

g. Au contraire, si la distribution a priori sur (θ, f) rend θ et f dépendants, et si f est connue et vaut f_0, la distribution a posteriori dépend de f_0. En déduire que cette dépendance viole le principe de vraisemblance.

[*Note :* La description simplifiée ci-dessus de Robins et Ritov, 1997, est tirée de Robins et Wasserman, 2000.]

1.8 Notes

1.8.1 Une brève histoire de la Statistique bayésienne

Différents livres ont été écrits sur l'histoire de la Statistique bayésienne, notamment Stigler (1986), Dale (1991), Lad (1996) et Hald (1998). Nous ne faisons ici que souligner quelques points forts du développement de cette discipline durant les deux cents dernières années.

Comme nous le détaillons dans ce chapitre, la formule de Bayes est apparue pour la première fois en 1761, dans le cadre de l'exemple binomial de la Section 1.2, exposé par le révérend Thomas Bayes devant la "Royal Society", et publié de façon posthume par son ami R. Price en 1763. Pierre Simon Laplace redécouvrit ensuite cette formule dans une plus grande généralité en 1773, sans, semble-t-il, avoir connaissance des travaux précédents de Bayes. L'utilisation du principe bayésien devint alors courant pendant le siècle suivant, comme le rapporte Stigler (1986), mais des critiques commencèrent à émerger vers la fin du XIXème siècle, comme par exemple dans Venn (1886) ou Bertrand (1889), en particulier sur le choix de la loi a priori uniforme et des paradoxes de reparamétrisation qui en résultent, voir Zabell (1989).

Puis, malgré des formalisations plus poussées du paradigme bayésien par Edgeworth et (Karl) Pearson au tournant du siècle et, plus tard, par Keynes (1921), le début du XXème siècle fut surtout marqué par, tout d'abord, Kolmogorov, qui proposa dans les années 1920 une axiomatisation de la théorie des probabilités semblant contredire le paradigme bayésien et la notion de probabilité subjective, ensuite par Fisher qui s'éloigna de l'approche bayésienne (Fisher, 1912) en définissant la fonction de vraisemblance (Fisher, 1922), puis en développant la Statistique fiduciaire (Fisher, 1930), et qui ne révisa jamais son opinion négative sur la Statistique bayésienne. Cette opposition paraît quelque peu paradoxale, car la Statistique fiduciaire tentait, en un certain sens, de surmonter la difficulté de choisir une loi a priori en la construisant à partir de la fonction de vraisemblance (Seidenfeld, 1992), dans le même esprit que les approches non informatives de Jeffreys (1939) et Bernardo (1979).

Par exemple, considérant la relation $O = P + \epsilon$ où ϵ est un terme d'erreur, la Statistique fiduciaire tient que, si P (la cause) est connu, O (l'effet) suit la loi définie par la relation ci-dessus. Réciproquement, si O est connu, $P = O - \epsilon$ est distribuée selon la distribution symétrique. De ce point de vue, les observations et les paramètres jouent un rôle *symétrique*, selon la façon dont on analyse le modèle, c'est-à-dire suivant ce qui est connu et ce qui ne l'est pas. Plus généralement, l'approche fiduciaire consiste à renormaliser la vraisemblance (1.3) afin de la transformer en densité *de* θ lorsque

$$\int_\Theta \ell(\theta|x)\, d\theta < +\infty,$$

donc en inversant effectivement les rôles de x et θ. Comme on peut le voir dans l'exemple précédent, le raisonnement sous-tendant cette inversion causale est complètement conditionnel : sachant P, on a $O = P+\epsilon$, et, sachant O, $P = O-\epsilon$. Bien entendu, ce raisonnement ne tient pas d'un point de vue probabiliste : si O est une variable aléatoire et P est un paramètre (constant), écrire $P = O - \epsilon$

n'implique pas que P soit une variable aléatoire. De plus, transformer $\ell(\theta|x)$ en une densité n'est pas toujours possible. L'approche fiduciaire a été abandonnée progressivement après la mise en évidence de paradoxes fondamentaux (voir Stein, 1959, Wilkinson, 1977, et les références dans Zabell, 1992).

Le livre de Jeffreys (1939) est le premier traité moderne de Statistique bayésienne : il couvre, en plus de la notion d'a priori non informatif, celles de loi prédictive, de facteur de Bayes et d'a priori impropre. Mais cet ouvrage publié au moment du développement par Fisher de la Statistique de la vraisemblance et des intervalles de confiance par Neyman (1934), ne rencontra pas le même succès. Les approches alternatives à la Statistique bayésienne devinrent alors standard dans les années 1930, avec l'introduction des estimateurs du maximum de vraisemblance et le développement d'une théorie formalisée de la Statistique mathématique, pour laquelle les lois a priori n'apparaissaient au mieux que comme une façon de construire des estimateurs optimaux, voir Wald (1950) ou Ibragimov et Has'minskii (Ibragimov et Has'minskii, 1981, Chapitre 6).

Les tentatives d'une formalisation plus poussée de l'approche bayésienne par Gini ou de Finetti, des années 1930 aux années 1970, ne se traduisirent pas par une plus grande popularité face à la théorie alors dominante de Neyman-Pearson, même si la communauté bayésienne s'accroissait et produisait des traités tels que ceux de Savage (1954) et de Lindley (1965, 1971).

On peut avancer que ce n'est que très récemment que la Statistique bayésienne a pris un nouvel élan, grâce au développement de nouveaux outils numériques– qui ont toujours joué un rôle central pour le paradigme bayésien–et l'intérêt vite croissant des praticiens pour cette approche de modélisation statistique, comme souligné dans l'article de Berger (2000) sur l'état présent et futur de la Statistique bayésienne[18].

La vitalité actuelle de la Statistique bayésienne peut être mise en évidence par le pourcentage élevé d'articles bayésiens publiés dans les revues statistiques ou concernant d'autres domaines scientifiques. Il semble donc que les praticiens de ce siècle prendront mieux en compte les avantages de la Statistique bayésienne que leurs prédécesseurs du XXème siècle.

1.8.2 Statistique bayésienne non paramétrique

Bien que ce livre se cantonne à l'approche paramétrique de la Statistique, il existe une littérature (de plus en plus) importante sur la Statistique bayésienne non paramétrique. Premièrement, les notions d'optimalité comme la minimaxité jouent un rôle central en estimation fonctionnelle ; de la même façon que dans le cadre paramétrique (voir le Chapitre 3), les estimateurs de Bayes peuvent être utilisés pour la détermination de bornes de minimaxité et d'estimateurs minimax.

Deuxièmement, et d'un point de vue nettement moins formel, il est parfois nécessaire de concevoir une modélisation bayésienne a priori dans un espace de dimension infinie. C'est bien entendu plus difficile, tant pour des raisons mathématiques que pour des raisons de construction de l'a priori. Mais une

[18] On pourra aussi consulter la revue de Fienberg (2005) sur la question historique suivante : à partir de quelle époque la méthodologie utilisant les principes bayésiens a-t-elle pris la dénomination de "bayésienne" ?

première solution est de se situer dans la zone grise entre Statistique paramétrique et non paramétrique comme dans l'Exemple 1.23 : le nombre de paramètres est fini mais croît vers l'infini avec le nombre d'observations. C'est le cas notamment pour l'estimation par noyau, où une densité est approchée par un mélange

$$\frac{1}{n\sigma}\sum_{i=1}^{n} K\left(\frac{x-x_i}{\sigma}\right) ,$$

où K est une densité, et σ peut être estimé d'une façon bayésienne, par des développements d'Hermite (Hjort, 1996), ou des bases d'ondelettes (Müller et Vidakovic, 1999, Chap. 1). Dans ce dernier cas, une fonction f est décomposée sur une base fonctionnelle,

$$f(x) = \sum_i \sum_j \omega_{ij} \Psi\left(\frac{x-\mu_i}{\sigma_j}\right) ,$$

où Ψ est une fonction particulière appelée *ondelette mère*, comme par exemple l'ondelette de Haar

$$\Psi(x) = \mathbb{I}_{[0,1/2)} - \mathbb{I}_{[1/2,1)} ,$$

les paramètres de position et d'échelle μ_i et σ_j étant fixés et connus. Les coefficients ω_{ij} peuvent être associés à une distribution a priori telle que (Abramovich *et al.*, 1998)

$$\omega_{ij} \sim \varrho_i \mathcal{N}(0, \tau_i^2) + (1-\varrho_i)\delta_0 ,$$

où δ_0 est la masse de Dirac en 0.

Une deuxième solution, lorsqu'on cherche à estimer une fonction de répartition F, est d'assigner une distribution a priori à celle-ci. Le choix le plus courant est la distribution de Dirichlet $\mathcal{D}(F_0, \alpha)$, F_0 étant la moyenne a priori et α la précision, comme introduit par Ferguson (1974). Cette loi a priori jouit d'une propriété de cohérence, c'est-à-dire si $F \sim \mathcal{D}(F_0, \alpha)$, le vecteur $(F(A_1), \ldots, F(A_p))$ est distribué selon une loi de Dirichlet au sens usuel du terme, $\mathcal{D}_p(\alpha F_0(A_1), \ldots, \alpha F_0(A_p))$ pour toute partition $(A_1, \ldots, A_p)$. Elle génère cependant des distributions a posteriori qui sont partiellement discrètes : si $x_1, \ldots, x_n$ sont distribués selon F et $F \sim \mathcal{D}(F_0, \alpha)$, la distribution marginale de x_1 conditionnellement à $(x_2, \ldots, x_n)$ est

$$\frac{\alpha}{\alpha+n-1} F_0 + \frac{1}{\alpha+n-1}\sum_{i=2}^{n} \delta_{x_i} .$$

(Voir aussi les Exercices 1.63 et 1.65 pour d'autres caractérisations.) L'approximation de la distribution a posteriori nécessite des outils numériques avancés que nous traiterons dans le Chapitre 6. (Voir Note 6.6.7 pour plus de détails.) D'autres types de distributions a priori ont été proposés dans la littérature comme les *distributions généralisées de Dirichlet* (Hjort, 1996), les *arbres de Pólya* (Fabius, 1964, Lavine, 1992), les *processus bêta* (Hjort, 1996), et les *processus de Lévy* (Phillips et Smith, 1996).

Pour conclure, mentionnons qu'une tendance récente de la Statistique bayésienne est de considérer des modèles de dimension variable, comme les mélanges, les modèles de chaînes de Markov cachées et d'autres modèles dynamiques, ainsi que les réseaux neuronaux, grâce à de nouveaux outils numériques développés par

Grenander et Miller (1994), Green (1995), Phillips et Smith (1996) ou Stephens (1997). C'est le cas, par exemple, pour les modèles de mélange,

$$\sum_{i=1}^{k} p_{ik}\varphi(x|\theta_{ik})$$

où $\varphi(\cdot|\theta)$ est une densité paramétrique, la somme des poids p_{ik} vaut 1 et le nombre de composants k est inconnu. Bien qu'il s'agisse d'un problème paramétrique bien défini, il s'approche plus des impératifs non paramétriques que de l'estimation paramétrique standard (voir Richardson et Green, 1997 ou Marin *et al.*, 2004).

1.8.3 Lois a posteriori propres

Nous savons depuis la Section 1.5 qu'un a priori impropre π ne peut être utilisé dans un but inférentiel que si (1.20) est vérifiée pour l'observation x disponible. Si ce n'est pas le cas, les quantités a posteriori telles que moyenne ou médiane n'ont pas de sens, puisque, par exemple, le rapport

$$\frac{\int_\Theta f(x|\theta)\pi(\theta)\,d\theta}{\int_\Theta \theta f(x|\theta)\pi(\theta)\,d\theta}$$

n'est pas défini. Vérifier la condition (1.20) peut se révéler relativement difficile pour des modèles complexes (voir les Exercices 1.60 et 1.61) ou même simplement impossible. Malheureusement, l'avènement de techniques informatiques comme l'échantillonnage de Gibbs (voir le Chapitre 6) autorise à ne se tenir qu'à la relation $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$ afin de simuler des valeurs de l'a posteriori $\pi(\theta|x)$ et les résultats de cette simulation ne mettent hélas pas toujours en évidence le fait que cet a posteriori n'existe pas (voir Hobert et Casella, 1996). Il existe effectivement des exemples dans la littérature de données analysées avec de telles lois a posteriori non définies, ce problème n'ayant été découvert que plusieurs années après.

Nous verrons cependant dans la Note 6.6.4 qu'il existe de bonnes raisons pour utiliser des a posteriori impropres sur des espaces étendus, c'est-à-dire pour une complétion de θ en (α, θ), tant que la distribution $\pi(\theta|x)$ reste propre.

1.8.4 Propriétés asymptotiques des estimateurs de Bayes

Nous ne développons pas le point de vue asymptotique dans ce livre pour deux raisons principales, la première étant que l'approche bayésienne est intrinsèquement conditionnelle. Lorsqu'on conditionne en x, qui peut être un échantillon $(x_1, \ldots, x_n)$, il n'y a aucune raison de se demander ce qui pourrait arriver si n tendait vers l'infini, puisque n est déterminé par la taille de l'échantillon. Conjecturer sur des valeurs futures des observations revient à mener une analyse fréquentiste, à l'opposé des impératifs de la perspective bayésienne. La seconde raison est que, même si elles ne sont pas construites dans ce but, les procédures bayésiennes ont de bonnes performances asymptotiques dans une large majorité des cas. Il n'est pas si paradoxal que la perspective bayésienne, et en particulier le choix d'un a priori, cessent le plus souvent de produire des résultats véritablement différents de ceux du maximum de vraisemblance lorsque le nombre d'observations devient infiniment plus grand que le nombre de paramètres. (Ce cadre idéal souffre d'exceptions bien connues,

comme le *problème de Neyman-Scott* de l'Exemple 3.35, voir Diaconis et Freedman, 1986, où le nombre de paramètres croît avec le nombre d'observations et donne des estimateurs de Bayes non convergents, voir aussi Robins et Ritov, 1997, et l'Exercice 1.67 qui s'y rapporte.)

Ibragimov et Has'minskii (1981, Chap. 1) démontrent que les estimateurs de Bayes sont *convergents* dans un cadre général, c'est-à-dire qu'ils convergent presque sûrement vers la vraie valeur du paramètre lorsque le nombre d'observations tend vers l'infini. C'est le cas notamment pour les estimateurs δ_α ($\alpha \geq 1$) qui minimisent le coût a posteriori (voir le Chapitre 2) associé à la fonction de perte $\mathrm{L}(\delta, \theta) = |\theta - \delta|^\alpha$, sous des conditions assez faibles sur la distribution a priori π et la densité d'échantillonnage $f(x|\theta)$. Ibragimov et Has'minskii (1981, Chap. 3) établissent aussi (sous des conditions plus fortes) l'efficacité asymptotique de certains estimateurs de Bayes, c'est-à-dire le fait que la distribution a posteriori converge vers la vraie valeur à la vitesse $n^{-1/2}$; voir Schervish (1995) pour plus de détails.

Barron *et al.* (1999) donnent des conditions générales pour la convergence d'une distribution a posteriori dans le sens suivant : la probabilité a posteriori de tout voisinage de Hellinger de la vraie distribution tend vers 1 presque sûrement lorsque la taille de l'échantillon tend vers l'infini. (La *distance de Hellinger* entre deux densités f_1 et f_2 (ou les distributions correspondantes) est définie par

$$d(f_1, f_2) = \int \left(f_1(x)^{1/2} - f_2(x)^{1/2} \right)^2 dx \,.$$

Nous l'utiliserons dans le cadre de la Théorie de la Décision dans le Chapitre 2.) L'hypothèse de base sur la distribution a priori π est qu'elle attribue une masse positive à tout voisinage de Kullback-Leibler de la vraie distribution. (Nous utiliserons aussi la pseudo-distance de Kullback-Leibler dans le Chapitre 2.)

Nous reviendrons cependant à l'asymptotique, dans le Chapitre 3, pour la définition de lois a priori non informatives via l'approximation asymptotique des comportements de queue et, dans le Chapitre 6, pour l'*approximation de Laplace* des intégrales de densités a posteriori.

2

Les bases de la Théorie de la Décision

> "Today would run out according to the Pattern. But over and over he mulled over the decisions he had made since he first entered the Waste. Could he have done something different, something that would have avoided this day, this place ? Next time, perhaps."
>
> **Robert Jordan**, *The Fires of Heaven*.

2.1 Évaluation des estimateurs

Considérant que l'objectif général de la plupart des études inférentielles est de fournir une *décision* au statisticien (ou au client), il semble raisonnable d'exiger un critère d'*évaluation* des procédures de décision qui prenne en compte les conséquences de chaque décision et dépende des paramètres du modèle, c'est-à-dire du vrai état du monde (ou de la nature). Ces *décisions* peuvent être de différents types, par exemple acheter des capitaux selon leurs futurs rendements θ, interrompre une expérience agricole sur une nouvelle culture de productivité θ, estimer la contribution de l'économie souterraine θ au PIB des États-Unis, déterminer si le nombre θ des sans domicile fixe a augmenté depuis le dernier recensement. Un autre type de décision est d'évaluer si une nouvelle théorie scientifique est compatible avec les données expérimentales disponibles. Si aucun critère d'évaluation n'est disponible, il est impossible de comparer différentes procédures décisionnelles et des solutions absurdes, comme l'estimateur $\hat{\theta} = 3$ ou pis encore, la réponse que quelqu'un veut imposer, ne peuvent être éliminées que par un raisonnement

ad hoc. Éviter ce type de raisonnement nécessite un renforcement de l'axiomatisation du cadre inférentiel statistique, appelé *Théorie de la Décision*. Cette structure théorique augmentée est nécessaire à la Statistique pour aboutir à une cohérence autrement inatteignable[19].

Bien que presque tout le monde s'accorde sur le besoin de tels critères d'évaluation, il existe une controverse importante autour du choix de ces critères, car les conséquences de cette décision ne sont pas négligeables. Ces difficultés amènent même certains statisticiens à rejeter complètement la Théorie de la Décision, en s'appuyant sur l'argument qu'une détermination pratique des critères d'évaluation du décideur est totalement impossible dans la plupart des cas.

Ce critère est habituellement appelé *coût* et est défini ci-dessous. L'ensemble des décisions possibles, $\mathscr{D}$, est appelé espace de *décision* et la plupart des exemples théoriques se concentrent sur le cas $\mathscr{D} = \Theta$, qui représente le cadre d'estimation standard.

Définition 2.1. *Une* fonction de coût *est une fonction* L *de* $\Theta \times \mathscr{D}$ *dans* $[0, +\infty)$.

La fonction de coût est censée évaluer la pénalité (ou l'erreur) $\mathrm{L}(\theta, d)$ associée à la décision d quand le paramètre prend la valeur θ. Dans un cadre traditionnel d'estimation du paramètre, lorsque $\mathscr{D}$ est Θ ou $h(\Theta)$, la fonction de coût $\mathrm{L}(\theta, \delta)$ mesure l'erreur commise en évaluant $h(\theta)$ par δ. La Section 2.2 présente un ensemble d'axiomes de rationalité qui garantissent l'existence d'une telle fonction dans un cadre décisionnel.

Dans la pratique, la détermination même de la fonction de coût est souvent difficile, en particulier parce que les conséquences de chaque action pour chaque valeur de θ sont souvent impossibles à déterminer quand $\mathscr{D}$ ou Θ sont de grands ensembles, par exemple quand ils contiennent un nombre infini d'éléments. De plus, dans les modèles qualitatifs, il peut être délicat de quantifier les conséquences de chaque décision. Nous verrons à travers des paradoxes comme le *paradoxe de Saint-Pétersbourg* que, même quand la fonction de coût semble évidente, par exemple lorsque des erreurs peuvent être exprimées comme pertes monétaires, la fonction de coût réelle peut être assez différente de son approximation linéaire et intuitive.

La complexité de la détermination de la fonction de coût subjective du décideur incite souvent le statisticien à recourir aux fonctions de coût classiques ou *canoniques*, choisies pour leur simplicité et leur souplesse mathématique. Ce type de fonction de coût est aussi nécessaire pour un traitement théorique de l'obtention des procédures optimales, quand il n'y a pas de motivation pratique pour le choix d'une fonction de coût en particulier.

[19] L'approche bayésienne est, de notre point de vue, l'étape ultime de cette recherche de cohérence.

Le terme *classique* est lié à une longue histoire qui remonte jusqu'à Laplace (1773) pour le coût absolu (2.4) et à Gauss (1810) pour le coût quadratique (2.2), à l'époque où l'*erreur* en termes de performance des estimateurs ou de conséquences des décisions était confondue avec l'*erreur* au sens de l'irréductible variabilité des variables aléatoires (variance). Mais cette caractéristique ne devrait pas être prise comme une quelconque validation, car une utilisation plus générale de ces coûts ne les légitime pas davantage. En réalité, le recours à ces coûts automatiques (ou génériques), bien que justifié dans la pratique–il vaut mieux malgré tout, prendre une décision en un temps fini en utilisant un critère approximatif que de passer un temps infini à déterminer exactement la fonction de coût correcte–a généré une grande partie des critiques envers la Théorie de la Décision.

Une base fondamentale de la Théorie de la Décision bayésienne est que l'inférence statistique devrait commencer par la détermination de trois facteurs :

(1) la famille des distributions pour les observations, $f(x|\theta)$;
(2) la distribution a priori pour les paramètres, $\pi(\theta)$;
(3) le coût associé aux décisions, $\mathrm{L}(\theta, \delta)$;

les distributions a priori et la fonction de coût, et parfois même la distribution d'échantillonnage résultant de considérations partiellement subjectives. Les partisans de la Théorie de la Décision classique omettent le deuxième point mais les critiques fréquentistes du paradigme bayésien échouent souvent à prendre en compte le problème de la construction de la fonction de coût, même si celle-ci est aussi compliquée que l'obtention de la distribution a priori. De plus, présupposer l'existence d'une fonction de coût implique qu'une certaine information sur le problème considéré est disponible. Cette information peut donc être utilisée plus efficacement pour développer une distribution a priori. En réalité, coût et a priori sont difficiles à dissocier et devraient être analysés simultanément (Lindley, 1985). Nous verrons dans la Section 2.4 un exemple de la *dualité* qui existe entre ces deux facteurs. Nous verrons aussi dans la Section 2.5.4 comment les coûts classiques peuvent être remplacés par des coûts plus intrinsèques (similaires aux lois a priori non informatives présentées dans le Chapitre 3), quand il n'y a aucune information disponible sur la pénalité associée à des décisions erronées ou même sur la paramétrisation d'intérêt.

Dans certains cas il est possible de réduire la classe des fonctions de coût acceptables par des considérations d'*invariance*, par exemple quand le modèle est invariant sous l'action d'un groupe de transformations. Ce type de considérations s'applique aussi au choix de la distribution a priori, comme on le verra dans le Chapitre 9. Il est important de souligner que ces motivations d'invariance sont souvent utilisées dans d'autres approches décisionnelles, lorsqu'une réduction drastique de la classe des procédures inférentielles se révéle nécessaire pour obtenir la "meilleure" solution.

Exemple 2.2. Soit le problème de l'estimation de la moyenne θ d'un vecteur normal, $x \sim \mathscr{N}_n(\theta, \Sigma)$, où Σ est une matrice diagonale connue avec pour éléments diagonaux σ_i^2 $(1 \leq i \leq n)$. Dans ce cas, $\mathscr{D} = \Theta = \mathbb{R}^p$ et δ représente une évaluation de θ. S'il n'y a pas d'information additionnelle disponible sur le modèle, il paraît logique de choisir une fonction de coût qui attribue le même poids à chaque composante, soit donc un coût de la forme

$$\sum_{i=1}^{n} \mathrm{L}\left(\frac{\delta_i - \theta_i}{\sigma_i}\right) ,$$

où L prend son minimum en 0. Effectivement, pour ce type de coût, les composantes ayant une grande variance ne biaisent pas fortement la sélection de l'estimateur résultant. En d'autres termes, les composantes avec une grande variance n'ont pas un poids trop important dès que les erreurs d'estimation $(\delta_i - \theta_i)$ sont normalisées par σ_i. Le choix habituel de L est le coût quadratique $\mathrm{L}(t) = t^2$, ce qui signifie que l'erreur d'estimation globale est la somme des carrés des erreurs de chaque composante. ‖

2.2 La fonction d'utilité

La notion d'utilité (définie comme l'opposé d'une fonction de coût) est utilisée non seulement en Statistique, mais aussi en économie et dans d'autres domaines comme la Théorie des Jeux où il est nécessaire d'*ordonner* les conséquences d'actions ou de décisions. *Conséquences* (ou *récompenses*) sont des notions génériques qui résument l'ensemble des résultats émanant de l'action du décideur. Dans les cas les plus simples, il peut s'agir d'un gain ou d'un coût financier dus à cette décision. Dans le cas de l'estimation, l'utilité peut être une mesure de la distance entre l'évaluation et la vraie valeur du paramètre, comme dans l'Exemple 2.2. Les bases axiomatiques de l'utilité ont été attribuées à von Neumann et Morgenstern (1947) et ont mené à de nombreuses extensions, particulièrement en Théorie des Jeux. Dans un cadre statistique, cette approche a été considérée par Wald (1950) et Ferguson (1967). Des extensions et des commentaires additionnels peuvent être trouvés dans (DeGroot, 1970, Chapitre 7) ; pour des références sur la théorie de l'utilité, voir Fishburn (1988) et Machina (1982, 1987). Voir aussi Chamberlain (2000) pour une connexion avec l'Économétrie.

Le cadre général sous-tendant la théorie de l'utilité considère $\mathscr{R}$, l'espace des *récompenses*, supposé complètement connu ; par exemple, $\mathscr{R} = \mathbb{R}$. Nous supposons aussi qu'*il est possible d'ordonner les récompenses*, donc qu'il existe un *ordre total*, noté $\preceq$, sur $\mathscr{R}$ tel que, si r_1 et r_2 sont dans $\mathscr{R}$,

(1) $r_1 \preceq r_2$ ou $r_2 \preceq r_1$; et
(2) si $r_1 \preceq r_2$ et $r_2 \preceq r_3$, alors $r_1 \preceq r_3$.

Ces deux propriétés paraissent être des conditions minimales dans un cadre décisionnel. En particulier, la *transitivité* (2) est absolument nécessaire pour permettre une comparaison entre les procédures de décision. Sinon, nous pouvons nous retrouver avec des cycles tels que $r_1 \preceq r_2 \preceq r_3 \preceq r_1$ et être dans l'incapacité de sélectionner la meilleure récompense parmi ces trois choix. (La Note 2.8.3 présente un critère qui est intransitif et ne se rapporte donc pas à la Théorie de la Décision.) Nous notons respectivement $\prec$ et $\sim$ l'ordre *strict* et la relation d'*équivalence* dérivés de $\preceq$. Cependant, une et seulement une des trois relations suivantes est satisfaite par tout couple (r_1, r_2) dans $\mathscr{R}^2$

$$r_1 \prec r_2, \qquad r_2 \prec r_1, \qquad r_1 \sim r_2.$$

Pour avancer davantage dans la construction de la fonction d'utilité, il est nécessaire d'étendre l'espace des récompenses de $\mathscr{R}$ à $\mathscr{P}$, l'espace des distributions de probabilité dans $\mathscr{R}$. Ceci permet aussi au décideur de prendre des décisions partiellement aléatoires ; de plus, l'espace des récompenses ainsi étendu est convexe.

Exemple 2.3. Dans toute situation réaliste, les récompenses associées à une action ne sont pas exactement connues au moment où la décision est prise ou, d'une façon équivalente, certaines décisions comportent une part de risque. Par exemple, en finance, le revenu financier $r \in \mathscr{R} = \mathbb{R}$ d'actions cotées en Bourse n'est pas garanti au moment où les actionnaires doivent déterminer les entreprises dont ils devront acheter des actions. Dans ce cas, $\mathscr{D} = \{d_1, \ldots, d_n\}$, où d_k représente l'action "acheter des actions de la compagnie k". Au moment de la décision, les gains associés aux différentes actions sont des dividendes aléatoires, connus seulement à la fin de l'année. ‖

La relation d'ordre $\preceq$ est supposée disponible également dans $\mathscr{P}$. Par exemple, quand la récompense est monétaire, la relation d'ordre dans $\mathscr{P}$ peut être obtenue en comparant la moyenne des rendements associés à la distribution P. Il est donc possible de comparer deux distributions de probabilité dans $\mathscr{R}$, P_1 et P_2. Nous supposons ainsi que $\preceq$ satisfait les extensions des deux hypothèses (1) et (2) sur $\mathscr{P}$:

(A_1) $P_1 \preceq P_2$ ou $P_2 \preceq P_1$; et

(A_2) si $P_1 \preceq P_2$ et $P_2 \preceq P_3$, alors $P_1 \preceq P_3$.

La relation d'ordre sur $\mathscr{R}$ apparaît alors comme un cas particulier d'ordre sur $\mathscr{P}$, via la considération des masses de Dirac δ_r ($r \in \mathscr{R}$).

L'existence de l'ordre $\preceq$ sur $\mathscr{P}$ est fondée sur l'hypothèse qu'il existe une récompense optimale, et donc qu'il existe au moins un ordre partiel sur les conséquences, même quand elles sont aléatoires. C'est évidemment le cas lorsqu'il existe une fonction U de $\mathscr{R}$ associée à $\preceq$, telle que $P_1 \preceq P_2$ est équivalente à

$$\mathbb{E}^{P_1}[U(r)] \leq \mathbb{E}^{P_2}[U(r)],$$

comme dans l'exemple monétaire ci-dessus. Cette fonction U est dite *fonction d'utilité*. Nous présentons maintenant un système axiomatique portant sur $\preceq$ qui assure l'existence de la fonction d'utilité.

Par souci de simplicité, nous considérons ici seulement le groupe des distributions *bornées*, $\mathscr{P}_{\mathscr{B}}$, correspondant aux distributions à support borné, pour lesquelles il existe r_1 et r_2 tels que

$$[r_1, r_2] = \{r : \ r_1 \preceq r \preceq r_2\} \quad \text{et} \quad P([r_1, r_2]) = 1.$$

Pour P_1, P_2 dans $\mathscr{P}_{\mathscr{B}}$, nous définirons le *mélange* $P = \alpha P_1 + (1-\alpha)P_2$ comme la distribution qui génère une récompense de P_1 avec probabilité α et une récompense de P_2 avec probabilité $(1-\alpha)$. Par exemple, $\alpha\delta_{r_1} + (1-\alpha)\delta_{r_2}$ est la distribution qui donne comme résultat la récompense r_1 avec probabilité α et la récompense r_2 avec probabilité $(1-\alpha)$. Deux hypothèses supplémentaires (ou axiomes) sont nécessaires pour obtenir l'existence d'une fonction d'utilité dans $\mathscr{R}$. Tout d'abord, il doit y avoir *respect de l'ordre sous des alternatives indifférentes* :

(A$_3$) si $P_1 \preceq P_2$, $\alpha P_1 + (1-\alpha)P \preceq \alpha P_2 + (1-\alpha)P$ pour tout $P \in \mathscr{P}$.

Par exemple, si les actionnaires de l'Exemple 2.3 peuvent comparer deux compagnies avec des distributions de dividendes P_1 et P_2, ils doivent pouvoir garder le même classement s'il y a une probabilité $(1-\alpha)$ que les deux dividendes soient remplacés par des bons du Trésor avec une distribution de dividendes P. La relation d'ordre doit être aussi *connexe* (ou *fermée*) :

(A$_4$) si $P_1 \preceq P_2 \preceq P_3$, il existe α et $\beta \in]0,1[$ tel que
$$\alpha P_1 + (1-\alpha)P_3 \preceq P_2 \preceq \beta P_1 + (1-\beta)P_3.$$

La dernière hypothèse implique alors le résultat suivant.

Lemme 2.4. *Si r_1, r_2, et r sont des récompenses dans $\mathscr{R}$ avec $r_1 \prec r_2$ et $r_1 \preceq r \preceq r_2$, il existe un seul v $(0 \le v \le 1)$ tel que $r \sim vr_1 + (1-v)r_2$.*

Le Lemme 2.4 est en réalité le point essentiel pour la construction de la *fonction d'utilité*, U, dans $\mathscr{R}$. En effet, pour r_1 et r_2, deux récompenses arbitraires telles que $r_2 \prec r_1$, nous pouvons définir U de la façon suivante. Pour chaque $r \in \mathscr{R}$, soit

(i) $U(r) = v$ si $r_2 \preceq r \preceq r_1$ et $r \sim vr_1 + (1-v)r_2$;
(ii) $U(r) = \frac{-v}{1-v}$ si $r \preceq r_2$ et $r_2 \sim vr_1 + (1-v)r$; et
(iii) $U(r) = \frac{1}{v}$ si $r_1 \preceq r$ et $r_1 \sim vr + (1-v)r_2$.

En particulier, $U(r_1) = 1$ et $U(r_2) = 0$. De plus, cette fonction U conserve la relation d'ordre sur $\mathscr{R}$ (voir DeGroot, 1970, p.105, pour une démonstration).

Lemme 2.5. *Si r_1, r_2 et r_3 sont trois récompenses dans $\mathscr{R}$ telles que $r_2 \sim \alpha r_1 + (1-\alpha)r_3$*

$$U(r_2) = \alpha U(r_1) + (1-\alpha)U(r_3).$$

En réalité, les axiomes (A_3) et (A_4) peuvent être davantage affaiblis. Il est effectivement suffisant qu'ils ne soient satisfaits que dans $\mathscr{R}$. L'extension de la définition de fonction d'utilité pour $\mathscr{P}_{\mathscr{B}}$ nécessite une hypothèse supplémentaire. Soit P tel que $P([r_1, r_2]) = 1$, définissons

$$\alpha(r) = \frac{U(r) - U(r_1)}{U(r_2) - U(r_1)}$$

et

$$\beta = \int_{[r_1, r_2]} \alpha(r)\, dP(r).$$

Alors, l'axiome additionnel

(A_5) $$P \sim \beta\delta_{r_2} + (1 - \beta)\delta_{r_1}$$

implique que, si r est équivalent à $\alpha(r)r_1 + (1-\alpha(r))r_2$ pour chaque $r \in [r_1, r_2]$, cette équivalence doit être vraie en moyenne. En effet, notons que β est obtenu à partir d'une utilité moyenne,

$$\beta = \frac{\mathbb{E}^P[U(r)] - U(r_1)}{U(r_2) - U(r_1)},$$

et cette hypothèse fournit une définition de U dans $\mathscr{P}_{\mathscr{B}}$. Comme dans le Lemme 2.5 où U est restreint à $\mathscr{R}$, et comme le montre le résultat suivant, l'axiome (A_5) indique que U permet une *linéarisation* (ou une paramétrisation linéaire) de la relation d'ordre $\preceq$ dans $\mathscr{P}_{\mathscr{B}}$. Bien que légèrement tautologique –puisqu'elle dépend dans sa formulation de la fonction d'utilité que nous essayons de construire–, (A_5) nous conduit effectivement à l'extension suivante du Lemme 2.5 à $\mathscr{P}_{\mathscr{B}}$.

Théorème 2.6. *Soient P_1 et P_2 sur $\mathscr{P}_{\mathscr{B}}$. Alors,*

$$P_1 \preceq P_2$$

si et seulement si

$$\mathbb{E}^{P_1}[U(r)] \leq \mathbb{E}^{P_2}[U(r)].$$

De plus, si U^* est une autre fonction d'utilité qui satisfait la relation d'équivalence présentée ci-dessus, il existe $a > 0$ et b tels que

$$U^*(r) = aU(r) + b.$$

Preuve. Soient r_1 et r_2 tels que

$$P_1([r_1, r_2]) = P_2([r_1, r_2]) = 1$$

(avec $r_1 \prec r_2$). Comme

$$P_1 \sim \frac{\mathbb{E}^{P_1}[U(r)] - U(r_1)}{U(r_2) - U(r_1)}\delta_{r_2} + \frac{U(r_2) - \mathbb{E}^{P_1}[U(r)]}{U(r_2) - U(r_1)}\delta_{r_1}$$

et

$$P_2 \sim \frac{\mathbb{E}^{P_2}[U(r)] - U(r_1)}{U(r_2) - U(r_1)}\delta_{r_2} + \frac{U(r_2) - \mathbb{E}^{P_2}[U(r)]}{U(r_2) - U(r_1)}\delta_{r_1},$$

$P_1 \preceq P_2$ est effectivement équivalent à

$$\frac{\mathbb{E}^{P_1}[U(r)] - U(r_1)}{U(r_2) - U(r_1)} \leq \frac{\mathbb{E}^{P_2}[U(r)] - U(r_1)}{U(r_2) - U(r_1)},$$

soit encore $\mathbb{E}^{P_1}[U(r)] \leq \mathbb{E}^{P_2}[U(r)]$. De plus, pour toute autre fonction d'utilité U^*, il existe a et b tels que $U^*(r_1) = aU(r_1) + b$, $U^*(r_2) = aU(r_2) + b$. L'extension de cette relation à chaque $r \in \mathscr{R}$ découle du Lemme 2.5. □

Notons que la construction ci-dessus n'implique aucune restriction sur la fonction U. Donc, celle-ci n'a pas besoin d'être bornée, bien que cette condition soit souvent mentionnée dans les livres. On peut avancer que cette généralité est artificielle et formelle, car les fonctions d'utilité subjectives sont toujours bornées. Par exemple, quand on considère une récompense monétaire, il existe un seuil psychologique, disons de 100 000 000 euros, au-dessus duquel (la plupart) des individus ont une fonction d'utilité presque constante.

Cependant, cette limite supérieure varie d'individu à individu, et la variation est encore plus grande entre des individus et des entreprises ou des États. Il est aussi important d'inclure les récompenses inacceptables, bien que l'hypothèse (A_4) empêche les récompenses d'utilité égale à $-\infty$. (Cette restriction implique que la mort d'un patient pendant une étude pharmaceutique ou un accident grave dans une centrale nucléaire ont une utilité finie.) De plus, la plupart des fonctions de coût théoriques ne sont pas bornées. Une contrepartie de cette généralité est que les résultats ci-dessus n'ont été établis que pour $\mathscr{P}_{\mathscr{B}}$. En réalité, ils peuvent être étendus à $\mathscr{P}_{\mathscr{E}}$, l'ensemble des distributions P dans $\mathscr{P}$ telles que $\mathbb{E}^P[U(r)]$ soit finie, sous l'hypothèse que (A_1)–(A_5) et deux conditions supplémentaires sont satisfaites par $\mathscr{P}_{\mathscr{E}}$ (voir l'Exercice 2.3).

Théorème 2.7. *Soient P et Q, deux distributions sur $\mathscr{P}_{\mathscr{E}}$. Alors, $P \preceq Q$ si et seulement si*

$$\mathbb{E}^P[U(r)] \leq \mathbb{E}^Q[U(r)].$$

Évidemment, le Théorème 2.7 ne parvient pas à traiter des distributions d'utilité infinies. Si de telles distributions existent, elles doivent être comparées entre elles et une fonction d'utilité doit être construite sur cette classe restreinte, puisque dans un sens il s'agit des seules distributions intéressantes.

Cependant, les fonctions de coût considérées par la suite seront minorées, le plus souvent par 0. Les fonctions d'utilité correspondantes, opposées aux fonctions de coût, sont donc toujours majorées et les paradoxes de récompense infinie peuvent être évités. (Rubin, 1984 et Fishburn, 1988, fournissent des

systèmes axiomatiques plus faibles assurant l'existence d'une fonction d'utilité.)

Plusieurs critiques ont été formulées, d'ordre théorique et psychologique, contre la notion de *rationalité des décideurs* et les axiomes associés (A_1)–(A_4). Premièrement, il paraît illusoire de croire que les individus peuvent comparer toutes les récompenses, c'est-à-dire qu'ils peuvent fournir un ordre total de $\mathscr{P}$ (ou même de $\mathscr{R}$), car leurs capacités de discernement sont forcément limitées, en particulier en ce qui concerne les alternatives contigües ou extrêmes.

L'hypothèse de *transitivité* est aussi trop forte, car les exemples en sport ou en politique montrent que l'ordre des préférences conduit souvent dans la pratique à une intransitivité, comme on le voit dans les paradoxes de *Condorcet et de Simpson* (voir Casella et Wells, 1993, et les Exercices 1.9 et 2.2). Plus fondamentalement, l'hypothèse que l'ordre peut être étendu de $\mathscr{R}$ à $\mathscr{P}$ a été fortement contestée, car elle implique que l'ordre social puisse être obtenu à partir d'un ensemble d'ordres individuels, et en général cela n'est pas possible (voir Arrow, 1956 ou Blyth, 1972a). Cependant, bien que reconnaissant ce fait, Rubin (1984) remarque que cette impossibilité implique simplement que l'utilité et l'a priori soient inséparables, non pas qu'une décision optimale (bayésienne) ne puisse pas être obtenue, et il donne une série restreinte d'axiomes se rapportant à cet objectif. En général, les critiques exprimées ci-dessus sont absolument valables, mais ne peuvent résister à l'argument de la nécessité absolue d'un cadre axiomatique qui valide la prise de décision dans un cadre incertain. Comme cela est déjà évoqué dans le Chapitre 1, la modélisation statistique *est et doit être* réductrice. Même si elle passe à côté d'une partie de la complexité du monde, cette représentation simplifiée du monde permet aux statisticiens et aux autres de prendre des décisions. La Théorie de la Décision décrit ainsi un cadre idéalisé, sous une rationalité fondamentale que les vrais décideurs n'arrivent pas à atteindre, mais qu'ils visent néanmoins[20].

D'un point de vue plus pratique, la construction de la fonction d'utilité décrite au-dessus peut être critiquée comme irréaliste. Berger (1985b) fournit quelques exemples fondés sur DeGroot (1970), où la fonction d'utilité est construite par des divisions successives de l'espace des récompenses (voir aussi Raiffa et Schlaifer, 1961). Cependant, si $\mathscr{R}$ est grand (par exemple, non dénombrable), U ne peut pas être évaluée pour chaque récompense r, même si la linéarité mise en avant dans le Lemme 2.5 permet des approximations quand $\mathscr{R} \subset \mathbb{R}$. Dans un cadre multidimensionnel, les approximations linéaires ne sont plus possibles, sauf si on utilise une combinaison linéaire d'utilités réelles, soit

[20] Pour emprunter à Smith (1984), critiquer les structures idéales de la Théorie de la Décision à cause des limitations humaines revient, d'une façon ou d'une autre, à remettre en cause l'intégration parce que quelques intégrales ne peuvent être résolues que numériquement.

$$U(r_1, r_2, \ldots, r_n) = \sum_{i=1}^{n} \alpha_i U_i(r_i)$$

(voir Raiffa, 1968, Keeney et Raiffa, 1976, ou Smith, 1988, pour une discussion). En général, les fonctions d'utilité pratiques ne seront que des approximations des vraies fonctions d'utilité.

Même quand la récompense est purement financière, une détermination rigoureuse de la fonction d'utilité s'impose, car U peut être loin d'être linéaire, en particulier pour de grandes récompenses. Cela signifie qu'un gain de \$3 000 avec une probabilité de 1/2 n'équivaut pas forcément à gagner \$1 500 sûrement. Pour résoudre ce paradoxe, Laplace (1795) introduit la notion d'*attente morale*, dérivée de la valeur relative d'une augmentation du gain *"la valeur absolue divisée par le gain total de la personne concernée"*. Laplace infère que l'attente morale *"coïncide avec l'attente mathématique quand le gain devient infini en comparaison avec les variations dues à l'incertitude"*, ce qui signifie que l'utilité n'est effectivement linéaire qu'au voisinage de 0. Sinon, les attitudes d'*aversion au risque* ralentissent la courbe d'utilité, qui est typiquement concave et majorée pour de grandes valeurs des récompenses. (Les personnes avec une fonction d'utilité convexe sont dites *amateurs de risque*, car elles préfèrent un gain aléatoire à l'espérance de ce gain. Notons que cette attitude est assez compréhensible au voisinage de 0.) Construire une fonction d'utilité monétaire est évidemment plus compliqué que d'utiliser une utilité linéaire, mais cette construction fournit une représentation plus précise de la réalité et peut même éviter des paradoxes comme celui présenté ci-dessous.

Exemple 2.8. (Paradoxe de Saint-Pétersbourg) Soit un jeu où une pièce est lancée jusqu'à ce que le côté *face* apparaisse. Quand cela arrive au n-ième jet, le gain du joueur est 3^n, ce qui donne un gain moyen de

$$\sum_{n=1}^{+\infty} 3^n \frac{1}{2^n} = +\infty.$$

Chaque joueur devrait donc être prêt à payer un droit d'entrée arbitrairement élevé pour jouer ce jeu, même s'il a moins de 0.05 de probabilité d'aller au-delà du cinquième jet ! Cette modélisation ne prend pas en compte le fait que la fortune d'un joueur est nécessairement bornée et qu'il ne peut jouer qu'un nombre limité de fois. Une solution à ce paradoxe est de substituer une fonction d'utilité bornée à la fonction d'utilité linéaire, comme

$$U(r) = \frac{r}{\delta + r} \qquad (\delta > 0,\ r > -\delta),$$

et $U(r) = -\infty$ sinon. Cette construction est assez similaire à l'attente morale de Laplace. Un droit d'entrée acceptable e sera alors tel que l'utilité moyenne du jeu est aussi grande que l'utilité de ne rien faire, soit

$$\mathbb{E}[U(r - e)] \geq U(0) = 0.$$

La Figure 2.1 représente l'utilité moyenne en fonction de δ, calculée par approximation numérique de la série

$$\sum_{n=1}^{+\infty} \frac{3^n}{\delta + 3^n} 2^{-n}.$$

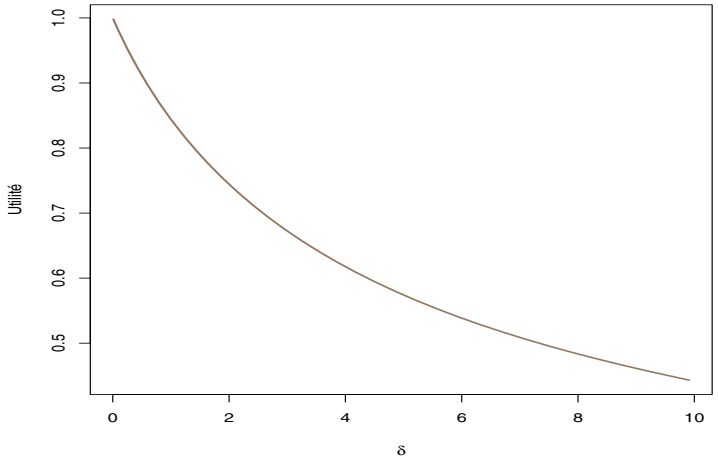

Fig. 2.1. Utilité moyenne pour le paradoxe de Saint-Pétersbourg.

Considérons maintenant une modification du jeu où le joueur peut se retirer à n'importe quel moment n et prendre le gain 3^n si le côté *pile* n'est pas encore apparu. Le gain moyen au temps n est alors

$$\frac{3^n}{\delta + 3^n} 2^{-n},$$

qui peut fournir un temps optimal n_0 pour quitter le jeu, dépendant du paramètre d'utilité δ, qui caractérise en quelque sorte l'*aversion au risque* du joueur (voir Smith, 1988, pour une description plus minutieuse). Par exemple, δ peut représenter la chance du joueur, car $U(\tau)$ tend vers $-\infty$ quand τ tend vers $-\delta$. Ce choix particulier de U peut bien sûr être critiqué, mais une représentation plus précise de la fonction d'utilité nécessite une analyse détaillée des motivations du joueur (voir aussi l'Exercice 2.9). ‖

Voir également Bernardo et Smith (1994) pour une analyse détaillée des bases de la théorie de l'utilité, avec une description particulière des *arbres de décision*.

2.3 Utilité et coût

Revenons à un cadre purement statistique. D'un point de vue décisionnel, le modèle statistique inclut maintenant trois espaces : $\mathscr{X}$, *espace* des observations, Θ, *espace* des paramètres, et $\mathscr{D}$, *espace* des décisions (ou *espace* d'action). L'inférence statistique consiste alors à prendre une décision $d \in \mathscr{D}$ par rapport au paramètre $\theta \in \Theta$, fondée sur l'observation $x \in \mathscr{X}$, x et θ étant reliés par la distribution $f(x|\theta)$. Dans la plupart des cas, la décision d devra évaluer (ou *estimer*) une fonction de θ, $h(\theta)$, le plus précisément possible. La Théorie de la Décision suppose de plus que chaque action d peut être évaluée (ce qui signifie que la précision peut être quantifiée) et conduit à une récompense r, avec une utilité $U(r)$ (qui existe sous l'hypothèse de rationalité des décideurs). Dorénavant, cette utilité sera notée $U(\theta, d)$ pour insister sur le fait qu'elle dépend uniquement de ces deux facteurs. Quand d'autres facteurs aléatoires r interviennent dans U, nous écrirons $U(\theta, d) = \mathbb{E}_{\theta,d}[U(r)]$. Donc, $U(\theta, d)$ peut être vue comme une mesure de proximité entre l'estimation proposée d et la vraie valeur $h(\theta)$.

Une fois que la fonction d'utilité a été construite (ou approchée), nous construisons la fonction de *coût* correspondante

$$\mathrm{L}(\theta, d) = -U(\theta, d).$$

En général, la fonction de coût est supposée positive, ce qui implique $U(\theta, d) \leq 0$, et donc il n'existe pas de décision ayant une utilité infinie. L'hypothèse de l'existence d'un minorant pour L peut être critiquée comme trop stricte, mais elle évite des paradoxes comme ceux mentionnés ci-dessus. On peut aussi soutenir que, d'un point de vue statistique, la fonction de coût L représente bien le *coût* (ou l'erreur) dû à une mauvaise évaluation de la fonction de θ d'intérêt, et donc que même la meilleure évaluation possible de cette fonction, soit, lorsque θ est connu, peut entraîner au mieux un coût nul. Dans le cas contraire, la fonction de coût perdrait sa continuité en $d = \theta$, ce qui pourrait même empêcher le choix d'une procédure de décision.

Évidemment, sauf pour les cas les plus triviaux, il est généralement impossible de minimiser uniformément (en d) la fonction de coût $\mathrm{L}(\theta, d)$ quand θ est inconnu. Pour obtenir un critère de comparaison utilisable à partir d'une fonction de coût dans un contexte aléatoire, l'approche *fréquentiste* propose de considérer plutôt le coût moyen (ou *risque fréquentiste*)

$$\begin{aligned} R(\theta, \delta) &= \mathbb{E}_\theta[\mathrm{L}(\theta, \delta(x))] \\ &= \int_{\mathscr{X}} \mathrm{L}(\theta, \delta(x)) f(x|\theta)\, dx, \end{aligned}$$

où $\delta(x)$ est la règle de décision, soit l'attribution d'une décision à chaque résultat $x \sim f(x|\theta)$ de l'expérience aléatoire. La fonction δ, de $\mathscr{X}$ dans $\mathscr{D}$, est habituellement appelée *estimateur* (tandis que la *valeur* $\delta(x)$ est appelée

estimation de θ). Quand il n'y a pas de risque de confusion, nous noterons aussi $\mathscr{D}$ l'ensemble des estimateurs.

Le paradigme fréquentiste repose sur cette notion pour comparer les estimateurs et, si possible, choisir le meilleur d'entre eux. Le raisonnement est que ces estimateurs sont évalués selon leurs performances à long terme pour toutes les valeurs possibles du paramètre θ. Notons cependant qu'il existe plusieurs difficultés liées à cette approche.

(1) L'erreur (coût) est moyennée sur toutes les valeurs de x, proportionnellement à la densité $f(x|\theta)$. Il semble donc que l'observation x ne soit plus prise en compte par la suite. Le critère de risque évalue les procédures selon leurs performances de long terme et non directement pour une observation x donnée. Une telle évaluation peut être satisfaisante pour un(e) statisticien(ne), mais elle n'est pas très convaincante pour un(e) client(e) qui cherche un résultat optimal pour ses données x, pas pour celles des autres !

(2) L'analyse fréquentiste du problème de décision suppose tacitement que le même problème sera rencontré de nombreuses fois pour que l'évaluation en fréquence ait un sens. En effet, $R(\theta, \delta)$ est approximativement le coût moyen sur les répétitions i.i.d. de la même expérience, selon la Loi des Grands Nombres. Cependant, d'un point de vue philosophique et pratique, il existe beaucoup de controverses sur la notion même de répétabilité des expériences (voir Jeffreys, 1961). En fait, si de nouvelles observations parviennent à un statisticien, celui-ci devrait les utiliser, ce qui pourrait modifier la façon dont l'expérience est conduite, comme par exemple dans les expériences médicales.

(3) Pour une procédure δ, le risque $R(\theta, \delta)$ est une *fonction* du paramètre θ. L'approche fréquentiste n'induit donc pas un ordre *total* sur l'ensemble des procédures. Il est généralement impossible de comparer les procédures de décision avec ces critères, car deux fonctions de risque qui se croisent empêchent la comparaison entre les estimateurs correspondants. Au mieux, on peut espérer trouver une procédure δ_0 qui minimise (en δ) uniformément (en θ) $R(\theta, \delta)$, mais ce type de situation se produit rarement, à moins que l'espace des procédures de décision ne soit très restreint. Les procédures optimales ne peuvent être obtenues que par une restriction plutôt artificielle à un ensemble de procédures autorisées.

Exemple 2.9. Soient x_1 et x_2, deux observations de

$$P_\theta(x = \theta - 1) = P_\theta(x = \theta + 1) = 0.5, \qquad \theta \in \mathbb{R}.$$

Le paramètre d'intérêt est θ (donc $\mathscr{D} = \Theta$) et il est estimé par δ sous le coût

$$\mathrm{L}(\theta, \delta) = 1 - \mathbb{I}_\theta(\delta),$$

appelée le plus souvent *coût* $0-1$, qui pénalise les erreurs d'estimation, quelle que soit leur magnitude, par 1. Considérons en particulier l'estimateur

$$\delta_0(x_1, x_2) = \frac{x_1 + x_2}{2},$$

dont la fonction de risque est

$$\begin{aligned} R(\theta, \delta_0) &= 1 - P_\theta(\delta_0(x_1, x_2) = \theta) \\ &= 1 - P_\theta(x_1 \neq x_2) = 0.5. \end{aligned}$$

Ce calcul montre que l'estimateur δ_0 est correct la moitié du temps. En réalité, cet estimateur est toujours correct quand $x_1 \neq x_2$, et toujours faux autrement. Cependant l'estimateur $\delta_1(x_1, x_2) = x_1 + 1$ a aussi une fonction de risque égale à 0.5, comme $\delta_2(x_1, x_2) = x_2 - 1$. Donc, δ_0, δ_1 et δ_2 ne peuvent pas être classés sous le coût $0 - 1$. ‖

En revanche, l'approche bayésienne de la Théorie de la Décision intègre sur l'espace Θ, car θ est inconnu, plutôt que de le faire sur l'espace $\mathscr{X}$, x étant connu. Il est fondé sur le *coût moyenne a posteriori*

$$\varrho(\pi, d|x) = \mathbb{E}^\pi[L(\theta, d)|x] = \int_\Theta \mathrm{L}(\theta, d)\pi(\theta|x)\, d\theta\,,$$

qui moyenne l'erreur (c'est-à-dire le coût) selon la distribution a posteriori du paramètre θ, *conditionnellement à la valeur observée* x. Pour un x donné, l'erreur moyenne résultant de la décision d est en réalité $\varrho(\pi, d|x)$. Le coût moyen a posteriori est ainsi une fonction de x mais cette dépendance n'est pas gênante, contrairement à la dépendance fréquentiste du risque au paramètre puisque x, à la différence de θ, est connu.

En se donnant une distribution a priori π, il est aussi possible de définir le *risque intégré*, qui est le risque fréquentiste moyenné sur les valeurs de θ selon leur distribution a priori

$$\begin{aligned} r(\pi, \delta) &= \mathbb{E}^\pi[R(\theta, \delta)] \\ &= \int_\Theta \int_\mathscr{X} \mathrm{L}(\theta, \delta(x))\, f(x|\theta)\, dx\ \pi(\theta)\, d\theta. \end{aligned}$$

Un intérêt particulier de ce deuxième concept est qu'il associe un nombre réel à chaque estimateur, et non une fonction de θ. Il induit donc un ordre total sur l'ensemble des estimateurs et permet une comparaison directe entre ces estimateurs. Cela implique que, quoique prenant en compte l'information a priori via la distribution a priori, l'approche bayésienne est suffisamment *réductrice* (dans un sens positif) pour atteindre une décision efficace. De plus, les deux notions ci-dessus sont équivalentes puisqu'elles conduisent à la même décision.

Théorème 2.10. *Un estimateur minimisant le risque intégré $r(\pi, \delta)$ est obtenu par sélection, pour chaque $x \in \mathscr{X}$, de la valeur $\delta(x)$ qui minimise le coût moyen a posteriori, $\varrho(\pi, \delta|x)$, puisque*

$$r(\pi, \delta) = \int_{\mathscr{X}} \varrho(\pi, \delta(x)|x) m(x)\, dx. \tag{2.1}$$

Preuve. L'égalité (2.1) découle directement du Théorème de Fubini, car, comme $\mathrm{L}(\theta, \delta) \geq 0$,

$$\begin{aligned} r(\pi, \delta) &= \int_{\Theta} \int_{\mathscr{X}} \mathrm{L}(\theta, \delta(x)) f(x|\theta)\, dx\ \pi(\theta)\, d\theta \\ &= \int_{\mathscr{X}} \int_{\Theta} \mathrm{L}(\theta, \delta(x)) f(x|\theta) \pi(\theta)\, d\theta\, dx \\ &= \int_{\mathscr{X}} \int_{\Theta} \mathrm{L}(\theta, \delta(x)) \pi(\theta|x)\, d\theta\ m(x)\, dx\ . \end{aligned}$$

□

Ce résultat mène à la définition suivante d'un estimateur de Bayes.

Définition 2.11. *Un* estimateur de Bayes *associé à une distribution a priori* π *et une fonction de coût* L *est un estimateur* δ^π *minimisant* $r(\pi, \delta)$*. Pour chaque* $x \in \mathscr{X}$*, ce dernier est donné par*

$$\delta^\pi(x) = \arg\min_d \varrho(\pi, d|x)\,.$$

La valeur $r(\pi) = r(\pi, \delta^\pi)$ *est alors appelée* risque de Bayes.

Le Théorème 2.10 fournit ainsi un outil constructif pour la détermination des estimateurs de Bayes. Notons que, d'un point de vue strictement bayésien, seul le coût moyen a posteriori $\varrho(\pi, \delta|x)$ compte, puisque le paradigme bayésien est fondé sur une approche conditionnelle. Faire la moyenne sur toutes les valeurs possibles de x, alors que nous connaissons la valeur observée de x, semble être une perte d'information. Néanmoins, l'équivalence présentée par le Théorème 2.10 est importante parce que, premièrement, elle montre que l'approche conditionnelle n'est pas nécessairement aussi dangereuse que les critiques fréquentistes peuvent l'indiquer. En effet, bien que l'approche bayésienne fonctionne de façon conditionnelle à l'observation présente x, elle inclut aussi les propriétés probabilistes de la distribution de l'observation, $f(x|\theta)$. Deuxièmement, cette équivalence fournit une connection entre les résultats classiques de la Théorie des Jeux (voir la Section 2.4) et l'approche axiomatique bayésienne, fondée sur la distribution a posteriori. Ceci explique aussi pourquoi les estimateurs de Bayes jouent un rôle important pour les critères d'optimalité fréquentistes.

Le résultat présenté ci-dessus est valable pour des a priori propres et impropres, *du moment que le risque de Bayes* $r(\pi)$ *est fini*. Dans le cas contraire, la notion d'estimateur (décisionnel) de Bayes est affaiblie. Nous définissons alors un *estimateur de Bayes généralisé* comme la quantité minimisant, pour

chaque x, le coût moyen a posteriori. En terme d'optimalité fréquentiste, nous verrons que la distinction entre a priori propres et impropres est beaucoup moins importante que celle entre estimateurs de Bayes réguliers et généralisés, puisque les premiers sont admissibles. Notons que, pour des fonctions de coût strictement convexes, les estimateurs de Bayes sont uniques.

Nous terminons cette partie par un exemple de construction d'une fonction de coût dans un cadre de calibrage d'expert. Les références dans ce domaine sont DeGroot et Fienberg (1983), Murphy (1984), Bayarri et DeGroot (1988) et Schervish (1989). Smith (1988) montre aussi comment l'évaluation d'un prévisionniste peut aider à améliorer l'estimation des probabilités a priori. Voir la Note 2.8.1 pour une illustration différente en traitement d'image.

Exemple 2.12. Les prévisions météorologiques aux États-Unis sont souvent données sous la forme de probabilités. Par exemple *"la probabilité de pluie pour demain est estimée à 0.4"*. De telles prévisions étant quantifiées, il est intéressant (pour leurs employeurs autant que pour les utilisateurs) d'évaluer les météorologistes à travers une fonction de coût.

Pour un météorologiste donné, soit N le nombre des différents pourcentages annoncés au moins une fois par an et soient p_i $(1 \leq i \leq N)$ les différents pourcentages. Par exemple, nous pouvons avoir $N = 5$ et

$$p_1 = 0, \quad p_2 = 0.45, \quad p_3 = 0.7, \quad p_4 = 0.9, \quad \text{et} \quad p_5 = 0.95.$$

Dans ce cas, on observe effectivement les paramètres θ_i, soit

$$\theta_i = \frac{\text{nombre de jours pluvieux pour lesquels } p_i \text{ est prédite}}{\text{nombre de jours pour lesquels } p_i \text{ est prédite}}$$

(plus exactement, ce rapport est une bonne approximation de θ_i).

Si q_i indique la proportion de jours où p_i est prédite, une fonction de coût possible pour les experts est

$$\mathrm{L}(\theta, p) = \sum_{i=1}^{N} q_i (p_i - \theta_i)^2 + \sum_{i=1}^{N} q_i \log(q_i).$$

Pour un ensemble donné de θ_i $(1 \leq i \leq N)$, le meilleur météorologiste est celui qui est parfaitement calibré, donc celui qui satisfait $p_i = \theta_i$ $(1 \leq i \leq N)$. De plus, parmi ces météorologistes parfaits, le meilleur est le mieux équiréparti, satisfaisant $q_i = 1/N$ $(1 \leq i \leq N)$, c'est-à-dire le météorologiste le plus audacieux, par opposition à celui qui veut donner toujours le même pronostic p_{i_0}, par conséquence du terme d'*entropie*, $\mathfrak{E}(\mathbf{q}) = \sum_i q_i \log(q_i)$. Cependant, la distance $(p_i - \theta_i)^2$ peut être remplacée par n'importe quelle autre fonction prenant son minimum en $p_i = \theta_i$ (voir les Exercices 2.12 et 2.14). Le poids q_i dans la première somme est aussi utilisé pour calibrer plus efficacement les météorologistes, afin de prévenir la sur pénalisation de prévisions plus rares.

Ce coût a été construit avec un biais en faveur des experts utilisant un grand N, car l'entropie $\mathfrak{E}(N)$ augmente avec N. Cependant, une meilleure performance pour un plus grand N nécessite que p_i soit (presque) égal à θ_i et que q_i soit proche de $1/N$. ‖

2.4 Deux optimalités : minimaxité et admissibilité

Cette section est consacrée à deux notions fondamentales de la Théorie de la Décision fréquentiste, présentées par Wald (1950) et Neyman et Pearson (1933a,b). Comme il a été mentionné auparavant, et contrairement à l'approche bayésienne, le paradigme fréquentiste n'est pas assez réducteur pour conduire à un seul estimateur optimal. Bien que dans ce livre nous nous intéressions surtout aux aspects bayésiens de la Théorie de la Décision, il est nécessaire malgré tout d'étudier ces notions fréquentistes en détail, parce qu'elles montrent que les estimateurs de Bayes sont souvent optimaux pour les concepts fréquentistes d'optimalité et devraient donc être utilisés même lorsque l'information a priori est omise. En d'autres termes, on peut refuser le paradigme bayésien et ignorer la signification d'une distribution a priori, tout en obtenant malgré tout des estimateurs corrects d'un point de vue fréquentiste par l'utilisation de cette distribution a priori. Donc, dans ce sens technique, les fréquentistes devraient aussi prendre en compte l'approche bayésienne, car elle fournit un *outil* pour la construction d'estimateurs optimaux (voir Brown, 1971, 2000, Strawderman, 1971, Berger, 1985b, ou Berger et Robert, 1990, pour des exemples). De plus, ces propriétés peuvent être utiles pour la sélection d'une distribution a priori, quand l'information a priori n'est pas suffisamment précise pour conduire à une distribution a priori unique (Chapitre 3).

2.4.1 Estimateurs randomisés

De même que pour l'étude de la fonction d'utilité, où nous étendons l'espace de récompense de $\mathscr{R}$ à $\mathscr{P}$, nous avons besoin d'étendre aussi l'espace de décision à l'ensemble des *estimateurs randomisés*, prenant leurs valeurs dans $\mathscr{D}^*$, l'espace des distributions de probabilité sur $\mathscr{D}$. Utiliser un estimateur randomisé δ^* signifie que l'action est générée selon la distribution de densité de probabilité $\delta^*(x,.)$, une fois que l'observation x a été recueillie. Le coût de l'estimateur randomisé δ^* est alors défini comme le coût moyen

$$\mathrm{L}(\theta, \delta^*(x)) = \int_{\mathscr{D}} \mathrm{L}(\theta, a)\delta^*(x, a)\, da.$$

Cette extension est nécessaire au traitement des notions de minimaxité et d'admissibilité. Évidemment, de tels estimateurs n'en sont pas moins à prescrire, en particulier parce qu'ils contredisent le principe de vraisemblance, en

donnant plusieurs réponses possibles pour la même valeur de x (et donc de $\ell(\theta|x)$). De plus, il semble assez paradoxal d'ajouter du bruit au phénomène étudié pour prendre une décision dans l'incertain !

Exemple 2.13. (Suite de l'Exemple 2.9) Soit l'estimateur randomisé

$$\delta^*(x_1,x_2)(t) = \begin{cases} \mathbb{I}_{(x_1+x_2)/2}(t) & \text{si } x_1 \neq x_2, \\ [\mathbb{I}_{(x_1-1)}(t) + \mathbb{I}_{(x_1+1)}(t)]/2 & \text{sinon,} \end{cases}$$

où $\mathbb{I}_v$ est la masse de Dirac sur v. En réalité, si $x_1 = x_2$, les deux valeurs $\theta_1 = x_1 - 1$ et $\theta_2 = x_1 + 1$ ont la même vraisemblance. Comparé avec δ_0 qui n'estime jamais correctement θ lorsque $x_1 = x_2$, δ^* est exact avec une probabilité de 1/2. Cependant, quand δ^* n'estime pas correctement θ, il est plus loin de θ que δ_0. Le choix de l'estimateur dépend alors de la fonction de coût, donc de la manière dont la distance (ou l'erreur) entre l'estimateur et le paramètre θ est mesurée. ‖

D'un point de vue fréquentiste, les estimateurs randomisés sont néanmoins nécessaires, par exemple pour la théorie des tests fréquentistes, car ils permettent d'obtenir des niveaux de confiance qui ne peuvent être atteints autrement (voir le Chapitre 5). L'ensemble $\mathscr{D}^*$ apparaît ainsi comme une complétion topologique de $\mathscr{D}$. Cependant, cette modification de l'espace de décision ne modifie aucunement les réponses bayésiennes, comme le montre le résultat suivant (où l'ensemble des fonctions prenant leurs valeurs dans $\mathscr{D}^*$ est aussi noté $\mathscr{D}^*$).

Théorème 2.14. *Pour toute distribution a priori π sur Θ, le risque de Bayes pour l'ensemble des estimateurs randomisés est le même que celui pour l'ensemble des estimateurs non randomisés, soit*

$$\inf_{\delta\in\mathscr{D}} r(\pi,\delta) = \inf_{\delta^*\in\mathscr{D}^*} r(\pi,\delta^*) = r(\pi).$$

Preuve. Pour tout $x \in \mathscr{X}$ et tout $\delta^* \in \mathscr{D}^*$, nous avons

$$\begin{aligned} &\int_\Theta \int_{\mathscr{D}} L(\theta,a)\delta^*(x,a)da\,\pi(\theta|x)d\theta \\ &\quad = \int_{\mathscr{D}} \int_\Theta L(\theta,a)\pi(\theta|x)d\theta\,\delta^*(x,a)da \\ &\quad \geq \int_{\mathscr{D}} \inf_a \left\{ \int_\Theta L(\theta,a)\pi(\theta|x)d\theta \right\} \delta^*(x,a)da \\ &\quad = \varrho(\pi,\delta^\pi|x). \end{aligned}$$

□

Ce résultat reste vrai même quand le risque de Bayes $r(\pi)$ est infini. La démonstration est fondée sur le fait qu'une procédure randomisée moyennise le risque des estimateurs non randomisés et ne peut ainsi faire mieux que ces derniers. Cependant, le fait qu'utiliser des procédures randomisées n'a pas de sens n'est pas pris en compte par le risque fréquentiste à moins que certaines conditions, comme la convexité, ne soient imposées à la fonction de coût.

2.4.2 Minimaxité

Le critère de minimaxité que nous présentons maintenant apparaît comme une assurance contre le pire, car il vise à minimiser le coût moyen dans le cas le moins favorable. Il représente aussi un effort fréquentiste pour éviter de recourir au paradigme bayésien, tout en engendrant un ordre (faible) sur $\mathscr{D}^*$.

Définition 2.15. *On appelle* risque minimax *associé à la fonction de coût* L *la valeur*

$$\bar{R} = \inf_{\delta \in \mathscr{D}^*} \sup_{\theta} R(\theta, \delta) = \inf_{\delta \in \mathscr{D}^*} \sup_{\theta} \mathbb{E}_\theta[L(\theta, \delta(x))],$$

et estimateur minimax *tout estimateur (éventuellement randomisé)* δ_0 *tel que*

$$\sup_{\theta} R(\theta, \delta_0) = \bar{R}.$$

Cette notion est validée par la Théorie des Jeux, où deux adversaires ("le statisticien" et la "Nature") s'affrontent. Une fois que le statisticien a choisi une procédure, la Nature choisit l'état de la nature (c'est-à-dire le paramètre θ) qui maximise l'erreur du statisticien. (Nous verrons ci-après que ce choix est en général équivalent à celui de la distribution a priori π. L'approche bayésienne n'entre donc pas dans ce cadre conflictuel, car la distribution a priori est aussi supposée connue.) En général, cette perspective antagoniste apparaît comme regrettable dans une analyse statistique. En effet, considérer la Nature (ou la réalité) comme un ennemi ne peut que biaiser vers les pires cas et empêcher le statisticien d'utiliser l'information disponible (pour une analyse et une défense de la minimaxité, voir Brown, 1993, et Strawderman, 2000.)

La notion de minimaxité fournit une bonne illustration des aspects conservateurs du paradigme fréquentiste. Puisque cette approche refuse de faire la moindre hypothèse sur le paramètre θ, elle doit considérer les pires cas comme également probables et nécessite alors de se fixer sur le risque maximal. En réalité, d'un point de vue bayésien, cela équivaut souvent à prendre une distribution a priori concentrée sur ces pires cas (voir la Section 2.4.3). Dans la plupart des cas, ce point de vue est trop conservateur parce que certaines valeurs du paramètre sont moins vraisemblables que d'autres.

Exemple 2.16. Les premières plates-formes pétrolières en mer du Nord ont été construites selon un principe de minimaxité. En effet, elles étaient supposées résister à l'action conjuguée des plus fortes houles et des plus fortes tempêtes jamais observées, sous une température minimale record. Cette stratégie donne évidemment une marge confortable de sécurité, mais elle est très coûteuse. Pour des plates-formes plus récentes, les ingénieurs ont pris en compte la distribution de ces phénomènes climatiques afin de réduire les coûts de construction. ‖

Exemple 2.17. Une file d'attente à un feu rouge est en général correctement représentée par une loi de Poisson. Le nombre de voitures qui arrivent durant le temps d'observation, N, est donc distribué selon $\mathscr{P}(\lambda)$, avec un paramètre de moyenne λ devant être estimé. Évidemment, les valeurs de λ au-dessus d'une certaine limite sont assez invraisemblables. Par exemple, si λ_0 est le nombre de voitures dans toute la ville, le nombre moyen de voitures qui attendent à un feu n'excédera pas λ_0. Cependant, il peut arriver que certains estimateurs ne soient pas minimax parce que leur risque dépasse $\bar{R}$ pour les plus grandes valeurs de λ. ‖

L'exemple ci-dessus n'est pas forcément une critique du principe minimax, mais illustre plutôt le fait qu'une certaine information résiduelle est disponible dans la plupart des problèmes et pourrait être utilisée, même de manière marginale. De la même façon, l'Exemple 2.18 exhibe deux estimateurs, δ_1 et δ_2, tels que δ_1 a un risque minimax constant de $\bar{R}$ et δ_2 a un risque qui peut être aussi bas que $\bar{R}/10$ mais dépasse légèrement $\bar{R}$ pour des valeurs plus larges du paramètre (voir la Figure 2.2). Donc, selon le principe minimax, δ_1 devrait être préféré à δ_2, même si les valeurs de θ pour lesquelles δ_1 domine δ_2 sont plus invraisemblables (voir l'Exercice 2.28 pour un autre exemple frappant).

Exemple 2.18. Pour des raisons exposées dans la Note 2.8.2, nous considérons l'estimateur suivant

$$\delta_2(x) = \begin{cases} \left(1 - \dfrac{2p-1}{||x||^2}\right) x & \text{si } ||x||^2 \geq 2p-1, \\ 0 & \text{sinon,} \end{cases}$$

pour estimer $\theta \in \mathbb{R}^p$ quand $x \sim \mathscr{N}_p(\theta, I_p)$. Cet estimateur, dit *partie positive de l'estimateur de James-Stein*, est évalué sous le *coût quadratrique*,

$$\mathrm{L}(\theta, d) = ||\theta - d||^2.$$

La Figure 2.2 donne une comparaison des fonctions de risque respectives de δ_2 et $\delta_1(x) = x$, estimateur du maximum de vraisemblance, pour $p = 10$. Cette figure montre que δ_2 ne peut pas être minimax, car le risque maximum de δ_2 est

supérieur au risque (constant) de δ_1, c'est-à-dire $R(\theta, \delta_2) = \mathbb{E}_\theta[||\theta - \delta_2(x)||^2] = p$. (Nous montrerons dans la Section 2.4.3 que δ_1 est en effet un estimateur minimax dans ce cas.) Mais l'estimateur δ_2 est clairement supérieur dans la partie la plus intéressante de l'espace des paramètres, le coût supplémentaire étant d'ailleurs relativement minime. ‖

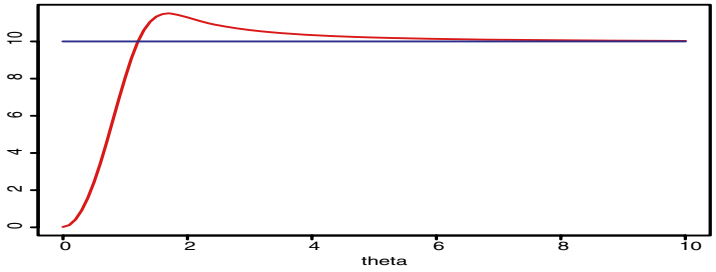

Fig. 2.2. Comparaison des risques des estimateurs δ_1 et δ_2.

Les divergences entre l'analyse bayésienne et l'analyse minimax sont illustrées par l'exemple suivant, emprunté à la Théorie des Jeux (puisqu'il n'y a ni observation ni modèle statistique).

Exemple 2.19. Deux personnes, A et B, suspectées d'être complices d'un cambriolage, sont arrêtées et placées dans des cellules séparées. Les deux suspects ont été interrogés et on leur a suggéré d'avouer le cambriolage. Bien qu'ils ne puissent pas être condamnés sans que l'un d'entre eux ait avoué, celui qui avoue le premier verra sa peine réduite. Le Tableau 2.1 fournit la perception de la récompense selon A (en années de liberté), où a_1 (resp. θ_1) représente le fait que A (resp. B) avoue. Les deux suspects ont un gain maximal s'ils se taisent tous les deux. Cependant, du point de vue de A, la stratégie minimax d'être le premier à parler, soit donc a_1, puisque $\max_\theta R(a_1, \theta) = 4$ et $\max_\theta R(a_2, \theta) = 10$. Par conséquent, les deux cambrioleurs se retrouveront en prison !

Tab. 2.1. Fonction d'utilité $U(\theta_i, a_j)$.

	a_1	a_2
θ_1	-4	-10
θ_2	8	30

Au contraire, si π est la probabilité (subjective) que A associe à l'événement *"B parle"*, soit, à θ_1, le risque de Bayes de a_1 est

$$r(\pi, a_1) = \mathbb{E}^\pi[-U(\theta, a_1)] = 4\pi - 8(1-\pi) = 12\pi - 8$$

et pour a_2,

$$r(\pi, a_2) = \mathbb{E}^\pi[-U(\theta, a_2)] = 10\pi - 30(1-\pi) = 40\pi - 30\,.$$

On vérifie simplement que, pour $\pi \leq 11/14$, $r(\pi, a_2)$ est plus petit que $r(\pi, a_1)$. Par conséquent, à moins que A ne soit persuadé que B va parler, il vaut mieux pour A ne rien dire. ‖

2.4.3 Existence d'une règle minimax et d'une stratégie maximin

Une difficulté importante liée à la notion de minimaxité est que les estimateurs minimax n'existent pas nécessairement. Ferguson (1967) et Berger (1985b, Chapitre 5) donnent des conditions suffisantes. En particulier, il existe une stratégie minimax quand Θ est fini et la fonction de coût est continue. Plus généralement, Brown (1976) (voir aussi Le Cam, 1986, et Strasser, 1985) considère l'espace de décision $\mathscr{D}$ comme plongé dans un autre espace de manière telle que l'ensemble des fonctions de risque sur $\mathscr{D}$ est compact dans ce grand espace. Dans cette perspective et sous des hypothèses supplémentaires, il est alors possible de construire des estimateurs minimax lorsque la fonction de coût est continue. Cependant, ces extensions impliquent l'utilisation de techniques topologiques trop avancées pour être considérées dans cet ouvrage. Par conséquent, nous ne donnerons ici que le résultat suivant (voir Blackwell et Girshick, 1954, pour une démonstration).

Théorème 2.20. *Si $\mathscr{D} \subset \mathbb{R}^k$ est convexe et compact et si $\mathrm{L}(\theta, d)$ est continue et convexe en tant que fonction de d, pour chaque $\theta \in \Theta$, il existe un estimateur minimax non randomisé.*

La restriction à des estimateurs non randomisés découle de *l'inégalité de Jensen*, puisque, lorsque la fonction de coût est convexe,

$$\mathrm{L}(\theta, \delta^*) = \mathbb{E}^{\delta^*}[\mathrm{L}(\theta, \delta)] \geq \mathrm{L}(\theta, \mathbb{E}^{\delta^*}(\delta)).$$

Ce résultat est un cas particulier du *théorème de Rao-Blackwell* (voir Lehmann et Casella, 1998).

Exemple 2.21. (Suite de l'Exemple 2.13) L'estimateur randomisé δ^* est uniformément dominé pour toute fonction de coût convexe par l'estimateur non randomisé $\mathbb{E}^{\delta^*}[\delta^*(x_1, x_2)]$, soit

$$\tilde{\delta}(x_1, x_2) = \begin{cases} \frac{1}{2}(x_1 + x_2) & \text{si } x_1 \neq x_2, \\ \frac{1}{2}(x_1 - 1) + \frac{1}{2}(x_1 + 1) = x_1 & \text{sinon,} \end{cases}$$

qui est en fait identique à l'estimateur δ_0 considéré initialement. Notons que cela n'est pas vrai pour le coût $0-1$, pour lequel δ^* domine $\tilde{\delta}$. ‖

Le résultat suivant met en avant la connexion entre approche bayésienne et principe minimax, dont la démonstration est immédiate.

Lemme 2.22. *Le risque de Bayes est toujours plus petit que le risque minimax,*

$$\underline{R} = \sup_{\pi} r(\pi) = \sup_{\pi} \inf_{\delta \in \mathscr{D}} r(\pi, \delta) \leq \bar{R} = \inf_{\delta \in \mathscr{D}^*} \sup_{\theta} R(\theta, \delta).$$

La première valeur est dite *risque maximin* et une distribution π^* telle que $r(\pi^*) = \underline{R}$ est appelée *distribution a priori la moins favorable*, quand de telles distributions existent. En général, la borne supérieure $r(\pi^*)$ est atteinte plutôt par une distribution impropre pouvant s'exprimer comme une limite de distributions a priori propres π_n. Mais ce phénomène n'empêche pas nécessairement la construction d'estimateurs minimax (voir le Lemme 2.27). Quand elles existent, les distributions les moins favorables sont celles qui ont le risque de Bayes le plus grand, donc aussi les moins intéressantes en terme de coût lorsqu'elles ne sont pas suggérées par l'information a priori disponible. Le résultat ci-dessus est assez logique au sens où l'information a priori ne peut qu'améliorer l'erreur d'estimation, même dans le pire des cas.

Un cas particulièrement intéressant correspond à la définition suivante.

Définition 2.23. *Un problème d'estimation est dit* admettre une valeur *si* $\underline{R} = \bar{R}$*, c'est-à-dire quand*

$$\sup_{\pi} \inf_{\delta \in \mathscr{D}} r(\pi, \delta) = \inf_{\delta \in \mathscr{D}^*} \sup_{\theta} R(\theta, \delta).$$

Quand le problème admet une valeur, certains estimateurs minimax sont des estimateurs de Bayes correspondant aux lois a priori les moins favorables. Cependant, ils peuvent être randomisés comme le démontre l'exemple suivant. Par conséquent le principe minimax ne fournit pas toujours des estimateurs acceptables.

Exemple 2.24. Soit[21] une observation de Bernoulli, $x \sim \mathscr{B}e(\theta)$ avec $\theta \in \{0.1, 0.5\}$. Quatre estimateurs non randomisés sont disponibles,

$$\delta_1(x) = 0.1, \qquad \delta_2(x) = 0.5,$$
$$\delta_3(x) = 0.1\,\mathbb{I}_{x=0} + 0.5\,\mathbb{I}_{x=1}, \quad \delta_4(x) = 0.5\,\mathbb{I}_{x=0} + 0.1\,\mathbb{I}_{x=1}.$$

Nous supposons de plus que la pénalité pour une réponse incorrecte est 2 quand $\theta = 0.1$ et 1 quand $\theta = 0.5$. Les *vecteurs de risque* $(R(0.1, \delta), R(0.5, \delta))$ des quatre estimateurs sont alors, respectivement, $(0, 1)$, $(2, 0)$, $(0.2, 0.5)$, et $(1.8, 0.5)$. Il est simple de voir que le vecteur de risque de chaque estimateur randomisé est une combinaison convexe de ces quatre estimateurs ou, d'une

[21] Les calculs dans cet exemple sont assez simples. Si besoin, voir le Chapitre 8 pour les détails.

façon équivalente, que l'*ensemble de risques*, $\mathscr{R}$, est l'enveloppe convexe des quatre vecteurs ci-dessus, comme le représente la Figure 2.3.

Dans ce cas, l'estimateur minimax est obtenu à l'intersection de la diagonale de $\mathbb{R}^2$ avec la frontière inférieure de $\mathscr{R}$. Comme le montre la Figure 2.3, cet estimateur δ^* est randomisé et prend la valeur $\delta_3(x)$ avec une probabilité $\alpha = 0.87$ et $\delta_2(x)$ avec une probabilité $1 - \alpha$. Le poids α est en effet obtenu par l'équation

$$0.2\alpha + 2(1 - \alpha) = 0.5\alpha.$$

Cet estimateur δ^* est aussi un *estimateur (randomisé) de Bayes* pour la loi a priori

$$\pi(\theta) = 0.22\, \mathbb{I}_{0.1}(\theta) + 0.78\, \mathbb{I}_{0.5}(\theta);$$

la probabilité a priori $\pi_1 = 0.22$ correspond à la pente entre $(0.2, 0.5)$ et $(2, 0)$, soit,

$$\frac{\pi_1}{1 - \pi_1} = \frac{0.5}{2 - 0.2}.$$

Notons que tout estimateur randomisé qui est une combinaison de δ_2 et de δ_3 est un estimateur de Bayes pour cette distribution, mais que seul δ^* est aussi un estimateur minimax. ‖

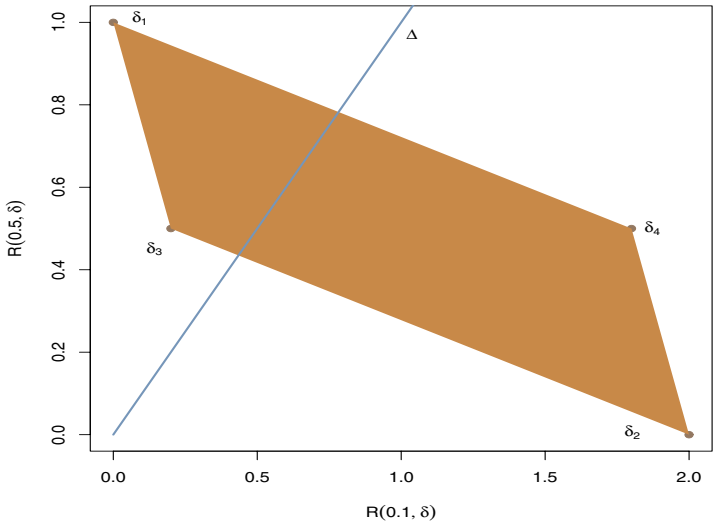

Fig. 2.3. Ensemble de risques pour l'estimation du paramètre de la distribution de Bernoulli et diagonale Δ.

À l'instar des estimateurs minimax, une distribution la moins favorable n'existe pas forcément, car son existence dépend d'un théorème d'hyperplan séparateur qui n'est pas toujours vérifié (voir Pierce, 1973, Brown, 1976, Berger, 1985b, et le Chapitre 8). De plus, Strawderman (1971) montre que, dans le cas particulier où $x \sim \mathcal{N}_p(\theta, I_p)$, il n'existe pas d'estimateur de Bayes régulier qui soit minimax lorsque $p \leq 4$.

D'un point de vue plus pratique, le Lemme 2.22 fournit des conditions suffisantes de minimaxité.

Lemme 2.25. *Si δ_0 est un estimateur de Bayes pour π_0 et si $R(\theta, \delta_0) \leq r(\pi_0)$ pour tout θ dans le support de π_0, δ_0 est minimax et π_0 est la distribution la moins favorable.*

Exemple 2.26. (Berger, 1985b) Soit $x \sim \mathcal{B}(n, \theta)$ où θ est à estimer sous un coût quadratique,

$$\mathrm{L}(\theta, \delta) = (\delta - \theta)^2 .$$

L'estimateur de Bayes est alors donné par les espérances a posteriori (voir la Section 2.5) et quand $\theta \sim \mathcal{B}e\left(\frac{\sqrt{n}}{2}, \frac{\sqrt{n}}{2}\right)$, l'espérance a posteriori est

$$\delta^*(x) = \frac{x + \sqrt{n}/2}{n + \sqrt{n}}.$$

De plus, cet estimateur a *un risque constant*, $R(\theta, \delta^*) = 1/4(1 + \sqrt{n})^2$. Par conséquent, en intégrant sur θ, $r(\pi) = R(\theta, \delta^*)$ et δ^* est minimax selon le Lemme 2.25. Notons la différence avec l'estimateur du maximum de vraisemblance, $\delta_0(x) = x/n$, pour des petites valeurs de n et la concentration irréaliste de l'a priori dans un voisinage de 0.5 pour les valeurs les plus grandes de n. ‖

Puisque les estimateurs minimax correspondent généralement à des *estimateurs de Bayes généralisés*, on doit souvent recourir à un argument limite pour établir la minimaxité, plutôt que de calculer directement le risque de Bayes comme dans le Lemme 2.25.

Lemme 2.27. *S'il existe une suite (π_n) de lois a priori propres telles que l'estimateur de Bayes généralisé δ_0 satisfasse*

$$R(\theta, \delta_0) \leq \lim_{n \to \infty} r(\pi_n) < +\infty$$

pour tout $\theta \in \Theta$, alors δ_0 est minimax.

Exemple 2.28. Quand $x \sim \mathcal{N}(\theta, 1)$, l'estimateur de maximum de vraisemblance $\delta_0(x) = x$ est un estimateur de Bayes généralisé par rapport à la mesure de Lebesgue sur $\mathbb{R}$, pour le coût quadratique. Puisque

$$R(\delta_0, \theta) = \mathbb{E}_\theta (x - \theta)^2 = 1 ,$$

ce risque est la limite du risque de Bayes $r(\pi_n)$ quand π_n est égal à $\mathcal{N}(0, n)$, comme

$$r(\pi_n) = \frac{n}{n+1}\,.$$

Par conséquent, l'estimateur de maximum de vraisemblance δ_0 est minimax. Notons que cet argument peut être étendu directement au cas $x \sim \mathcal{N}_p(\theta, I_p)$ pour établir que δ_0 est minimax pour tout p. ‖

Quand l'espace Θ est compact, une description exacte des règles (ou des estimateurs) de Bayes minimax est disponible. Ceci découle du *principe des zéros séparés* pour les nombres complexes : si la fonction $R(\theta, \delta^\pi)$ n'est pas constante et est analytique, l'ensemble des θ tels que $R(\theta, \delta^\pi)$ est maximal est un ensemble *séparé* et, dans le cas d'un ensemble compact Θ, forcément fini.

Théorème 2.29. *Considérons un problème statistique admettant simultanément une valeur, une loi la moins favorable π_0, et un estimateur minimax δ^{π_0}. Alors, si $\Theta \subset \mathbb{R}$ est compact et si $R(\theta, \delta^{\pi_0})$ est une fonction analytique de θ, soit π_0 a un support fini, soit $R(\theta, \delta^{\pi_0})$ est constant.*

Exemple 2.30. Soit $x \sim \mathcal{N}(\theta, 1)$, avec $|\theta| \leq m$, c'est-à-dire $\theta \in [-m, m]$. Alors, selon le Théorème 2.29, les lois les moins favorables ont nécessairement un support fini, $\{\pm\theta_i,\ 1 \leq i \leq \omega\}$, avec un cardinal 2ω ou $2\omega - 1$ et des points de support θ_i dépendant de m. En effet, le seul estimateur à risque constant est $\delta_0(x) = x$, qui n'est pas minimax dans ce cas. En général, la détermination exacte de n et des points de θ_i ne peut être faite que numériquement. Par exemple, quand $m \leq 1.06$, la loi a priori avec pour poids $1/2$ en $\pm m$ est la *seule* distribution a priori la moins favorable. Pour $1.06 \leq m \leq 2$, le support de π contient $-m$, 0, et m. Voir Casella et Strawderman (1981) et Bickel (1981) pour plus de détails, et Johnstone et MacGibbon (1992) pour un traitement similaire du modèle de Poisson. ‖

Les exemples ci-dessus montrent pourquoi le principe minimax, bien qu'étroitement lié au paradigme bayésien, n'est pas nécessairement attirant d'un point de vue bayésien. En effet, mis à part le fait que les estimateurs minimax sont parfois randomisés, comme dans l'Exemple 2.24, les Exemples 2.26 et 2.30 montrent que les lois a priori les moins favorables sont souvent irréalistes, car conduisant à un fort biais a priori vers quelques points de l'espace d'échantillonnage. Pour l'Exemple 2.30, Gatsonis *et al.* (1987) ont montré que les lois a priori uniformes sont de bons substituts à des lois a priori à support discret, même si elles ne sont pas minimax.

Des extensions du Théorème 2.29 au cas non compact sont données dans Kempthorne (1988). Dans un cadre multidimensionnel, quand le problème est invariant par rotation, les lois les moins favorables sont uniformes sur une suite de sphères imbriquées (voir Robert *et al.*, 1990). Le problème pratique de la

détermination des points du support est considéré par Kempthorne (1988) et Eichenauer et Lehn (1989).

Lorsqu'un problème admet une valeur, il est souvent difficile de construire la loi la moins favorable. Des méthodes alternatives pour obtenir un estimateur minimax sont alors nécessaires. Le Chapitre 9 montre comment la détermination de certaines structures d'invariance du modèle peut conduire à l'identification du meilleur estimateur équivariant et à un estimateur minimax (théorème de Hunt-Stein). Malheureusement, les conditions sous lesquelles ce théorème peut s'appliquer sont difficiles à vérifier et sont rarement satisfaites.

Finalement, une fois qu'on a obtenu un estimateur minimax, il reste à déterminer s'il est optimal ou non : plusieurs estimateurs minimax peuvent exister simultanément et certains d'entre eux peuvent dominer uniformément d'autres. Il est alors nécessaire de présenter un deuxième critère, plus local, pour comparer les estimateurs minimax, qui sont des estimateurs ayant de bonnes performances globales.

2.4.4 Admissibilité

Ce deuxième critère fréquentiste induit un ordre partiel sur $\mathscr{D}^*$ en comparant les risques fréquentistes des estimateurs $R(\theta, \delta)$.

Définition 2.31. *Un estimateur δ_0 est* inadmissible *s'il existe un estimateur δ_1 qui domine δ_0, c'est-à-dire tel que pour tout θ,*

$$R(\theta, \delta_0) \geq R(\theta, \delta_1)$$

et, pour au moins une valeur θ_0 du paramètre,

$$R(\theta_0, \delta_0) > R(\theta_0, \delta_1)\,.$$

Sinon, δ_0 est dit admissible.

Ce critère est particulièrement intéressant pour son action *réductrice*. Effectivement, du moins en théorie, il semble logique de soutenir que les estimateurs inadmissibles ne devraient pas être considérés du tout, puisqu'ils peuvent être améliorés uniformément. Par exemple, le théorème de Rao-Blackwell implique alors que, pour des fonctions de coût convexes, les estimateurs randomisés et plus généralement ceux dépendant d'autres quantités que les statistiques exhaustives sont inadmissibles. Cependant, l'admissibilité à elle seule n'est pas suffisante pour valider l'utilisation d'un estimateur. Par exemple, les estimateurs constants $\delta(x) = \theta_0$ sont en général admissibles parce qu'ils fournissent une valeur exacte pour $\theta = \theta_0$. D'un point de vue fréquentiste, il est donc important de chercher des estimateurs qui satisfassent les deux optimalités : minimaxité et admissibilité. Dans cette optique, on peut mentionner les résultats suivants.

Proposition 2.32. *S'il existe un unique estimateur minimax, cet estimateur est admissible.*

Preuve. Si δ^* est le seul estimateur minimax, pour tout estimateur $\tilde{\delta} \neq \delta^*$,

$$\sup_\theta R(\theta, \tilde{\delta}) > \sup_\theta R(\theta, \delta^*).$$

Donc, $\tilde{\delta}$ ne peut pas dominer δ^*. □

Notons que la réciproque de ce résultat est fausse, car il peut exister plusieurs estimateurs minimax admissibles. Par exemple, dans le cas $\mathcal{N}_p(\theta, I_p)$, il existe des estimateurs de Bayes réguliers minimax pour $p \geq 5$ (Strawderman, 1971 et Fourdrinier *et al.*, 1998). Quand la fonction de coût L est absolument convexe (en d), la caractérisation suivante est aussi possible.

Proposition 2.33. *Si δ_0 est admissible de risque constant, δ_0 est l'unique estimateur minimax.*

Preuve. Pour tout $\theta_0 \in \Theta$, $\sup_\theta R(\theta, \delta_0) = R(\theta_0, \delta_0)$. Alors, s'il existe δ_1 tel que $\bar{R} \leq \sup_\theta R(\theta, \delta_1) < R(\theta_0, \delta_0)$, δ_0 ne peut pas être admissible. De la même façon, si

$$\bar{R} = \sup_\theta R(\theta, \delta_1) = R(\theta_0, \delta_0)$$

et si θ_1 est tel que $R(\theta_1, \delta_1) < \bar{R}$, δ_1 domine δ_0. Par conséquent, quand δ_0 est admissible, le seul cas possible est qu'il existe δ_1 tel que $R(\theta, \delta_1) = R(\theta, \delta_0)$ pour tout $\theta \in \Theta$. Ce qui est aussi impossible quand δ_0 est admissible (voir l'Exercice 2.36). □

Remarquons à nouveau que la réciproque de ce résultat est fausse. Il peut exister des estimateurs minimax ayant un risque constant qui soient inadmissibles. En fait, ils sont inadmissibles dès qu'il existe d'autres estimateurs minimax. C'est le cas par exemple pour $\delta_0(x) = x$ quand $x \sim \mathcal{N}_p(\theta, I_p)$ et $p \geq 3$ (voir la Note 2.8.2 et l'Exercice 2.57). Il y a aussi des cas où il n'existe pas d'estimateur minimax admissible (il faut pour cela qu'il n'existe pas de *classe minimale complète*, voir le Chapitre 8).

Nous avons vu dans la section précédente que la minimaxité pouvait être parfois considérée, dans une perspective bayésienne, comme un choix par la "Nature" d'une stratégie maximin (loi la moins favorable), π, donc que *certains* estimateurs minimax sont des estimateurs de Bayes. La notion d'admissibilité est encore plus fortement liée au paradigme bayésien au sens où, dans la plupart des problèmes statistiques, les estimateurs de Bayes "engendrent" la classe des estimateurs admissibles, c'est-à-dire que ces derniers peuvent être écrits comme des estimateurs de Bayes (ou estimateurs de Bayes généralisés) ou comme limites d'estimateurs de Bayes. Le Chapitre 8 est consacré plus en détail aux relations entre estimateurs de Bayes et admissibilité. Nous ne donnerons ici que deux résultats importants.

Proposition 2.34. *Si la distribution a priori π est strictement positive sur Θ, de risque de Bayes fini, et la fonction de risque $R(\theta, \delta)$ est une fonction continue de θ pour tout δ, l'estimateur de Bayes δ^π est admissible.*

Preuve. Supposons δ^π inadmissible. Soit δ' un estimateur dominant uniformément δ^π. Alors, pour tout θ, $R(\theta, \delta') \leq R(\theta, \delta^\pi)$ et, dans tout *ensemble ouvert* C de Θ, $R(\theta, \delta') < R(\theta, \delta^\pi)$. Par intégration de cette inégalité, on obtient

$$r(\pi, \delta') < r(\pi, \delta^\pi) = \int_\Theta R(\theta, \delta^\pi)\pi(\theta)\, d\theta,$$

ce qui est impossible. □

Proposition 2.35. *Si l'estimateur de Bayes associé à une loi a priori π est unique, il est admissible.*

La démonstration de ce résultat est similaire à celle de la Proposition 2.32. Même si l'estimateur de Bayes n'est pas unique, il reste possible de présenter au moins *un* estimateur de Bayes admissible. Quand la fonction de coût est strictement convexe, l'estimateur de Bayes est nécessairement unique et donc admissible, selon la proposition ci-dessous.

Exemple 2.36. (Suite de l'Exemple 2.26) L'estimateur δ^* est un estimateur de Bayes régulier, donc admissible, et de risque constant. Par conséquent, il est l'*estimateur minimax unique* sous le coût quadratique. ‖

Notons que la Proposition 2.34 fait intervenir l'hypothèse d'un risque de Bayes fini. Autrement, tout estimateur est, dans un certain sens, un estimateur de Bayes (voir l'Exercice 2.43). D'un autre côté, quelques résultats d'admissibilité peuvent être établis pour des lois a priori impropres. C'est la raison pour laquelle nous préférons appeler *estimateurs de Bayes généralisés* ceux associés à un risque de Bayes infini, plutôt que les estimateurs correspondant à une loi a priori impropre. Ce choix implique que les estimateurs de Bayes de différentes quantités associés à la même loi a priori peuvent être respectivement estimateurs de Bayes réguliers et estimateurs de Bayes généralisés, suivant ce qu'ils estiment. Ceci assure aussi que les estimateurs de Bayes réguliers seront toujours admissibles, comme le démontre le résultat suivant.

Proposition 2.37. *Si un estimateur de Bayes, δ^π, associé à une loi a priori (propre ou impropre) π, est tel que le risque de Bayes,*

$$r(\pi) = \int_\Theta R(\theta, \delta^\pi)\pi(\theta)\, d\theta,$$

soit fini, δ^π est admissible.

Exemple 2.38. Soit $x \sim \mathcal{N}(\theta, 1)$, et on veut tester l'hypothèse nulle $H_0 : \theta \leq 0$ contre l'hypothèse alternative $H_1 : \theta > 0$. Ce problème de test est un problème d'*estimation* si nous considérons l'estimation de la fonction indicatrice $\mathbb{I}_{H_0}(\theta)$. Sous le coût quadratique

$$\left(\mathbb{I}_{H_0}(\theta) - \delta(x)\right)^2,$$

nous pouvons proposer l'estimateur suivant

$$\begin{aligned} p(x) &= P_0(X > x) \qquad (X \sim \mathcal{N}(0,1)) \\ &= 1 - \Phi(x), \end{aligned}$$

dit *p-value*, qui est considéré comme une bonne réponse fréquentiste au problème de test (voir Kiefer, 1977 et Casella et Berger, 1987). En utilisant l'Exemple 1.25, il est facile de montrer que p est un estimateur de Bayes sous la mesure de Lebesgue et un coût quadratique, car $\pi(\theta|x)$ est la distribution $\mathcal{N}(x, 1)$ et

$$\begin{aligned} p(x) &= \mathbb{E}^\pi[\mathbb{I}_{H_0}(\theta)|x] = P^\pi(\theta < 0|x) \\ &= P^\pi(\theta - x < -x|x) = 1 - \Phi(x). \end{aligned}$$

De plus, le risque de Bayes de p est fini (Exercice 2.34). Par conséquent la p-value, en tant qu'estimateur de $\mathbb{I}_{H_0}$, est admissible. (Voir la Section 5.4 pour une analyse approfondie des propriétés de la p-value.) ‖

Exemple 2.39. Dans le cadre de l'exemple précédent, si θ est le paramètre d'intérêt, $\delta_0(x) = x$ est un estimateur de Bayes généralisé sous le coût quadratique, car

$$r(\pi, \delta_0) = \int_{-\infty}^{+\infty} R(\theta, \delta_0)\, d\theta = \int_{-\infty}^{+\infty} 1\, d\theta = +\infty\,.$$

La Proposition 2.35 ne permet donc pas dans ce cas de déterminer l'admissibilité de δ_0. Bien que δ_0 soit en réalité admissible, son admissibilité doit être établie à l'aide d'une suite de lois a priori propres, comme nous le montrerons dans le Chapitre 8. ‖

Exemple 2.40. Soit $x \sim \mathcal{N}_p(\theta, I_p)$. Si le paramètre d'intérêt est $||\theta||^2$ et la loi a priori est la mesure de Lebesgue sur $\mathbb{R}^p$, puisque $\mathbb{E}^\pi[||\theta||^2|x] = \mathbb{E}[||y||^2]$, avec $y \sim \mathcal{N}_p(x, I_p)$, l'estimateur de Bayes sous le coût quadratique est

$$\delta^\pi(x) = ||x||^2 + p.$$

Cet estimateur de Bayes généralisé n'est pas admissible parce qu'il est dominé par $\delta_0(x) = ||x||^2 - p$ (Exercice 2.35). (Puisque le risque classique est $R(\theta, \delta^\pi) = \mathrm{var}(\|x\|^2) + 4p^2$, le risque de Bayes est bien infini.) Ce phénomène montre que la mesure de Lebesgue n'est pas nécessairement le meilleur choix d'une mesure a priori non informative quand le paramètre d'intérêt est un sous-vecteur du paramètre (voir le Chapitre 3). ‖

2.5 Fonctions de coût usuelles

Quand le contexte d'une expérience ne permet pas une détermination de la fonction d'utilité (manque de temps, information, etc.), une alternative courante est de faire appel à des fonctions de coût classiques, qui sont mathématiquement simples et de propriétés connues. Bien entendu, cette approche est une approximation sous-jacente du modèle statistique et ne devrait être utilisée que quand la fonction d'utilité n'est pas disponible. Nous finissons cette section par une note sur des fonctions de coût plus intrinsèques, même si celles-ci sont rarement utilisées en pratique. (Voir aussi la Note 2.8.1 pour une description des fonctions de coût utilisées en analyse d'image.)

2.5.1 Le coût quadratique

Introduit par Legendre (1805) et Gauss (1810), ce coût est sans conteste le critère d'évaluation le plus commun. Fondant sa validité sur l'ambiguïté de la notion d'*erreur* dans un contexte statistique (soit erreur de mesure, soit variation aléatoire), il a aussi donné lieu à de nombreuses critiques, la plus fréquente étant sans doute le fait que le coût quadratique

$$\mathrm{L}(\theta, d) = (\theta - d)^2 \tag{2.2}$$

pénalise trop fortement les grandes erreurs.

Cependant, les fonctions de coût convexes comme (2.2) ont l'avantage incomparable d'éviter le paradoxe des *amateurs de risque* (traduction de *risk lovers*) et d'exclure les estimateurs randomisés. Une autre justification habituelle pour le coût quadratique est que celui-ci peut être vu comme le développement limité d'un coût symétrique plus complexe (voir l'Exercice 4.15 pour un contre-exemple). Dans son article de 1810, Gauss reconnaissait déjà l'arbitraire du coût quadratique mais le défendait au nom de la simplicité. Bien que les critiques concernant l'utilisation systématique de la fonction de coût quadratique soient fondées, son usage est néanmoins très répandu, car il donne en général des solutions bayésiennes qui sont celles naturellement fournies comme estimateurs pour une inférence non décisionnelle fondée sur une distribution a priori. En effet, les estimateurs de Bayes associés au coût quadratique sont les moyennes a posteriori. Cependant, notons que le coût quadratique n'est pas le seul coût à avoir cette caractéristique. Les fonctions de coût conduisant à la moyenne a posteriori comme estimateur de Bayes sont appelées *fonctions de coût propres* et ont été identifiées par Lindley (1985), Schervish (1989), der Meulen B. (1992), et Hwang et Pemantle (1994) (voir aussi l'Exercice 2.15).

Proposition 2.41. *L'estimateur de Bayes δ^π associé à la loi a priori π et au coût quadratique* (2.2) *est la moyenne a posteriori*

$$\delta^\pi(x) = \mathbb{E}^\pi[\theta|x] = \frac{\int_\Theta \theta f(x|\theta)\pi(\theta)\,d\theta}{\int_\Theta f(x|\theta)\pi(\theta)\,d\theta}.$$

Preuve. Comme

$$\mathbb{E}^\pi[(\theta-\delta)^2|x] = \mathbb{E}^\pi[\theta^2|x] - 2\delta\mathbb{E}^\pi[\theta|x] + \delta^2,$$

le minimum du coût a posteriori est effectivement atteint par $\delta^\pi(x) = \mathbb{E}^\pi[\theta \mid x]$. □

Les corollaires suivants se déduisent de manière immédiate.

Corollaire 2.42. *L'estimateur de Bayes δ^π associé à π et au coût quadratique pondéré*

$$\mathrm{L}(\theta,\delta) = \omega(\theta)(\theta-\delta)^2, \tag{2.3}$$

où $\omega(\theta)$ est une fonction positive, est

$$\delta^\pi(x) = \frac{\mathbb{E}^\pi[\omega(\theta)\theta|x]}{\mathbb{E}^\pi[\omega(\theta)|x]}.$$

Corollaire 2.43. *Quand $\Theta \in \mathbb{R}^p$, l'estimateur de Bayes δ^π associé à π et au coût quadratique,*

$$\mathrm{L}(\theta,\delta) = (\theta-\delta)^t Q(\theta-\delta),$$

est la moyenne a posteriori, $\delta^\pi(x) = \mathbb{E}^\pi[\theta|x]$, pour toute matrice Q $p \times p$ symétrique définie positive.

Le Corollaire 2.42 exhibe une *dualité* (faible) entre coût et loi a priori, au sens où il revient au même d'estimer θ sous (2.3) avec la loi π, ou sous (2.2) avec la loi $\pi_\omega(\theta) \propto \pi(\theta)\omega(\theta)$. De plus, bien que la notion d'admissibilité soit indépendante de la fonction ω, l'estimateur de Bayes en dépend fortement. Par exemple, δ^π peut ne pas exister si ω croît trop vite vers $+\infty$. D'un autre côté, le Corollaire 2.43 montre la robustesse de l'estimateur de Bayes par rapport à la forme quadratique de Q. (Shinozaki, 1975, a aussi montré que le caractère admissible ne dépend pas de Q.)

Le coût quadratique est particulièrement intéressant lorsque l'espace des paramètres est borné et le choix d'un coût plus subjectif est impossible. En effet, ce coût est assez simple d'utilisation et l'erreur d'approximation est alors de faible importance. L'indétermination de la fonction de coût (et son remplacement par une approximation quadratique) est fréquente en *évaluation de la précision*, qui inclut par exemple l'*estimation du coût* (Rukhin, 1988a,b, Lu et Berger, 1989a,b, Hwang *et al.*, 1992, Robert et Casella, 1993, 1994, et Fourdrinier et Wells, 1993).

Exemple 2.44. (Suite de l'Exemple 2.21) Nous cherchons à évaluer la performance de l'estimateur

$$\delta(x_1, x_2) = \begin{cases} \dfrac{x_1 + x_2}{2} & \text{si } x_1 \neq x_2, \\ x_1 + 1 & \text{sinon,} \end{cases}$$

par $\alpha(x_1, x_2)$ sous le critère quadratique

$$[\mathbb{I}_\theta(\delta(x_1, x_2)) - \alpha(x_1, x_2)]^2 ,$$

où $\mathbb{I}_\theta(v)$ vaut 1 si $v = \theta$, 0 sinon ; la fonction α estime donc d'une certaine façon la probabilité que δ prenne la vraie valeur θ. (Ceci est un cas particulier d'estimation de coût, pour la fonction de coût $1 - \mathbb{I}_\theta(\delta)$.) Deux estimateurs peuvent être considérés :

(i) $\alpha_0(x_1, x_2) = 0.75$, qui donne l'espérance de $\mathbb{I}_\theta(\delta(x_1, x_2))$; et

(ii) $\alpha_1(x_1, x_2) = \begin{cases} 1 & \text{si } x_1 \neq x_2, \\ 0.50 & \text{si } x_1 = x_2. \end{cases}$

Le risque des deux estimateurs est alors

$$\begin{aligned} R(\theta, \alpha_0) &= \mathbb{E}_\theta \left(\mathbb{I}_\theta(\delta(x_1, x_2)) - 0.75\right)^2 \\ &= 0.75 - (0.75)^2 = 0.1875 \end{aligned}$$

et

$$\begin{aligned} R(\theta, \alpha_1) &= \mathbb{E}_\theta \left(\mathbb{I}_\theta(\delta(x_1, x_2)) - \alpha_1(x_1, x_2)\right)^2 \\ &= (0.5)^2 \frac{1}{2} = 0.125 \ . \end{aligned}$$

Par conséquent, α_1 est un meilleur estimateur des performances de δ que α_0. Présenté dans Berger et Wolpert (1988), ce résultat de domination est assez logique et suggère qu'une évaluation conditionnelle des estimateurs est plus appropriée. ‖

2.5.2 L'erreur de coût absolu

Une solution alternative au coût quadratique en dimension un est d'utiliser le coût absolu,

$$\mathrm{L}(\theta, d) = \mid \theta - d \mid, \tag{2.4}$$

déjà considéré par Laplace (1773) ou, plus généralement, une fonction linéaire par morceaux

$$\mathrm{L}_{k_1,k_2}(\theta, d) = \begin{cases} k_2(\theta - d) & \text{si } \theta > d, \\ k_1(d - \theta) & \text{sinon.} \end{cases} \tag{2.5}$$

De telles fonctions croissent plus lentement que le coût quadratique. Par conséquent, tout en restant convexes, elles ne surpénalisent pas des erreurs

grandes mais peu vraisemblables. Huber (1964a) propose aussi un mélange des fonctions coûts absolues et quadratiques, pour maintenir une pénalisation quadratique aux alentours de 0,

$$\tilde{\mathrm{L}}(\theta,d)=\begin{cases}(d-\theta)^2 & \text{si } \mid d-\theta \mid < k,\\ 2k \mid d-\theta \mid -k^2 & \text{sinon.}\end{cases}$$

Bien que convexe[22], le coût mixte ralentit la progression du coût quadratique pour des grandes erreurs et robustifie son effet. Malheureusement, il n'existe pas en général de formule explicite des estimateurs de Bayes sous cette fonction de coût.

Proposition 2.45. *L'estimateur de Bayes associé à la loi a priori π et à la fonction de coût linéaire par morceaux* (2.5) *est le fractile* $(k_2/(k_1+k_2))$ *de* $\pi(\theta|x)$.

Preuve. L'équation classique suivante,

$$\begin{aligned}\mathbb{E}^\pi[\mathrm{L}_{k_1,k_2}(\theta,d)|x] &= k_1\int_{-\infty}^{d}(d-\theta)\pi(\theta|x)\,d\theta + k_2\int_{d}^{+\infty}(\theta-d)\pi(\theta|x)\,d\theta\\ &= k_1\int_{-\infty}^{d}P^\pi(\theta<y|x)\,dy + k_2\int_{d}^{+\infty}P^\pi(\theta>y|x)\,dy,\end{aligned}$$

est obtenue par une intégration par parties. Dérivant en d, on obtient

$$k_1P^\pi(\theta<d|x)-k_2P^\pi(\theta>d|x)=0,$$

soit encore

$$P^\pi(\theta<d|x)=\frac{k_2}{k_1+k_2}.$$

□

En particulier, si $k_1=k_2$, soit, dans le cas du coût absolu, l'estimateur de Bayes est la médiane a posteriori, qui est l'estimateur obtenu par Laplace (voir l'Exemple 1.11). Notons que, quand π a un support non connexe, la Proposition 2.45 fournit des exemples d'estimateurs de Bayes multiples pour certaines valeurs de x (voir l'Exercice 2.40).

[22]De nouveau, si nous insistons sur la *convexité*, c'est parce qu'elle assure que les estimateurs randomisés sont sous-optimaux d'un point de vue fréquentiste. Par conséquent, une approche décisionnelle statistique qui voudrait rester le plus fidèle possible au principe de vraisemblance impose nécessairement d'avoir une fonction de coût convexe. Bien évidemment, cette exigence exclut les fonctions de coût bornées.

2.5.3 Le coût $0-1$

Ce coût est surtout utilisé dans l'approche classique des tests d'hypothèse, proposée par Neyman et Pearson (voir la Section 5.3). Plus généralement, c'est un exemple typique d'un coût non quantitatif. En effet, pour ce coût, la pénalité associée à un estimateur δ est 0 si la réponse est correcte et 1 sinon.

Exemple 2.46. Soit le test de $H_0 : \theta \in \Theta_0$ contre $H_1 : \theta \notin \Theta_0$. Alors $\mathscr{D} = \{0, 1\}$, où 1 représente l'acceptation de H_0 et 0 son rejet. (En d'autres termes, la fonction de θ estimée est $\mathbb{I}_{\Theta_0}(\theta)$.) Pour la fonction de coût $0-1$, qui vaut

$$\mathrm{L}(\theta, d) = \begin{cases} 1-d & \text{si } \theta \in \Theta_0 \\ d & \text{sinon,} \end{cases} \tag{2.6}$$

le risque associé est

$$\begin{aligned} R(\theta, \delta) &= \mathbb{E}_\theta[\mathrm{L}(\theta, \delta(x))] \\ &= \begin{cases} P_\theta(\delta(x) = 0) & \text{si } \theta \in \Theta_0, \\ P_\theta(\delta(x) = 1) & \text{sinon,} \end{cases} \end{aligned}$$

ce qui donne exactement les *erreurs de première et deuxième espèce* qui sous-tendent la Théorie de Neyman-Pearson . ‖

Ce coût n'est pas très intéressant, de par son caractère non quantitatif, et nous verrons au Chapitre 5 quelques théories alternatives pour le test d'hypothèses. Les estimateurs de Bayes associés reflètent aussi l'aspect primitif d'un tel coût (voir aussi l'Exercice 2.41).

Proposition 2.47. *L'estimateur de Bayes associé à π et au coût* (2.6) *est*

$$\delta^\pi(x) = \begin{cases} 1 & \textit{si } P(\theta \in \Theta_0|x) > P(\theta \notin \Theta_0|x), \\ 0 & \textit{sinon,} \end{cases}$$

donc $\delta^\pi(x)$ vaut 1 si et seulement si $P(\theta \in \Theta_0|x) > 1/2$.

2.5.4 Coûts intrinsèques

Il peut arriver que certains problèmes soient tellement non informatifs que non seulement la fonction de coût soit inconnue, mais aussi que le modèle n'admette pas une paramétrisation naturelle. Ce type de situation apparaît quand c'est la loi $f(x|\theta)$ elle-même qui nous intéresse, par exemple dans un contexte de prévision.

Cependant, comme nous l'avons évoqué dans la section précédente, le choix de la paramétrisation est important, car, contrairement à l'approche du maximum de vraisemblance, si g est une transformation bijective de θ, l'estimateur de Bayes de $g(\theta)$ est généralement différent de la transformation par g de l'estimateur de Bayes de θ sous le même coût (voir l'Exercice 2.36). Ce manque d'invariance, bien qu'il soit perturbant pour les néophytes, n'est généralement pas préoccupant pour les décideurs, car il montre comment le paradigme bayésien peut s'adapter à un problème d'estimation donné *et* à une fonction de coût donnée, tandis que l'estimation par maximum de vraisemblance n'est pas capable de tenir compte de la notion de coût. Mais les quelques cas où la fonction de coût et la paramétrisation naturelle sont absolument indisponibles peuvent nécessiter ce type d'invariance ultime. (Voir Wallace et Boulton, 1975, pour une autre approche.)

Dans un tel contexte non informatif, il semble naturel d'utiliser des coûts comparant directement les distributions $f(\cdot|\theta)$ et $f(\cdot|\delta)$ associées au vrai paramètre θ et l'estimateur δ. De telles fonctions de coût,

$$\mathrm{L}(\theta,\delta) = d(f(\cdot|\theta), f(\cdot|\delta)),$$

sont effectivement indépendantes de la paramétrisation. Deux distances standard pour les distributions sont

(1) la *distance entropique*

$$\mathrm{L_e}(\theta,\delta) = \mathbb{E}_\theta\left[\log\left(\frac{f(x|\theta)}{f(x|\delta)}\right)\right], \tag{2.7}$$

dite aussi divergence de Kullback-Leibler et qui n'est pas une distance au sens mathématique à cause de son asymétrie ; et

(2) la *distance de Hellinger*

$$\mathrm{L_H}(\theta,\delta) = \frac{1}{2}\,\mathbb{E}_\theta\left[\left(\sqrt{\frac{f(x|\delta)}{f(x|\theta)}} - 1\right)^2\right]. \tag{2.8}$$

Exemple 2.48. Soit $x \sim \mathcal{N}(\theta, 1)$. On a alors

$$\begin{aligned}\mathrm{L_e}(\theta,\delta) &= \frac{1}{2}\mathbb{E}_\theta[-(x-\theta)^2 + (x-\delta)^2] = \frac{1}{2}(\delta-\theta)^2,\\ \mathrm{L_H}(\theta,\delta) &= 1 - \exp\{-(\delta-\theta)^2/8\}.\end{aligned}$$

Dans le cas normal où $\pi(\theta|x)$ est une loi $\mathcal{N}(\mu(x), \sigma^2)$, il est trivial de démontrer que l'estimateur de Bayes est $\mu(x)$ dans les deux cas. ‖

Le coût de Hellinger est sans doute plus intrinsèque que le coût entropique, ne serait-ce que parce qu'il existe toujours (notons que (2.8) est majoré par 1). Malheureusement, bien qu'il mène à des expressions explicites de $\mathrm{L_H}(\theta,\delta)$

pour les familles de distributions usuelles, il ne permet pas de calcul explicite des estimateurs de Bayes, sauf dans le cas particulier traité ci-dessus. En revanche, pour les *familles exponentielles*, le coût entropique fournit un estimateur explicite qui est la moyenne a posteriori du *paramètre naturel* (voir le Chapitre 3). De plus, bien qu'il soit assez différent du coût de Hellinger, le coût entropique fournit des réponses similaires pour les familles de distributions habituelles (voir Robert, 1996b). Il y a aussi plusieurs raisons théoriques pour défendre l'utilisation de la distance de Kullback-Leibler, allant de la théorie de l'information (Exercice 2.48) à l'importance de la règle du score logarithmique et de l'invariance de position et d'échelle, comme le détaillent Bernardo et Smith (1994).

2.6 Critiques et alternatives

Quelques critiques des notions fréquentistes de minimaxité et d'admissibilité ont été mentionnées dans les sections précédentes. Ces concepts ont en réalité peu d'importance d'un point de vue purement bayésien. D'une part, l'admissibilité est automatiquement satisfaite par la plupart des estimateurs de Bayes. D'autre part, la minimaxité est en quelque sorte incompatible avec le paradigme bayésien, car, sous une loi a priori, les valeurs du paramètre ne peuvent pas être pondérées de façon égale. Cependant, la minimaxité peut être pertinente en termes de robustesse, c'est-à-dire quand l'information a priori n'est pas suffisamment précise pour déterminer la loi a priori.

Il arrive parfois que le décideur soit incapable de construire précisément la fonction de coût. Par exemple, quand le décideur est un comité composé de plusieurs experts, il n'est pas rare que ceux-ci soient en désaccord sur le choix de la fonction de coût (et parfois même de la distribution a priori). Partant d'Arrow (1956), la littérature sur ces extensions de la Théorie de la Décision est assez vaste (voir Genest et Zidek, 1986, Rubin, 1984, et Van Eeden et Zidek, 1993, pour des détails et références).

Lorsque la fonction de coût n'a pu être entièrement déterminée, on peut supposer qu'elle appartient à une famille paramétrique de fonctions de coût, le décideur choisissant le paramètre le plus approprié. Mis à part les coûts L_p, deux autres possibilités sont

$$\mathrm{L}_1(\theta,\delta) = \log(\alpha||\theta-\delta||^2+1), \qquad \mathrm{L}_2(\theta,\delta) = 1-\exp\{-c||\theta-\delta||^2\}.$$

Une approche alternative plus en accord avec le paradigme bayésien est de considérer que, du moment que le coût est partiellement inconnu, cette incertitude peut être représentée par une *fonction de coût aléatoire* $\mathrm{L}(\theta,\delta)$. L'évaluation des estimateurs est alors obtenue en intégrant par rapport à cette variable additionnelle : si F est la distribution du coût, la fonction à minimiser (en δ) est

$$\int_\Theta \int_\Omega \mathrm{L}(\theta, \delta, \omega) dF(\omega)\, d\pi(\theta|x), \tag{2.9}$$

où F dépend éventuellement de θ ou même de x. En réalité, ce cas est la seule extension intéressante, car, sinon, minimiser (2.9) revient à utiliser le coût moyen

$$\bar{\mathrm{L}}(\theta, \delta) = \int_\Omega \mathrm{L}(\theta, \delta, \omega)\, dF(\omega).$$

Une autre approche du problème de manque de précision de la fonction de coût consiste à considérer simultanément un ensemble de fonctions de coût et à construire des estimateurs ayant de bonnes performances pour toutes ces fonctions. Évidemment, ce critère multidimensionnel n'engendre qu'un ordre *partiel* sur les estimateurs. On pourra consulter Abraham et Daurés (2000) et Abraham (2001) pour des perspectives intéressantes sur cette approche robuste des coûts.

Exemple 2.49. Soit $x \sim \mathcal{N}_p(\theta, I_p)$. Le paramètre θ est estimé sous un coût quadratique. Si la matrice des coûts Q n'est pas déterminée exactement, une alternative robuste est de considérer les coûts associés aux matrices Q telles que $Q_1 \preceq Q \preceq Q_1$ (où $A \preceq B$ signifie que la matrice $B - A$ est définie positive). Notons que, selon le Corollaire 2.43, l'estimateur de Bayes est le même pour tous les Q. ‖

Exemple 2.50. Dans le cadre de l'exemple ci-dessus, Brown (1975) montre qu'un estimateur à rétrécisseur de la forme $(1 - h(x))x$ domine $\delta_0(x) = x$ pour une classe de coûts quadratiques, c'est-à-dire une classe de matrices Q, si et seulement si

$$\mathrm{tr}(Q) - 2\lambda_{\max}(Q) > 0 \tag{2.10}$$

pour toute matrice dans la classe (où $\lambda_{\max}(Q)$ désigne la plus grande valeur propre de la matrice Q). Notons que cette condition exclut le cas $p \leq 2$, pour lequel δ_0 est en réalité admissible. La constante $\mathrm{tr}(Q) - 2\lambda_{\max}(Q)$ apparaît aussi dans la constante de majoration de $||x||^2 h(||x||^2)$ (voir le Théorème 2.52). Par conséquent, (2.10) est à la fois une condition nécessaire et suffisante pour avoir un phénomène de Stein (voir l'Exemple 2.18 et la Note 2.8.2). ‖

Le critère ultime pour la robustesse de la fonction de coût est celui de la *domination universelle* introduit par Hwang (1985). En effet, ce critère considère l'ensemble de toutes les fonctions de coût $\ell(||\delta - \theta||_Q)$, pour une norme donnée $||x||_Q = x^t Q x$ et toutes les fonctions croissantes ℓ. Un estimateur δ_1 est dit *dominer universellement* un autre estimateur δ_2 si, pour tout ℓ,

$$\mathbb{E}_\theta[\ell(||\delta_1(x) - \theta||_Q)] \leq \mathbb{E}_\theta[\ell(||\delta_2(x) - \theta||_Q)].$$

Un deuxième critère est celui de la *domination stochastique* : δ_1 *domine stochastiquement* δ_2 si, pour tout $c > 0$,

$$P_\theta(||\delta_1(x) - \theta||_Q \leq c) \geq P_\theta(||\delta_2(x) - \theta||_Q \leq c).$$

Bien que ce critère paraisse plus intrinsèque et moins lié à la Théorie de la Décision que la domination universelle, Hwang (1985) a montré que les deux critères sont en réalité équivalents.

Théorème 2.51. *Un estimateur δ_1 domine universellement un estimateur δ_2 si et seulement si δ_1 domine stochastiquement δ_2.*

Preuve. L'estimateur δ_1 domine stochastiquement δ_2 si, pour tout $c > 0$,

$$P_\theta(||\delta_1(x) - \theta||_Q \leq c) \geq P_\theta(||\delta_2(x) - \theta||_Q \leq c).$$

Ce qui s'écrit

$$\mathbb{E}_\theta \left[\mathbb{I}_{[c,+\infty[}(||\delta_1(x) - \theta||_Q)\right] \leq \mathbb{E}_\theta \left[\mathbb{I}_{[c,+\infty[}(||\delta_2(x) - \theta||_Q)\right].$$

Comme $\ell(t) = \mathbb{I}_{[c,+\infty[}(t)$ est une fonction croissante de t, la domination universelle implique la domination stochastique. La réciproque découle du fait que deux variables aléatoires stochastiquement ordonnées ont également leurs premiers moments ordonnés. □

De plus, ces deux critères ne sont pas vides, car Hwang (1985) a établit le résultat de domination suivant : Si $x \sim \mathscr{T}_\alpha(\mu, \sigma^2)$, loi de Student à α degrés de liberté, certains estimateurs à rétrécisseur dominent universellement $\delta_0(x) = x$. Si la dimension n'est pas trop petite (normalement, $p = 4$ suffit), Brown et Hwang (1989) ont prouvé que, si $x \sim \mathscr{N}_p(\theta, \Sigma)$, l'estimateur $\delta_0(x)$ est admissible par domination universelle si et seulement si $Q = \Sigma$. Pour d'autres choix de la matrice Q et p assez grand, δ_0 est dominé stochastiquement. Cependant, même si ces critères sont moins discriminants que les coûts habituels, ils permettent d'effectuer des comparaisons, et même de faire apparaître des phénomènes de Stein (Note 2.8.2), car les estimateurs classiques ne sont pas nécessairement optimaux.

L'étude des fonctions de coût multiples n'est pas très développée d'un point de vue bayésien, car les estimateurs de Bayes varient en général avec un changement de fonction de coût. Cependant, dans des cas très particuliers, Rukhin (1978) a montré que les estimateurs de Bayes peuvent être *indépendants de la fonction de coût*. Sous certaines hypothèses de régularité, ce cas correspond aux densités vérifiant des équations de la forme

$$\log f(x|\theta) + \log \pi(\theta) = A_1(x)e^{\alpha\theta} + A_2(x)e^{-\alpha\theta} + A_3(x),$$

où π est la distribution a priori. Donc, pour cette *famille exponentielle* (voir la Section 3.3.3),

$$f(x|\theta) = \frac{B(x)}{\pi(\theta)} \exp\{A_1(x)e^{\alpha\theta} + A_2(x)e^{-\alpha\theta}\}, \tag{2.11}$$

les estimateurs de Bayes sont *universels*, parce qu'ils ne dépendent pas de la fonction de coût choisie.

2.7 Exercices

Section 2.2

2.1 Montrer que, si la fonction d'utilité de U est convexe, tout $P \in \mathscr{P}_{\mathscr{E}}$ satisfait

$$\mathbb{E}^P[r] = \int_{\mathscr{R}} r\, dP(r) \preceq P.$$

En déduire qu'une fonction de coût concave n'est pas réaliste.

2.2 Soient quatre dés avec les chiffres suivants sur leurs faces respectives : $(4,4,4,4,0,0)$, $(3,3,3,3,3,3)$, $(6,6,2,2,2,2)$, $(1,1,1,5,5,5)$. Deux joueurs lancent un dé chacun et comparent leurs résultats. Montrer que la relation *le dé* $[i]$ *l'emporte sur le dé* $[j]$ est intransitive, c'est-à-dire pour chaque choix du premier joueur, le deuxième peut choisir un dé de manière à ce que la probabilité de gagner soit supérieure à 0.5. Relier cet exemple au concept de proximité de Pitman présenté dans la Note 2.8.3.

2.3 *Montrer que $\mathscr{P}_{\mathscr{B}} \subset \mathscr{P}_{\mathscr{E}}$, c'est-à-dire que les distributions bornées ont une utilité moyenne finie.

2.4 Démontrer les Lemmes 2.4 et 2.5.

2.5 *(DeGroot, 1970) Afin de démontrer l'extension du Théorème 2.6 de $\mathscr{P}_{\mathscr{B}}$ à $\mathscr{P}_{\mathscr{E}}$, considérons une suite décroissante s_m (pour $\preceq$) dans $\mathscr{R}$ telle que, pour tout $r \in \mathscr{R}$, il existe m avec $s_m \preceq r$. Si $P \in \mathscr{P}_{\mathscr{E}}$ et si $P(\{s_m \preceq r\}) > 0$, on note P_m la distribution conditionnelle

$$P_m(A) = \frac{P(A \cap \{s_m \preceq r\})}{P(\{s_m \preceq r\})}.$$

De même, si t_n est une suite croissante de $\mathscr{R}$ telle que, pour tout $r \in \mathscr{R}$, il existe n tel que $r \preceq t_n$, on définit P^n par

$$P^n(A) = \frac{P(A \cap \{r \preceq t_n\})}{P(\{r \preceq t_n\})},$$

pour $P(\{r \preceq t_n\}) > 0$. On supposera que de telles suites existent dans $\mathscr{R}$.

a. Montrer que P^n et P_m sont inclus dans $\mathscr{P}_{\mathscr{B}}$.

On ajoute l'hypothèse supplémentaire :

(A_6) *Pour tous* $P,\ Q \in \mathscr{P}_{\mathscr{E}}$*, tels qu'il existe* $r_0 \in \mathscr{R}$ *vérifiant* $P(\{r \preceq r_0\}) = Q(\{r_0 \preceq r\}) = 1$*, l'ordre* $P \preceq Q$ *est nécessairement satisfait.*

b. Montrer que (A_6) est en fait satisfait par $\mathscr{P}_{\mathscr{B}}$.

c. Montrer que, pour tout $P \in \mathscr{P}_{\mathscr{E}}$,

$$\mathbb{E}^P[U(r)] = \lim_{m \to +\infty} \mathbb{E}^{P_m}[U(r)] = \lim_{n \to +\infty} \mathbb{E}^{P^n}[U(r)].$$

d. Soient $P \in \mathscr{P}_{\mathscr{E}}$ et $m < m_1,\ n < n_1$ tels que $P(\{s_m \preceq r\}) > 0$ et $P(\{r \preceq t_n\}) > 0$. Montrer que

$$P^n \preceq P^{n_1} \preceq P \preceq P_{m_1} \preceq P_m.$$

La deuxième hypothèse additionnelle :

(A_7) *Soient P et Q dans $\mathscr{P}_{\mathscr{E}}$. S'il existe m_0 tel que $P_m \succeq Q$ pour $m \geq m_0$, alors $P \succeq Q$. De plus, il existe n_0 tel que $P^n \preceq Q$ pour $n \geq n_0$, alors $P \preceq Q$,*
est supposée vraie ci-dessous.

e. Soient P et Q dans $\mathscr{P}_{\mathscr{E}}$ avec r_1, r_2 dans $\mathscr{R}$ tels que

$$P(\{r_1 \preceq r\}) = Q(\{r_2 \preceq r\}) = 1.$$

Montrer que $P \preceq Q$ si et seulement si $\mathbb{E}^P[U(r)] \leq \mathbb{E}^Q[U(r)]$. (*Indication :* Soient les suites P^n, P_m, et $a_m = \mathbb{E}^{P_m}[U(r)]$, $b_n = \mathbb{E}^{P^n}[U(r)]$. Utiliser l'hypothèse ($A_4$) et les questions c. et d.)

f. Déduire de la question ci-dessus que, si $P,\ Q \in \mathscr{P}_{\mathscr{E}}$, $P \preceq Q$ si et seulement si $\mathbb{E}^P[U(r)] \leq \mathbb{E}^Q[U(r)]$.

2.6 Dans le cadre de l'Exemple 2.8 sur le paradoxe de Saint-Pétersbourg, déterminer l'utilité moyenne d'un joueur pour $\delta = 1$ et $\delta = 10$. Calculer le nombre moyen de jeux qu'un joueur est prêt à jouer dans le jeu modifié.

2.7 *(Smith, 1988) Un expert a un ordre de préférence tel que les récompenses $\alpha\delta_{(x+h)} + (1-\alpha)\delta_{(x-h)}$ et x sont équivalentes, avec α indépendant de x. Montrer que la fonction d'utilité est, soit linéaire (quand $\alpha = 1/2$), soit de la forme e^{cx} ($c > 0$) ($\alpha < 1/2$), soit de la forme $1 - e^{-cx}$ ($\alpha > 1/2$).

2.8 (Raiffa, 1968) Dans un premier cas, une personne doit choisir entre un gain certain de \$10 000 (a_1) et un gain aléatoire de \$50 000 avec probabilité 0.89 et 0 sinon (a_2). Le deuxième cas est tel qu'un gain de \$50 000 avec une probabilité 0.1 (a_3) est opposé à un gain de \$10 000 avec probabilité 0.11 (a_4). Montrer que, même s'il paraît naturel de préférer a_1 à a_2 et a_3 à a_4, il n'existe pas de fonction d'utilité qui garantisse l'ordre $a_1 \preceq a_2$ et $a_3 \preceq a_4$.

2.9 Dans le cadre du paradoxe de Saint-Pétersbourg, défini dans l'Exemple 2.8, considérer les trois classes de fonction d'utilité suivantes :
(i) $U(r) = \log(\delta + r)$;
(ii) $U(r) = (\delta + r)^{\varrho}$ $(0 < \varrho < 1)$; et
(iii) $U(r) = 1 - e^{\delta + r}$.
Pour chaque classe, déterminer les prix d'entrée maximaux et le nombre optimal de jeux.

Section 2.3

2.10 (Casella, 1990) Montrer que, si la fonction r, de $\mathbb{R}_+$ dans $\mathbb{R}_+$, est concave, alors $r(t)$ est strictement décroissante et $r(t)/t$ décroissante.

2.11 Considérant la fonction de coût proposée dans l'Exemple 2.12, montrer qu'un expert parfait pour $N = 2$ domine un expert parfait pour $N = 1$. Ce même phénomène peut-il se produire pour $N = 3$?

2.12 (Smith, 1988) En utilisant les notations de l'Exemple 2.12, le *score de Brier* est défini comme la fonction de coût

$$\mathrm{L}(\theta, p) = \sum_{i=1}^{N} q_i (p_i - \theta_i)^2 + \bar{q}(1 - \bar{q}) - \sum_{i=1}^{N} q_i (p_i - \bar{q})^2,$$

avec $\bar{q} = \sum_{i=1}^{N} q_i \theta_i$, la proportion de jours pluvieux. Montrer qu'un expert parfait P_1 est meilleur qu'un expert parfait P_2 si sa "résolution"

$$R = \sum_{i=1}^{N} q_i(\theta_i - \bar{q})^2$$

est plus grande. Discuter l'expression de la fonction de coût.

2.13 Montrer que, pour une fonction de coût $\mathrm{L}(\theta, d)$ strictement croissante dans $|d-\theta|$ telle que $\mathrm{L}(\theta,\theta) = 0$, il n'existe pas de procédure statistique uniformément optimale. Donner un contre-exemple quand

$$\mathrm{L}(\theta, \varphi) = \theta(\mathbb{I}_{\mathbb{R}^*}(\theta) - \varphi)^2.$$

2.14 En relation avec l'Exemple 2.12, le *score* d'un météorologiste est la somme, tout au long de l'année, des erreurs $(\mathbb{I}_{A_{ij}} - p_i)^2$ pour tous les jours dont la probabilité p_i a été annoncée et pour lesquels A_{ij} est l'évènement qu'il pleuve effectivement. Si n_i est le nombre de jours où p_i a été prévu, montrer que le score se décompose en

$$\sum_{i=1}^{N}\sum_{j=1}^{n_i}(\mathbb{I}_{A_{ij}} - \theta_i)^2 + \sum_{i=1}^{N} n_i(\theta_i - p_i)^2.$$

2.15 *(Schervish, 1989) Soit un problème inférentiel où la probabilité p d'un évènement E doit être prédite, comme par exemple la probabilité de pluie. La réponse $\delta \in [0,1]$ d'un météorologiste est évaluée via un *score* $\mathrm{L}(E,\delta)$, qui prend la valeur $g_i(\delta) \geq 0$ si $\mathbb{I}_E = i \quad (i = 0, 1)$. Le score est dit *correct* si l'erreur moyenne

$$m(\delta) = pg_1(\delta) + (1-p)g_0(\delta)$$

est minimisée en $\delta = p$.

a. Montrer que, pour un score correct, g_0 est croissante et g_1 est décroissante.

b. Montrer que, si les g_i sont dérivables, le score est correct si et seulement si

$$-pg_1'(p) = (1-p)g_0'(1-p)$$

pour tout p dans $[0,1]$.

c. En déduire que, quand le score est correct, il existe une fonction positive h, intégrable sur $[0,1]$, telle que

$$g_0(r) = \int_{[0,r]} h(t)\,dt \qquad \text{et} \qquad g_1(r) = \int_{[1-r,1]} \frac{t}{1-t} h(t)\,dt.$$

2.16 Montrer à l'aide d'exemples discrets et continus qu'un estimateur de Bayes peut correspondre à plusieurs distributions a priori pour la même fonction de coût et, symétriquement, à plusieurs fonctions de coût pour une même loi a priori.

2.17 Deux experts doivent fournir une estimation de $p \in [0,1]$ sous la fonction de coût $(\delta - p)^2$. Ils ont pour distributions a priori respectivement π_1 et π_2, égales à $\mathscr{B}e(1,2)$ et $\mathscr{B}e(2,3)$.

a. Donner les deux estimations δ_1 et δ_2 quand les experts répondent séparément (sans observation).

b. L'expert 1 connaît la valeur de δ_2. On suppose que la quantité p est observée après coup et que le meilleur expert reçoit une amende de $(\delta_i - p)^2$, et l'autre une amende d'un montant fixe A. Montrer que la fonction de coût pour l'expert 1 est

$$(\delta_1 - p)^2 \mathbb{I}_{|\delta_1 - p| \leq |\delta_2 - p|} + A \mathbb{I}_{|\delta_1 - p| > |\delta_2 - p|}.$$

Déduire que, si A est suffisamment grand, la réponse optimale pour l'expert 1 est $\delta_1 = \delta_2$.

c. Modifier la fonction de coût ci-dessus afin de forcer l'expert 1 à donner une réponse honnête, qui est la valeur initiale δ_1.

2.18 (Raiffa et Schlaifer, 1961) Pour une fonction de coût $\mathrm{L}(\theta, d)$ donnée, définir la décision optimale comme la décision d_θ qui minimise $\mathrm{L}(\theta, d)$ pour un θ donné. Le *coût d'opportunité* est alors défini comme $\mathrm{L}^*(\theta, d) = \mathrm{L}(\theta, d) - \mathrm{L}(\theta, d_\theta)$.

a. Montrer que ceci est équivalent à supposer que $\inf_\theta \mathrm{L}(\theta, d) = 0$ pour tout θ.

b. Montrer que l'ensemble des procédures classiques (fréquentistes) optimales (au sens, respectivement, de l'admissibilité et de la minimaxité) est le même pour L et L^*.

c. Montrer que les procédures de Bayes sont les mêmes pour L et L^*.

2.19 (Raiffa et Schlaifer, 1961) Pour une fonction de coût $\mathrm{L}(\theta, d)$ et une distribution a priori π données, la *décision a priori optimale* est d^π qui minimise $\mathbb{E}^\pi[\mathrm{L}(\theta, d)]$.

a. Soit $\mathscr{D} = \{d_1, d_2\}$ et $\mathrm{L}(\theta, d_1) = 0.5 + \theta$, $\mathrm{L}(\theta, d_2) = 2 - \theta$. Donner les décisions a priori optimales quand π est $\mathscr{B}e(1,1)$ et $\mathscr{B}e(2,2)$.

b. *La valeur de l'information de l'échantillon* x est définie comme

$$\nu(x) = \mathbb{E}^\pi[\mathrm{L}(\theta, d^\pi)|x] - \mathbb{E}^\pi[\mathrm{L}(\theta, \delta^\pi(x))|x],$$

où $\delta^\pi(x)$ est un estimateur de Bayes régulier de θ. Indiquer pourquoi $\nu(x) \geq 0$ et donner la valeur de l'information de l'échantillon quand $x \sim \mathscr{B}(n, \theta)$ pour les fonctions de coût et a priori ci-dessus.

c. Quand $\Theta = \mathscr{D} = \mathbb{R}$, $x \sim \mathscr{N}(\theta, 1)$, et $\theta \sim \mathscr{N}(\theta_0, 10^2)$, montrer que la décision a priori optimale sous l'erreur quadratique est $d^\pi = \theta_0$ et que la valeur de l'information de l'échantillon est $(\theta_0 - x)^2$. Conclure en commentant la cohérence de cette notion.

2.20 Une stratégie d'investissement peut être mise en œuvre selon deux stratégies différentes, d_1 et d_2. Le profit (ou utilité) de l'investissement dépend d'un paramètre de rentabilité $\theta \in \mathbb{R}$ et vaut $U(\theta, d_i) = k_i + K_i\theta$.

a. Pour une loi a priori donnée π sur θ, quelle est la décision a priori optimale ?

b. Soit $x \sim \mathscr{N}(\theta, 1)$ et $\theta \sim \mathscr{N}(0, 10)$. Donner les stratégies a priori et a posteriori optimales. Exprimer l'amélioration apportée par l'observation de x en termes d'utilité et d'utilité espérée.

c. Si l'observation de x a un coût c_s, déterminer le coût c_s à partir duquel observer x n'est plus avantageux.

2.21 (Raiffa et Schlaifer, 1961) Dans un cadre semblable à celui de l'exercice précédent, on considère l'espace de décision $\mathscr{D} = \{d_1, d_2\}$ et le paramètre $\theta \in [0, 1]$. La fonction d'utilité est $\mathrm{L}(\theta, d_i) = k_i + K_i\theta$.

a. Si on définit $\varphi = (k_1 - k_2)/(K_1 - K_2)$, montrer que $\varphi \notin (0, 1)$ implique que l'une des deux décisions est toujours optimale. Dans les questions suivantes, nous supposons que $\varphi \in (0, 1)$.

b. Soit $x|\theta \sim \mathscr{B}(n,\theta)$ et soit $\theta \sim \mathscr{B}e(r, n'-r)$. Calculer les décisions a priori et a posteriori optimales et l'amélioration moyenne (de l'utilité) obtenue par l'observation de x.

c. Pour un coût d'observation K donné pour chaque variable aléatoire de Bernoulli, déterminer la taille d'échantillon optimale pour l'espérance moyenne.

Section 2.4.1

2.22 Démontrer le Théorème 2.14 lorsque $r(\pi)$ est fini.

2.23 Comparer δ_0 et δ^* dans l'Exemple 2.9 sous le coût $0-1$. Est-ce que ce résultat contredit le théorème de Rao-Blackwell (Théorème 2.20) ?

Section 2.4.2

2.24 Construire un exemple semblable à l'Exemple 2.19, mais où A serait forcé de se confesser d'un point de vue bayésien.

2.25 Considérer le cas où $\Theta = \{\theta_1, \theta_2\}$ et $\mathscr{D} = \{d_1, d_2, d_3\}$, pour la fonction de coût suivante

	d_1	d_2	d_3
θ_1	2	0	0.5
θ_2	0	2	1

a. Déterminer les procédures minimax.

b. Identifier la distribution a priori la moins favorable. (*Indication :* Représenter l'espace des risques associé aux trois actions de la même façon que dans l'Exemple 2.24.)

2.26 Considérer la fonction de risque suivante pour $\Theta = \{\theta_1, \theta_2\}$ et $\mathscr{D} = \{d_1, d_2, d_3\}$

	d_1	d_2	d_3
θ_1	1	2	1.75
θ_2	2	1	1.75

a. Dessiner le diagramme des risques de la même façon que dans l'Exemple 2.24 et en déduire les estimateurs minimax.

b. Déduire de cet exemple que la minimaxité n'est pas *cohérente* au sens suivant : d_1, d_2, d_3 peuvent être telles que $\max_\theta R(\theta, d_1) \geq \max_\theta R(\theta, d_3)$ et $\max_\theta R(\theta, d_2) \geq \max_\theta R(\theta, d_3)$, alors que l'estimateur minimax est de la forme $\alpha d_1 + (1-\alpha) d_2$.

Section 2.4.3

2.27 Démontrer le Lemme 2.22.

2.28 Considérer $x \sim \mathscr{B}(n,\theta)$, avec n connu.

a. Si $\pi(\theta)$ est la distribution bêta $\mathscr{B}e(\sqrt{n}/2, \sqrt{n}/2)$, donner la distribution a posteriori associée $\pi(\theta|x)$ et l'espérance a posteriori $\delta^\pi(x)$.

b. Montrer que, lorsque $\mathrm{L}(\delta,\theta) = (\theta-\delta)^2$, la fonction de risque δ^π est constante. Conclure que δ^π est minimax.

c. Comparer la fonction de risque pour δ^π avec celle de $\delta_0(x) = x/n$ pour $n = 10, 50$ et 100. Conclure sur l'intérêt de δ^π.

2.29 Démontrer les Lemmes 2.25 et 2.27.

2.30 Soient $x \sim \mathscr{N}(\theta, 1)$ et $\theta \sim \mathscr{N}(0, n)$. Montrer que le risque quadratique bayésien vaut $n/(n+1)$. Conclure sur la minimaxité de $\delta_0(x) = x$.

2.31 *Donner la densité de la distribution uniforme sur la sphère de rayon c et calculer la distribution marginale de $x \sim \mathcal{N}_p(\theta, I_p)$, lorsque θ est distribué uniformément sur cette sphère. Calculer l'espérance a posteriori $\delta^\pi(x)$ et étudier ses propriétés.

2.32 Construire un exemple équivalent à l'Exemple 2.28 lorsque $x \sim \mathcal{P}(\lambda)$, c'est-à-dire lorsque $\delta_0(x) = x$ est minimax. (*Indication :* Noter que δ_0 est un estimateur de Bayes généralisé pour $\pi(\lambda) = 1/\lambda$ et utiliser une suite de lois a priori $\mathcal{G}(\alpha, \beta)$.)

2.33 Établir les Propositions 2.32, 2.35 et 2.37.

Section 2.4.4

2.34 Dans l'Exemple 2.38, nous souhaitons prouver que le risque bayésien de $p(x)$ est fini.

a. Montrer que

$$\tau(\pi) = \int_{\mathbb{R}^2} \left\{\Phi^2(x) - 2\Phi(x)\mathbb{I}_{\theta\leq 0} + \mathbb{I}_{\theta\leq 0}\right\} \frac{e^{-(x-\theta)^2/2}}{\sqrt{2\pi}} d\theta dx$$

quand $\pi(\theta) = 1$.

b. En déduire que

$$\begin{aligned}\tau(\pi) &= \int_{-\infty}^{+\infty} \Phi(x)\Phi(-x)dx \\ &= 2\int_0^{+\infty} \Phi(x)\Phi(-x)dx\end{aligned}$$

en intégrant d'abord par rapport θ.

c. Montrer que

$$\int_0^{+\infty} \Phi(-x)dx = \int_0^{+\infty} y\frac{e^{-y^2}}{\sqrt{2\pi}}dy.$$

d. En déduire que $\tau(\pi)$ est fini.

2.35 Soit $x \sim \mathcal{N}_p(\theta, I_p)$. Une classe d'estimateurs de $||\theta||^2$ est donnée par

$$\delta_c(x) = ||x||^2 + c, \quad c \in \mathbb{R}.$$

a. Montrer que, sous le coût quadratique, δ_{-p} minimise la fonction de risque pour tout θ, au sein des estimateurs δ_c. Est-ce que ce problème d'estimation a un intérêt pratique ?

b. Comment choisir $\omega(\theta)$ de façon telle que la fonction de risque de δ_{-p} soit bornée uniformément pour le coût quadratique pondéré par $\omega(\theta)$? Conclure sur la minimaxité de δ_{-p}.

c. Montrer que δ_{-p} n'est pas admissible, et proposer un estimateur qui domine δ_{-p} uniformément.

2.36 Montrer que, sous la fonction de coût quadratique, si deux estimateurs à valeurs réelles δ_1 et δ_2 sont distincts et satisfont

$$R(\theta, \delta_1) = (\theta - \delta_1(x))^2 = R(\theta, \delta_2) = (\theta - \delta_2(x))^2,$$

l'estimateur δ_1 n'est pas admissible. (*Indication :* Considérer $\delta_3 = (\delta_1 + \delta_2)/2$ ou $\delta_4 = \delta_1^\alpha \delta_2^{1-\alpha}$.) Étendre ce résultat à toutes les fonctions de coût strictement convexes et construire un contre-exemple pour une fonction de coût non convexe.

2.37 Soit $\Theta = \{\theta_1, \theta_2\}$. On considère le cas où l'ensemble des risques est $\mathscr{R} = \{(r_1, r_2);\ (r_1-2)^2 + (r_2-2)^2 < 2, r_1 \leq 2, r_2 \leq 2\}$.

a. Tracer $\mathscr{R}$ et en déduire l'existence d'un point de minimaxité.

b. Donner les deux règles de décision admissibles pour ce problème.

c. Que peut-on dire sur l'existence de procédures bayésiennes ?

2.38 Deux experts ont des fonctions de coût différentes, données dans la table suivante pour $\mathscr{D} = \{d_1, d_2, d_3\}$ et $\Theta = \{\theta_1, \theta_2\}$.

L_1/L_2	d_1	d_2	d_3
θ_1	1/1	2.5/1.5	2/2.5
θ_2	1.5/4	2/3.5	3/3

a. Tracer les ensembles des risques pour les deux experts et identifier les procédures minimax et admissibles dans les deux cas.

b. Il y a plusieurs façons de combiner les opinions d'expert, c'est-à-dire de construire une fonction de coût unique. Pour chacun des choix suivants, donner l'ensemble des risques et les règles de décision optimales :

$$\text{(i) } L = (L_1 + L_2)/2 \qquad \text{(ii) } L = \sup(L_1, L_2) \qquad \text{(iii) } L = \sqrt{L_1 L_2}.$$

c. Pour quel choix de L les règles admissibles le sont aussi pour l'un des deux coûts initiaux ? Sous quelles conditions l'ensemble des risques est-il convexe ?

Section 2.5

2.39 *Établir les Propositions 2.41, 2.42, et 2.43. Démontrer le lemme de Shinozaki (1975) : *si δ est admissible pour le coût quadratique usuel, il l'est aussi pour tout coût quadratique*.

2.40 Soient $\pi(\theta) = (1/3)(\mathscr{U}_{[0,1]}(\theta) + \mathscr{U}_{[2,3]}(\theta) + \mathscr{U}_{[4,5]}(\theta))$ et $f(x|\theta) = \theta e^{-\theta x}$. Montrer que, sous le coût (2.5), il existe, pour tout x, des valeurs de k_1 et k_2 telles que l'estimateur de Bayes ne soit pas unique.

2.41 Établir la Proposition 2.47 et montrer que la fonction de coût L considérée dans l'Exemple 2.48 est équivalente à l'estimateur $\mathbb{I}_{H_0}(\theta)$ sous le *coût absolu*,

$$L(\theta, \delta) = |\theta - \delta|.$$

Calculer l'estimateur de Bayes associé au coût quadratique.

2.42 *(Zellner, 1986a) Soit la fonction de coût dite LINEX sur $\mathbb{R}$, définie par

$$L(\theta, d) = e^{c(\theta - d)} - c(\theta - d) - 1.$$

a. Montrer que $L(\theta, d) > 0$ et représenter ce coût comme une fonction de $(\theta - d)$ lorsque $c = 0.1, 0.5, 1, 2$.

b. Donner l'expression des estimateurs de Bayes sous cette fonction de coût.

c. Pour $x_1, \ldots, x_n \sim \mathscr{N}(\theta, 1)$ et $\pi(\theta) = 1$, donner l'estimateur de Bayes associé.

2.43 (Berger, 1985b) Soient $x \sim \mathscr{N}(\theta, 1)$, $\theta \sim \mathscr{N}(0, 1)$ et la fonction de coût

$$L(\theta, \delta) = e^{3\theta^2/2}(\theta - \delta)^2.$$

a. Montrer que $\delta^\pi(x) = 2x$.

b. Montrer que δ^π est dominé uniformément par $\delta_0(x) = x$ et que $r(\pi) = +\infty$.

2.44 Déterminer l'estimateur de Bayes associé avec le coût absolu sur $\mathbb{R}^k$,

$$\mathrm{L}(\theta,\delta) = ||\theta - \delta||.$$

2.45 Considérer les questions suivantes pour le coût entropique et le coût intrinsèque de Hellinger.

a. Montrer que $\mathrm{L_e}$ (resp. $\mathrm{L_H}$) est positive, qu'elle est nulle si $d = \theta$ et déterminer sous quelle condition $d = \theta$ est l'unique solution de $\mathrm{L_e}(\theta, d) = 0$ (resp. de $\mathrm{L_H}(\theta, d) = 0$).

b. Donner les expressions de ces deux fonctions de coût lorsque $x \sim \mathscr{N}(0,\theta)$ et $x \sim \mathscr{B}e(n,\theta)$.

c. Montrer que, si $x \sim \mathscr{G}(\alpha,\theta)$ et $\theta \sim \mathscr{G}(\nu, x_0)$, l'estimateur de Bayes de θ sous le coût de Hellinger est de la forme $k/(x_0 + x)$.

2.46 *(Wells, 1992) Comme cela est mentionné dans la Section 2.5.4, les estimateurs de Bayes ne sont pas invariants sous une reparamétrisation arbitraire. Dans le cas gaussien, $x \sim \mathscr{N}(\theta, 1)$, déterminer si les seules transformations de θ pour lesquelles les estimateurs de Bayes sont invariants sous le coût quadratique sont les transformations affines, $\eta = a\theta + b$. [*Note* : La réponse est non.]

2.47 *(Efron, 1992) Calculer les estimateurs de Bayes de θ lorsque $\theta|x \sim \mathscr{N}(\mu(x), 1)$ et lorsque la fonction de coût est *quadratique asymétrique*,

$$\mathrm{L}(\theta,\delta) = \begin{cases} \omega(\theta-\delta)^2 & \text{si } \delta < \theta, \\ (1-\omega)(\theta-\delta)^2 & \text{sinon.} \end{cases}$$

2.48 (Robert, 1996a) Montrer que les coûts entropiques et de Hellinger sont équivalents localement au coût quadratique associé à l'information de Fisher,

$$I(\theta) = \mathbb{E}_\theta \left[\frac{\partial \log f(x|\theta)}{\partial \log} \left(\frac{\partial \log f(x|\theta)}{\partial \log} \right)^t \right],$$

c'est-à-dire

$$L_e(\theta,\delta) = L_e(\theta-\delta)^t I(\theta)^{-1}(\theta-\delta) + O(\|\theta-\delta\|^2)$$

et

$$L_H(\theta,\delta) = c_H(\theta-\delta)^t I(\theta)^{-1}(\theta-\delta) + O(\|\theta-\delta\|^2),$$

où c_e et c_H sont des constantes.

2.49 Soit $y = x + \epsilon$ avec ϵ et x variables aléatoires indépendantes et $\mathbb{E}[\epsilon] = 0$.

a. Montrer que $\mathbb{E}[y|x] = x$.

b. Montrer que la réciproque n'est pas vraie : $\mathbb{E}[x|y]$ n'est pas toujours égal à y. (*Indication :* Considérer, par exemple, le cas où $x \sim p\mathscr{N}(\theta_1, 1) + (1-p)\mathscr{N}(\theta_2, 1)$ et $\epsilon \sim \mathscr{N}(0,1)$.)

Section 2.6

2.50 Montrer que, pour les distributions universelles (Rukhin, 1978), les estimateurs de Bayes sont effectivement indépendants de la fonction de coût. Dans le cas particulier où $x \sim \mathscr{G}(\nu, 1/\nu)$, identifier θ, $A_1(x)$, $A_2(x)$ et l'a priori universel $\pi(\theta)$.

Note 2.8.1

2.51 Montrer que l'estimateur de Bayes associé à la fonction de coût L^0 est l'estimateur du maximum a posteriori (MAP).

2.52 Montrer que l'estimateur de Bayes associé à la fonction de coût L^1 est le vecteur des estimateurs MAP pour chaque composante.

2.53 Si $\mathscr{D}$ est un sous-ensemble de $\{1,\ldots,N\}$, notons $\mathbf{e} = \{e_i, i \in \mathscr{D}\}$, le vecteur des classifications erronées et $m_{\mathscr{D}}$ leur nombre.

a. Montrer que $p(m_{\mathscr{D}})$ peut s'écrire

$$p(m_{\mathscr{D}}) = 1 - \prod_{i\in\mathscr{D}} (1 - e_i)\,.$$

b. Soit $q(m_{\mathscr{D}})$ la fonction qui vaut 1 si et seulement si $m_{\mathscr{D}} = |\mathscr{D}|$. Montrer que

$$q(m_{\mathscr{D}}) = \prod_{i\in\mathscr{D}} e_i\,.$$

c. Montrer que

$$p(m_{\mathscr{D}}) = \sum_{k=1}^{|\mathscr{D}|} (-1)^{k+1} \sum_{\omega\in\mathscr{P}_k(\mathscr{D})} q(m_\omega)\,.$$

Note 2.8.2

2.54 Montrer que le paradoxe de Stein ne peut avoir lieu lorsque δ_0 est un estimateur de Bayes au sens strict, quelle que soit la dimension p. [*Note* : Brown (1971) a montré que certains estimateurs de Bayes généralisés jouissent de cette propriété.]

2.55 Montrer que la constante de majoration dans le Théorème 2.52 peut être remplacée par

$$c = 2\frac{q-2\alpha}{p-q+4\beta}.$$

(*Indication :* Majorer d'abord $h^2(t,u)$ par $c(u/t)h(t,u)$.) Comparer les deux bornes.

2.56 *(Stein, 1973) Établir le lemme de Stein : *Si $x \sim \mathscr{N}(\theta,1)$ et f est continue et presque partout dérivable, alors*

$$\mathbb{E}_\theta[(x-\theta)f(x)] = \mathbb{E}_\theta[f'(x)].$$

En déduire que, si $x \sim \mathscr{N}_p(\theta,\Sigma)$, $\delta(x) = x + \Sigma\gamma(x)$, et $\mathrm{L}(\theta,\delta) = (\delta-\theta)^t Q(\delta-\theta)$, avec γ dérivable, alors

$$R(\theta,\delta) = \mathbb{E}_\theta\left[\mathrm{tr}(Q\Sigma) + 2\,\mathrm{tr}(J_\gamma(x)Q^*) + \gamma(x)^t Q^*\gamma(x)\right],$$

où $\mathrm{tr}(A)$ est la trace de A, $Q^* = \Sigma Q \Sigma$ et $J_\gamma(x)$ est la matrice formée des éléments $\frac{\partial}{\partial x_i}\gamma_j(x)$. [*Note* : Cette représentation de la fonction de risque est liée à la technique d'*estimation sans biais du risque*, qui est centrale pour la construction de conditions suffisantes de domination d'estimateurs usuels. Voir Berger, 1985b, et Johnstone, 1998.]

2.57 *(**Suite de l'Exercice 2.56**) Utiliser la représentation sans biais fournie par le lemme de Stein pour montrer que, si $x \sim \mathcal{N}_p(\theta, \Sigma)$, $\delta(x) = x + \gamma(x)$ et $\mathrm{L}(\theta, \delta) = ||\delta - \theta||^2$, l'estimateur associé à $\gamma(x) = 2(2-p)/||x||^2$ a un risque constant égal à p.

Note 2.8.3
Les exercices suivants (2.58–2.63) *traitent du critère de proximité de Pitman. Un estimateur δ_1 de θ domine* au sens de Pitman *un estimateur δ_2, ce qui est noté $\delta_1 \overset{\mathrm{P}}{\succ} \delta_2$, si, pour tout $\theta \in \Theta$,*

$$P_\theta(|\delta_1(X) - \theta| < |\delta_2(X) - \theta|) > 0.5.$$

*La notion d'*admissibilité de Pitman *en découle directement.*

2.58 *Soit un estimateur sans biais médian δ^M, qui donc satisfait

$$\forall \theta, \qquad P_\theta(\delta^M(x) \leq \theta) = 0.5.$$

a. Montrer que δ^M est le meilleur estimateur (sous le critère de Pitman) au sein des estimateurs linéaires $\delta^M(x) + K$, $K \in \mathbb{R}$.
b. Si $\theta > 0$ et $\delta^M > 0$, montrer que δ^M est aussi le meilleur estimateur (pour le critère de Pitman) au sein des estimateurs $K\delta^M$, $K > 0$.

2.59 *Soient $X = \theta U$, $\theta > 0$, $U \sim \mathcal{U}(-0.9, 1.1)$. Montrer que

$$X \overset{\mathrm{P}}{\succ} 0.9|X| \overset{\mathrm{P}}{\succ} 3.2|X| \overset{\mathrm{P}}{\succ} X \ .$$

2.60 *(Robert *et al.*, 1993b) Soit $X \sim f(x - \theta)$, avec

$$\int_{-\infty}^{0} f(u)\, du = 1/2$$

et $f(0) > 0$. Si F est la fonction de répartition de X pour $\theta = 0$, la fonction $\epsilon(\theta)$ est définie par

$$F(-\theta) = \begin{cases} P_0(0 < X < \epsilon(\theta)) & \text{si } \theta > 0, \\ 1 - P_0(0 > X > -\epsilon(\theta)) & \text{si } \theta < 0, \end{cases}$$

et $\epsilon(0) = 0$. Soit

$$\theta_1 = \mathrm{Arg}\{\min_{\theta > 0} |\theta + \epsilon(\theta)|\}, \quad \theta_2 = \mathrm{Arg}\{\min_{\theta < 0} |\theta - \epsilon(\theta)|\}.$$

La version tronquée de ϵ est définie par

$$\epsilon^*(\theta) = \begin{cases} \epsilon(\theta) & \text{si } \theta > \theta_1 \text{ ou } \theta < \theta_2 \\ \theta_1 + \epsilon(\theta_1) - \theta & \text{si } 0 < \theta < \theta_1 \\ \theta + \epsilon(\theta_2) - \theta_2 & \text{si } 0 > \theta > \theta_2. \end{cases}$$

L'ensemble A vérifie

$$(x, \theta) \in A \quad \text{ si et seulement si } \quad \theta < x \leq \theta + \epsilon^*(\theta)$$

pour $\theta > 0$, et

$$(x, \theta) \in A \quad \text{ si et seulement si } \quad \theta - \epsilon^*(\theta) \leq x < \theta.$$

pour $\theta < 0$.

a. Justifier la troncature de ϵ et représenter A dans un cas particulier où le calcul de ϵ^* est faisable.
b. Montrer que, si $\delta(x)$ est une fonction croissante telle que $(x, \delta(x)) \in A$, alors $\delta \overset{\mathrm{P}}{\succ} \delta_0(x) = x$.
c. Montrer que, si $F(c) - F(-c) = 1/2$, tout estimateur δ tel que

$$\delta(x) = 0 \quad \text{quand} \quad |x| < c\,, \tag{2.12}$$

est admissible au sens de Pitman.
d. Lorsque δ est monotone, vérifie (2.12) et est dans A, montrer que δ est admissible au sens de Pitman et domine δ_0 au sens de Pitman. Montrer que

$$c < \theta_1 + \epsilon(\theta_1) \quad \text{et} \quad -c > \theta_2 - \epsilon(\theta_2)$$

et conclure à propos de l'existence de tels estimateurs.

2.61 Soit un couple de variables aléatoires (x, y) de fonction de répartition jointe

$$F_\alpha(x, y) = \frac{xy}{1 + \alpha(1-x)(1-y)} \mathbb{I}_{[0,1]^2}(x, y).$$

a. Montrer que F_α est effectivement une fonction de répartition et en déduire la densité $f_\alpha(x, y)$.
b. Donner la distribution marginale de x et y.
c. Supposons que deux estimateurs δ_1 et δ_2 soient distribués selon $\theta^{-2} f_\alpha(\delta_1/\theta, \delta_2/\theta)$. Que peut-on dire à propos de la proximité de Pitman à θ ? (*Indication :* Calculer $P(|\delta_1 - \theta| < |\delta_2 - \theta|)$.)

2.62 *Montrer que, si $X_1, X_2 \sim f(x|\theta)$, alors

$$\overline{X} \overset{\mathrm{P}}{\succ} X_1\,.$$

Appliquer ce résultat à la loi de Cauchy. Montrer que, pour tout réel η, $\overline{X}$ est plus proche au sens de Pitman de η que X_1, même si η est quelconque. [*Note* : Cette propriété n'est pas spécifique à la proximité de Pitman, puisqu'elle est aussi satisfaite par le coût quadratique.]

2.63 *(Robert *et al.*, 1993b) Montrer que (ou utiliser directement le résultat), si $\chi^2_\alpha(p, \lambda)$ est l'α-quantile d'une distribution du khi deux décentré, $\chi^2_p(\lambda)$, il vérifie

$$p - 1 + \lambda \leq \chi^2_{0.5}(p, \lambda) \leq \chi^2_{0.5}(p, 0) + \lambda.$$

a. Déduire de cette inégalité que les estimateurs de James-Stein

$$\delta_h(x) = \left(1 - \frac{h(x)}{||x||^2}\right) x$$

dominent au sens de Pitman δ_0 lorsque $x \sim \mathcal{N}(\theta, I_p)$ et

$$0 < h(x) \leq 2(p-1).$$

b. Montrer que cette condition est aussi nécessaire lorsque h est constante.

2.8 Notes

2.8.1 Fonctions de coût pour l'analyse d'image

Une image, représentée sur l'écran d'un ordinateur, est un tableau à deux dimensions $\mathbf{x}$ contenant des pixels de couleurs différentes (ou niveaux de gris, pour les images en noir et blanc). Une image est souvent observée avec du bruit, provenant éventuellement des imperfections du dispositif d'acquisition, comme pour un appareil de photo qui n'est pas mis au point, des perturbations dans la transmission, ou de défauts de l'image elle-même, comme par exemple des nuages dans une image satellite. L'analyse d'image bayésienne cherche, entre autres choses, à reconstruire l'image initiale.

L'image observée, $\mathbf{x}$, peut aussi s'écrire sous la forme d'un vecteur $(x_1 \dots, x_N)$, chaque x_i prenant ses valeurs dans $\{0, 1, \dots, C-1\}$, l'ensemble des couleurs. La vraie image est notée θ et $\mathbf{x}$ suit la loi $\mathbf{x} \sim f(\mathbf{x}|\theta)$.

La fonction de coût la plus rudimentaire dans ce cadre est la fonction dichotomique "0–1" $\mathrm{L}^0(\theta, \delta) = 0$ si $\theta = \delta$ et $\mathrm{L}^0(\theta, \delta) = 1$ sinon. Pour un a priori $\pi(\theta)$, l'estimateur de Bayes δ^π associé au coût 0–1 est l'image qui maximise la densité a posteriori $\pi(\theta|\mathbf{x})$, dite aussi *estimateur MAP*. Comme il a été noté par Rue (1995), cette fonction de coût est extrêmement sensible aux erreurs de classification, et entraîne un surlissage de l'image, gommant de petites structures qui sont importantes dans des applications comme la reconnaissance de forme.

La seconde fonction de coût standard est le *taux d'erreur de classification*, c'est-à-dire le nombre de classifications erronées, obtenu à partir du vecteur $\mathbf{e}$, qui est défini, pour un estimateur δ et une vraie image θ, comme $e_i = \mathbb{I}_{\delta_i \neq \theta_i}$ $(i = 1, \dots, N)$. Le nombre de classifications erronées est alors

$$\mathrm{L}^1(\theta, \delta) = \sum_{i=1}^{N} e_i \,.$$

Étant donné la structure additive de cette fonction de coût, le coût a posteriori est la somme des coûts pour chaque site $\mathbb{E}[e_i|\mathbf{x}]$ et l'estimateur de Bayes est donc le vecteur des estimateurs MAP marginaux. Le défaut de cette fonction de coût est donc l'inverse de celui du coût précédent : elle entraîne une estimation trop locale et ne prend pas en compte les interactions entre des sites voisins.

Rue (1995) introduit une nouvelle famille de fonctions de coûts pour la construction d'estimateurs d'images bayésiens, qui prennent en compte les différents traits caractéristiques de l'image. Si $\mathscr{D}$ est un sous-ensemble de $\{1, \dots, N\}$, et $m_{\mathscr{D}}$ le nombre de classifications erronées dans $\mathscr{D}$,

$$m_{\mathscr{D}} = \sum_{i \in \mathscr{D}} e_i \,,$$

$p(m_{\mathscr{D}})$ vaut 0 si $m_{\mathscr{D}} = 0$, 1 sinon, $R_\phi \mathscr{D}$ est l'ensemble $\mathscr{D}$ tourné d'un angle $\phi \in \{0, \pm\pi/2, \pi\}$ et $T_s \mathscr{D}$ est $\mathscr{D}$ translaté de s (dans sa représentation à deux dimensions). Si on note $\mathscr{P}_j(\mathscr{D})$ l'ensemble des sous-ensembles $\mathscr{D}$ de taille j, les fonctions de coût sont construites à partir (i) d'un ensemble de sous-ensembles de base de $\{1, \dots, N\}$, et (ii) de coefficients de pénalité $t_{ij,}$ tels que la pénalité associée à une région $\mathscr{B}_i$ soit

$$P_i(m_{\mathscr{B}_i}) = \sum_{j=1}^{|\mathscr{B}_i|} t_{i,j} \sum_{\omega \in \mathscr{P}_j(\mathscr{B}_i)} p(m_\omega)\,.$$

La fonction de coût est alors

$$\mathrm{L}(\theta, \delta) = \sum_{i=1}^{n} \sum_{s,\phi} P_i(m_{T_s R_\phi \mathscr{B}_i})\,, \tag{2.13}$$

où la seconde somme est restreinte aux couples (s, ϕ) tels que $T_s R_\phi \mathscr{B}_i$ soit un sous-ensemble de $\{1, \ldots, N\}$, c'est-à-dire est à l'intérieur de l'image initiale.

La motivation pour recourir à une telle combinaison devient plus claire lorsque, à l'instar de Rue (1995), on prend $n = 1$ et $\mathscr{B}_1$ est la région 2×2 constituée par les quatre voisins d'un point arbitraire. Dans ce cas particulier, Rue (1995) propose de prendre $t_{1,1} = 1$ afin de pénaliser une classification erronée en un site et de choisir une pénalité supplémentaire $t_{1,2} > 0$ pour le cas où deux sites voisins sont simultanément mal classés, tandis que $t_{1,3} = t_{1,4} = 0$. La fonction de coût résultante est alors le nombre de sites mal classés, plus $t_{1,2}$ fois le nombre de couples de voisins simultanément mal classés.

Comme le détaille Rue (1995), les problèmes de résolution minimales, de reconnaissance de forme et les modèles d'Ising sont d'autres exemples de ce cadre général. Par exemple, les sous-ensembles de base $\mathscr{B}_i$ peuvent inclure des formes particulières, comme des voitures pour le contrôle du traffic, ou des tumeurs pour le traitement d'images radiologiques. Bien entendu, le calcul de l'estimateur de Bayes associé à (2.13) n'est pas aussi simple que pour L^0 et L^1, et Rue (1995) propose une méthode itérative fondée sur une chaîne de Markov (voir le Chapitre 6).

2.8.2 Le phénomène de Stein

S'il existe un unique estimateur minimax, celui-ci est admissible, selon la Proposition 2.32. Réciproquement, si un estimateur minimax δ_0 est inadmissible, il existe des estimateurs minimax qui dominent δ_0 (sous certaines conditions de régularité faible, voir Brown, 1976). En particulier, si l'estimateur minimax à risque constant est inadmissible, il s'agit du pire estimateur minimax au sens où tout autre estimateur minimax a un risque uniformément plus petit. Jusqu'en 1955, on supposait que l'estimateur des moindres carrés, $\delta_0(x) = x$, lorsque $x \sim \mathscr{N}_p(\theta, I_p)$, était admissible et, puisque sa fonction de risque était constante, qu'il s'agissait de l'unique estimateur minimax. Stein (1955a) a montré que ceci n'est vrai que pour $p = 1, 2$ et mis ainsi en lumière "le *phénomène de Stein*", c'est-à-dire le supposé paradoxe de l'inadmissibilité d'estimateurs standards.

Formellement, le paradoxe de Stein peut être exprimé de la façon suivante. Si un estimateur standard $\delta^*(x) = (\delta_0(x_1), \ldots, \delta_0(x_p))$ est évalué sous le coût quadratique pondéré

$$\sum_{i=1}^{p} \omega_i (\delta_i - \theta_i)^2, \tag{2.14}$$

où $\omega_i > 0$ $(i = 1, \ldots, p)$, il existe p_0 tel que δ^* ne soit pas admissible pour $p \geq p_0$, *bien que les composantes* $\delta_0(x_i)$ *soient, priss séparément, admissibles pour l'estimation des* θ_i. Le phénomène de Stein est dû à l'utilisation de la

fonction de coût jointe (2.14), qui permet à l'estimateur dominant de tirer profit des autres composantes, même si celles-ci sont indépendantes et correspondent à des problèmes d'estimation sans rapport entre eux.

La littérature sur le phénomène de Stein et les phénomènes qui lui sont associés est désormais trop vaste pour que nous puissions en présenter tous les résultats ici. Nous renvoyons les lecteurs à Judge et Bock (1978), Lehmann (1983) et Berger (1985b) pour une bibliographie plus détaillée. Nous développerons dans le Chapitre 10 une analyse bayésienne du phénomène de Stein. Cette note présente brièvement les résultats principaux sur le phénomène de Stein, d'un point de vue fréquentiste.

Initialement, bien que la démonstration d'inadmissibilité de Stein (1955a) soit non constructive, James et Stein (1961) exhibèrent un estimateur qui domine uniformément $\delta_0(x) = x$ sous le coût quadratique pour $p \geq 3$ dans le cas gaussien, donc tel que, pour tout θ,

$$p = \mathbb{E}_\theta[||\delta_0(x) - \theta||^2] > \mathbb{E}_\theta[||\delta^{JS}(x) - \theta||^2].$$

Cet estimateur,

$$\delta^{JS}(x) = \left(1 - \frac{p-2}{||x||^2}\right) x, \tag{2.15}$$

est désormais appelé l'*estimateur de James-Stein*. Notons le comportement curieux de δ^{JS} lorsque x tend vers 0 : Le facteur

$$1 - \frac{p-2}{||x||^2}$$

devient négatif et tend même vers $-\infty$ lorsque $||x||$ tend vers 0. Cependant, δ^{JS} domine δ_0 pour tout θ. (Ceci est une conséquence du Théorème 2.52 ci-dessous.) Baranchick (1970) corrigea ce comportement paradoxal en montrant que les estimateurs *tronqués*

$$\begin{aligned}\delta_c^+(x) &= \left(1 - \frac{c}{||x||^2}\right)^+ x \\ &= \begin{cases} (1 - \frac{c}{||x||^2})x & \text{si } ||x||^2 > c, \\ 0 & \text{sinon,} \end{cases}\end{aligned} \tag{2.16}$$

dominent uniformément leurs équivalents non tronqués pour $p-2 \leq c \leq 2(p-2)$. En particulier, δ_{p-2}^+ domine δ^{JS}. Ces estimateurs sont de plus non comparables (pour différentes valeurs de c). Cette classe d'estimateurs est importante parce que, bien qu'elle soit constituée d'estimateurs non admissibles (voir le Chapitre 8), il est difficile de construire des estimateurs qui les dominent et ces derniers ne réduisent pas de manière significative la fonction de risque (Shao et Strawderman, 1996). En revanche, les estimateurs de James-Stein tronqués (ou *positive-part*) permettent une réduction significative du risque par rapport aux estimateurs des moindres carrés, comme l'illustre la Figure 2.2 pour $p = 10$ et $c = 2p - 1$.

À la suite de James et Stein (1961), des classes plus générales d'estimateurs dominant δ_0 ont été proposées par Alam (1973), Berger et Bock (1976), Judge et Bock (1978), Stein (1981), George (1986a,b), et Brandwein *et al.* (1992).

Ces estimateurs sont appelés *estimateurs à rétrécisseur* parce que, à l'instar de (2.15) et (2.16), ils rétrécissent x vers 0. Des phénomènes de Stein ont été aussi mis en évidence pour des distributions non normales et d'autres coûts que la fonction quadratique, voir Berger (1975b), Brandwein et Strawderman (1980), Hwang (1982a), Ghosh *et al.* (1983), Bock (1985), Haff et Johnstone (1986), Srivastava et Bilodeau (1988), Brandwein et Strawderman (1990). Certaines restrictions sur les classes d'estimateurs à rétrécisseur ont été proposées, qui permettent d'intégrer les contraintes d'admissiblité (Brown, 1971, Alam, 1973, Strawderman, 1974, Brown, 1975, Berger et Srinivasan, 1978, Brown et Hwang, 1982, Das Gupta et Sinha, 1986, Brown, 1988, et Fraisse *et al.*, 1998). Bondar (1987) montre que l'amélioration (en termes de risque) apportée par les estimateurs à rétrécisseur n'est significative que sur une petite partie de l'espace des paramètres, mais George (1986a,b) montre qu'il est possible d'étendre cette région grâce au concept d'*estimateur à rétrécisseur multiple* (voir l'Exercice 10.38).

Le phénomène de Stein peut aussi être considéré comme *robuste* au sens où il dépend principalement de la fonction de coût, plutôt que de la distribution exacte des observations, comme cela a été montré par Brown (1975), Shinozaki (1980, 1984), Berger (1980b,a), Das Gupta (1958), Bilodeau (1988), Cellier *et al.* (1989), Brandwein et Strawderman (1990) ou Kubokawa *et al.* (1991, 1992, 1993b). Il ne se restreint pas à l'estimation ponctuelle, et apparaît aussi dans le cadre des régions de confiance (Stein, 1962a, Hwang et Casella, 1982, Hwang et Casella, 1984, Casella et Hwang, 1983, Casella et Hwang, 1987, Robert et Casella, 1990, Hwang et Ullah, 1994), et dans celui de l'estimation de la précision (ou du coût) (Johnstone, 1998, Rukhin, 1988a,b, Lu et Berger, 1989a,b, Robert et Casella, 1993, Fourdrinier et Wells, 1993, George et Casella, 1994). En revanche, Gutmann (1982) a établi que le phénomène de Stein ne peut avoir lieu pour des espaces de paramètres finis. Brown (1971) (voir aussi Srinivasan, 1981, Johnstone, 1984, et Eaton, 1992) a prouvé que l'admissibilité est reliée à un processus stochastique associé à l'estimateur et Brown (1980) prouve le résultat surprenant suivant, appelé *phénomène de Berger*, d'après Berger (1980a) : il existe toujours une fonction de coût telle que la *frontière* entre admissiblité et inadmissibilité pour l'estimateur standard soit une dimension p_0 arbitraire donnée.

Ce survol rapide ne fait pas justice à la richesse des travaux sur le phénomène de Stein. Les avancées dans ce domaine sur les trente dernières années ont beaucoup apporté à la Théorie de la Décision, notamment à sa branche bayésienne. En effet, une des conséquences importantes du paradoxe de Stein a été de marquer la fin de l'âge d'or de la Statistique classique, puisqu'il montre que la quête du *meilleur* estimateur, c'est-à-dire d'un estimateur minimax admissible unique, est sans espoir, à moins qu'on ne restreigne la classe des estimateurs à considérer, ou qu'on ne prenne en compte une information a priori. Les travaux sur le phénomène de Stein ont donc mené à l'abandon progressif de *l'estimation sans biais*, à une compréhension plus profonde de la minimaxité et de l'admissibilité, et à une amélioration des techniques fréquentistes de calcul de risque (poursuivant l'idée de Stein, 1973, d'une *estimation sans biais du risque*). Cependant, son apport principal a été de renforcer l'interface entre les approches

fréquentiste et bayésienne[23], en incitant les fréquentistes à recourir aux techniques bayésiennes (voir, par exemple, les estimateurs pseudo-bayésiens de Bock, 1988) et les bayésiens à rendre les estimateurs plus robustes à l'égard de leurs performances fréquentistes, et de l'incertitude portant sur le choix de l'a priori (Berger, 1980a, 1982b, 1984, George, 1986a,b, Lu et Berger, 1989a,b, Berger et Robert, 1990). Nous renvoyons les lecteurs aux livres mentionnés ci-dessus ainsi qu'à Brandwein et Strawderman (1990) et Lehmann et Casella (1998) pour des références additionnelles.

Nous concluons cette note par la démonstration de l'inadmissibilité de $\delta_0(x) = x$ pour l'estimation du paramètre θ d'une distribution *à symétrie sphérique*, de densité $f(||x-\theta||)$ sur $\mathbb{R}^p$ ($p \geq 3$). Kelker (1970), Eaton (1986) et Fan et Anderson (1990) (voir aussi l'Exercice 1.1) fournissent des références sur ces distributions généralisant la loi normale dans les modèles de régression linéaire. Ce résultat a été établi pour la première fois par Cellier *et al.* (1989).

Théorème 2.52. *Soit $z = (x^t, y^t)^t \in \mathbb{R}^p$, de loi*

$$z \sim f(||x-\theta||^2 + ||y||^2), \tag{2.17}$$

avec $x \in \mathbb{R}^q$ et $y \in \mathbb{R}^{p-q}$. Un estimateur

$$\delta_h(z) = (1 - h(||x||^2, ||y||^2))x$$

domine δ_0 sous le coût quadratique usuel s'il existe α, $\beta > 0$ tels que :
(1) *$t^\alpha h(t,u)$ est une fonction croissante de t pour tout u ;*
(2) *$u^{-\beta} h(t,u)$ est une fonction croissante de u pour tout t ; et*
(3) $0 \leq (t/u)h(t,u) \leq \dfrac{2(q-2)\alpha}{p-q-2+4\beta}$.

Les conditions sur h données ci-dessus ne font donc pas intervenir f dans (2.17), qui n'a pas besoin d'être connue ; de plus, elles sont identiques à celles obtenues dans le cas normal (voir Brown, 1975). La présence d'un phénomène de Stein est donc robuste dans la classe des distributions à symétrie sphérique admettant un coût quadratique fini.

Preuve. Les conditions (1) et (2) impliquent que

$$\begin{cases} t\frac{\partial}{\partial t}h(t,u) \geq -\alpha h(t,u), \\ u\frac{\partial}{\partial u}h(t,u) \leq \beta h(t,u). \end{cases}$$

La fonction de coût δ_h peut s'écrire :

$$\begin{aligned} R(\theta,\delta_h) &= \mathbb{E}_\theta\left[\sum_{i=1}^q \left\{x_i - \theta_i - h(||x||^2, ||y||^2)x_i\right\}^2\right] \\ &= \mathbb{E}_\theta\left[\sum_{i=1}^q (x_i-\theta_i)^2\right] - 2\mathbb{E}_\theta\left[\sum_{i=1}^q h(||x||^2, ||y||^2)x_i(x_i-\theta_i)\right] \\ &\qquad + \mathbb{E}_\theta\left[h^2(||x||^2, ||y||^2)||x||^2\right]. \end{aligned}$$

[23] Le développement des techniques *bayésiennes empiriques* en est un exemple typique, voir le Chapitre 10.

On montre par une intégration par parties que

$$\int_{-\infty}^{+\infty} h(||x||^2, ||y||^2)x_i(x_i-\theta_i)f(||x-\theta||^2+||y||^2)\,dx_i$$
$$= \int_{-\infty}^{+\infty} \frac{\partial}{\partial x_i}\left[h(||x||^2, ||y||^2)x_i\right]\bar{F}(||x-\theta||^2+||y||^2)\,dx_i,$$

où

$$\bar{F}(t) = \int_t^{+\infty} f(u)du.$$

Donc

$$\mathbb{E}_\theta\left[\sum_{i=1}^{q} h(||x||^2, ||y||^2)x_i(x_i-\theta_i)\right]$$
$$= \int_{\mathbb{R}^p}\left[qh(||x||^2, ||y||^2) + 2h_1'(||x||^2, ||y||^2)||x||^2\right]\bar{F}(||x-\theta||^2+||y||^2)\,dz\,,$$

où $h_1'(t,u) = \frac{\partial}{\partial t}h(t,u)$. De même,

$$\mathbb{E}_\theta[h^2(||x||^2, ||y||^2)||x||^2] = \mathbb{E}_\theta\left[\frac{||x||^2}{||y||^2}h^2(||x||^2, ||y||^2)||y||^2\right]$$
$$= \int_{\mathbb{R}^p} ||x||^2 \sum_{j=1}^{p-q}\frac{\partial}{\partial y_j}\left(h^2(||x||^2, ||y||^2)\frac{y_j}{||y||^2}\right)\bar{F}(||x-\theta||^2+||y||^2)\,dz$$
$$= \int_{\mathbb{R}^p} ||x||^2\Bigg[4h(||x||^2, ||y||^2)h_2'(||x||^2, ||y||^2)||x||^2$$
$$+\,(p-q-2)h^2(||x||^2, ||y||^2)\frac{1}{||y||^2}\Bigg]\bar{F}(||x-\theta||^2+||y||^2)\,dz,$$

où $h_2'(t,u) = \frac{\partial}{\partial u}h(t,u)$. La différence des risques vaut alors

$$R(\theta,\delta_0) - R(\theta,\delta_h) =$$
$$\int_{\mathbb{R}^p}\Bigg\{2\left[qh(||x||^2, ||y||^2) + 2h_1'(||x||^2, ||y||^2)||x||^2\right]||x||^2 h(||x||^2, ||y||^2)$$
$$\left[4h_2'(||x||^2, ||y||^2) - (p-q-2)h(||x||^2, ||y||^2)\frac{1}{||y||^2}\right]\Bigg\}$$
$$\times\bar{F}(||x-\theta||^2+||y||^2)\,dz$$
$$\geq \int_{\mathbb{R}^p} h(||x||^2, ||y||^2)\left[-h(||x||^2, ||y||^2)\frac{||x||^2}{||y||^2}(p-q-2+4\beta)\right.$$
$$\left.+2(q-2\alpha)\right]\bar{F}(||x-\theta||^2+||y||^2)\,dz > 0,$$

ce qui conclut la démonstration. □

Notez que ce résultat de domination inclut comme cas particulier l'estimation d'un vecteur normal moyen lorsque la variance est connue à une constante multiplicative près, soit le problème considéré initialement par James et Stein (1961). Lorsque $h(t,u) = au/t$, a est borné par $2(q-2)/(p-q+2)$, comme l'ont démontré James et Stein (1961).

2.8.3 Proximité de Pitman

Une approche alternative à la Théorie de la Décision standard a été développée par Pitman (1937). Afin de comparer deux estimateurs δ_1 et δ_2 de θ, il a proposé de comparer les distributions de leurs distances (ou *proximité*) à θ, soit,

$$P_\theta \left(||\delta_1(x) - \theta|| \leq ||\delta_2(x) - \theta|| \right).$$

Si cette probabilité est uniformément plus grande que 0.5, δ_1 *domine* δ_2 *au sens de Pitman*, avec le message implicite que δ_1 devrait alors être préféré à δ_2. Quoique formellement semblable à la domination stochastique, ce critère, dit *proximité de Pitman*, présente des défauts majeurs, et nous déconseillons son utilisation comme critère de comparaison. Néanmoins, la littérature sur ce sujet est assez vaste (voir, par exemple, Blyth, 1972b, Rao, 1980, 1981, Blyth et Pathak, 1985, Rao *et al.*, 1986, Keating et Mason, 1988, Peddada et Khattree, 1986, Sen *et al.*, 1989, Ghosh et Sen, 1989). Ces articles étudient les propriétés de la proximité de Pitman et mettent en avant son *caractère intrinsèque*, puisqu'elle fait intervenir la distribution complète de $||\delta_1(x) - \theta||$ (par opposition à l'évaluation réductrice à travers une fonction de coût, quadratique par exemple). À l'opposé, Robert *et al.* (1993b) exposent les défauts fondamentaux de ce critère. Nous présentons ici deux points caractéristiques (voir les Exercices 2.58-2.63 pour d'autres exemples).

Une première critique importante de la proximité de Pitman concerne sa *non-transitivité*. De fait, ce critère ne fournit pas de moyen de déterminer un estimateur optimal ou même de comparer des estimateurs entre eux. Pitman (1937) avait déjà remarqué cette difficulté, mais certains partisans de ce critère (voir notamment Blyth, 1972a) affirment de manière paradoxale que cette propriété est un avantage supplémentaire, puisqu'elle reflète la complexité du monde. Comme nous l'avons déjà vu, il peut effectivement arriver qu'un ordre de préférence raisonnable ne soit pas toujours transitif. Mais le besoin aigu de réduire une telle complexité mis à part, notons que la proximité de Pitman est mise en avant comme un critère de comparaison, une alternative aux fonctions de coûts usuelles : lorsqu'il y a non-transitivité, l'ordre déduit de ce critère n'est pas absolu puisque, comme le montre l'exemple suivant, il y a toujours une possibilité d'obtenir un cycle de préférence. Dans de tels cas, ce critère ne peut pas fournir d'estimateur optimal.

Exemple 2.53. Soient $U \sim \mathcal{U}_{[-0.9,1.1]}$ et $x = \theta U$. On peut alors prouver que, au sens de Pitman, $\delta_0(x) = x$ domine $\delta_1(x) = 0.9|x|$, δ_1 domine $\delta_2(x) = 3.2|x|$, et δ_2 domine δ_0. Si on doit choisir l'un de ces trois estimateurs, ce critère n'apporte aucune aide. ‖

Bien entendu, la non-transitivité du critère de Pitman l'empêche d'être équivalent à une fonction de coût ; à ce titre, il ne peut pas relever de la Théorie de la Décision. Pour la même raison, il ne peut pas être équivalent à la domination stochastique. En fait, Blyth et Pathak (1985) fournissent un exemple où ces deux critères produisent des ordres opposés. Il est de même impossible de définir un estimateur de Bayes (décisionnel) pour le critère de Pitman (bien qu'un estimateur a posteriori de Pitman puisse exister. Voir Bose, 1992 et Ghosh *et al.*, 1993).

Un second défaut majeur de la proximité de Pitman est qu'elle peut exclure certains estimateurs classiques, même si ces derniers sont admissibles sous coût quadratique. Par exemple, Efron (1975) remarque qu'il est possible de dominer $\delta_0(x) = x$ au sens de Pitman dans le cas gaussien, $x \sim \mathcal{N}(\theta, 1)$. Robert *et al.* (1993b) montrent qu'un phénomène de Stein affecte $\mathcal{N}_p(\theta, I_p)$ pour $p \geq 2$ et que la condition de domination ne fait intervenir qu'une majoration pour la fonction de rétrécissement h (voir aussi Sen *et al.*, 1989 et l'Exercice 2.63). Le résultat suivant étend celui d'Efron (1975) au cas général où $x \sim f(x - \theta)$ et θ est la médiane de la distribution (voir l'Exercice 2.60 pour une démonstration).

Proposition 2.54. *Sous les conditions ci-dessus, l'estimateur $\delta_0(x) = x$ n'est pas admissible au sens de Pitman.*

De plus, les estimateurs dominants peuvent avoir des comportements indésirables, par exemple être nuls sur de grandes parties de l'espace des observations (voir l'Exercice 2.60).

Ces multiples défauts semblent indiquer clairement que la proximité de Pitman n'est pas une alternative viable à la Théorie de la Décision. Cet échec renforce notre conviction que la Théorie de la Décision est la formalisation adéquate d'une prise de décision dans un cadre incertain.

Comme nous l'avons souligné dans l'introduction, la détermination de la fonction de coût est une étape importante de la modélisation. Cette étape est trop souvent ignorée, au profit des fonctions de coût classiques, et il serait intéressant d'étudier la robustesse de ce choix, à l'instar de celle de la distribution a priori (voir la Section 3.5). Cependant, cette difficulté pratique ne justifie pas à elle seule de recourir à des critères exotiques, comme la proximité de Pitman, intrinsèquement incohérents.

3

Des informations a priori aux lois a priori

> "In the meantime, there was so much information to gather, so many puzzles to solve. Their house was the perfect place for Moraine to find the information she needed. Except that it was not there."
>
> **Robert Jordan**, *The Great Hunt.*

3.1 La difficulté du choix d'une loi a priori

Sans conteste, le point le plus criticable et le plus critiqué de l'analyse bayésienne est le choix de la loi a priori. Car, une fois que cette loi a priori est connue, l'inférence peut être conduite d'une façon quasi mécanique en minimisant le coût a posteriori, en calculant les régions de plus forte densité a posteriori ou en intégrant les paramètres pour obtenir la distribution prédictive. La loi a priori est la clé de voute de l'inférence bayésienne et sa détermination est donc l'étape la plus importante dans la mise en œuvre de cette inférence. Dans une certaine mesure, c'est aussi la plus difficile. Évidemment, dans la pratique, il est rare que l'information a priori soit suffisamment précise pour conduire à une détermination exacte de la loi a priori, au sens où plusieurs lois de probabilité peuvent être compatibles avec cette information. Il y a plusieurs raisons pour cela : le décideur, le client ou le statisticien n'a pas forcément le temps ou les ressources (ni souvent la volonté) de chercher à construire un a priori exact (qui, de toute façon, peut tout simplement ne pas exister, au vu de l'information disponible) et doit compléter l'information partielle qu'il a rassemblée à l'aide de données subjectives afin d'obtenir une loi a priori.

Il est donc nécessaire le plus souvent de faire un choix (partiellement) arbitraire de loi a priori, ce qui peut avoir un impact considérable sur l'inférence qui en découle. En particulier, l'utilisation systématique de lois usuelles (normale, gamma, bêta, etc.) et la restriction plus forte encore aux lois *conjuguées* (définies plus loin, dans la Section 3.3) ne sont pas toujours justifiées, car la détermination subjective de la loi a priori qui en résulte se fait au prix d'un traitement analytique plus fruste du problème, puisque ignorant une partie de l'information a priori. Certaines situations requièrent cependant une détermination partiellement automatisée de la loi a priori comme dans le cas extrême où l'information a priori est complètement absente. Nous considérerons deux techniques usuelles : l'approche a priori conjuguée (Section 3.3), qui nécessite une quantité limitée d'information, et l'approche non informative (Section 3.5), qui est obtenue à partir de la distribution de l'échantillon.

Historiquement, les détracteurs du paradigme bayésien ont concentré leurs critiques sur le choix de la loi a priori, en commençant par celui effectué par Laplace. Tandis que Bayes pouvait justifier sa modélisation a priori des boules de billard par un raisonnement physique (voir la Section 1.2), la modélisation abstraite par Laplace de la distribution des boules blanches dans une urne (Exemple 1.9), ou de la proportion de garçons (Exemple 1.11), *les deux étant fondées sur le principe de la raison insuffisante.* se prêtaient plus facilement à des critiques, qui d'ailleurs n'ont pas tardé à apparaître (voir Boole, 1854, Bertrand, 1889, et Chrystal, 1891).

Ces critiques contre l'approche bayésienne ont une certaine validité au sens où elles attirent l'attention sur le fait qu'il n'y a pas une façon unique de choisir une loi a priori, et que le choix de cette loi a un impact sur l'inférence résultante. Cet impact peut être négligeable, modéré ou énorme, puisqu'il est toujours possible de choisir une loi a priori qui donnera la réponse qu'on souhaite obtenir. Mais le point essentiel est ici que, premièrement, les lois a priori non fondées fournissent des inférences a posteriori non justifiées et, deuxièmement, le concept d'une loi a priori *unique* n'a pas de sens, sauf dans des cas très particuliers. Après des années de critiques (voir la Note 1.8.1), le travail de Jeffreys (1946) sur les a priori non informatifs apparut comme un don du ciel pour la communauté bayésienne, car il propose une méthode de construction de la loi a priori directement déduite de la distribution des observations. Certains bayésiens sont cependant en désaccord avec l'utilisation de méthodes automatisées (voir, par exemple, Lindley, 1971, 1990). Plus récemment, les avancées théoriques en *robustesse* et *analyse de sensibilité* ont aussi fourni une base solide à l'analyse bayésienne dans les cas d'information a priori incomplète, tandis que l'introduction de la modélisation hiérarchique (Chapitre 10) permet de placer la sélection d'un a priori à un niveau plus éloigné, avec une diminution notable de l'impact sur l'inférence résultante.

3.2 Détermination subjective et approximations

3.2.1 Existence

À moins que le décideur (ou le statisticien) ne soit informé sur le mécanisme (physique, économique, biologique, etc.) sous-jacent de génération du paramètre θ, il est généralement très difficile de proposer une forme exacte ou même paramétrée pour la distribution a priori sur θ. En fait, dans la plupart des cas, θ n'a pas de réalité propre (intrinsèque), mais correspond plutôt à une paramétrisation de la loi décrivant le phénomène aléatoire observé. La loi π est alors un moyen de résumer l'information disponible sur ce phénomène, ainsi que l'incertitude liée à cette information. Ces situations impliquent évidemment des approximations de la vraie distribution a priori—si une *vraie* loi existe ! Effectivement, et comme cela est discuté dans le Chapitre 1, les modèles statistiques sont le plus souvent des représentations simplifiées de ces phénomènes aléatoires et, puisqu'il n'existe pas de vrai modèle—mais seulement un modèle le plus proche du phénomène pour une distance appropriée— il est conceptuellement difficile de parler de la *vraie* valeur de θ et, a fortiori, d'une *vraie* loi a priori.

Exemple 3.1. (Dupuis, 1995b) Dans une expérience de capture-recapture (les détails de ces expériences seront abordés à la Section 4.3.3) de lézards, des biologistes s'intéressent aux migrations de ces lézards entre des zones de leur territoire (autour du mont Lozère). L'information disponible auprès des biologistes sur les probabilités de capture et de survie, respectivement p_t et q_{it}, où t et i sont les indices correspondant au temps et à la région considérés, est représentée dans le Tableau 3.1 par une moyenne a priori et un intervalle de confiance a priori de 95% pour ces probabilités. Plusieurs distributions a priori sont compatibles avec cette information a priori. Par exemple, puisque la distribution bêta $\mathscr{B}e(\alpha, \beta)$ peut être caractérisée par sa moyenne et un intervalle de confiance (voir l'Exercice 3.1), le statisticien choisit la distribution a priori bêta présentée dans le Tableau 3.2. ‖

Tab. 3.1. Information a priori sur les paramètres de capture et de survie pour différents temps et sites de capture. (*Source :* Dupuis, 1995a.)

Épisode	2	3	4	5	6
Moyenne	0.3	0.4	0.5	0.2	0.2
int. 95%	[0.1,0.5]	[0.2,0.6]	[0.3,0.7]	[0.05,0.4]	[0.05,0.4]
Site	A			B	
Épisode	$t = 1, 3, 5$	$t = 2, 4$	$t = 1, 3, 5$	$t = 2, 4$	
Moyenne	0.7	0.65	0.7	0.7	
int. 95%	[0.4,0.95]	[0.35,0.9]	[0.4,0.95]	[0.4,0.95]	

Tab. 3.2. Modèle a priori de capture et de survie correspondant à l'information du Tableau 3.1. (*Source :* Dupuis, 1995a.)

Épisode	2	3	4	5	6
Dist.	$\mathscr{B}e(6,14)$	$\mathscr{B}e(8,12)$	$\mathscr{B}e(12,12)$	$\mathscr{B}e(3.5,14)$	$\mathscr{B}e(3.5,14)$

Site	A		B	
Épisode	t=1,3,5	t=2,4	t=1,3,5	t=2,4
Dist.	$\mathscr{B}e(6.0,2.5)$	$\mathscr{B}e(6.5,3.5)$	$\mathscr{B}e(6.0,2.5)$	$\mathscr{B}e(6.0,2.5)$

Exemple 3.2. Un décideur veut modéliser les distributions des observations et du paramètre comme des lois normales : $x_1,\ldots,x_n \sim \mathscr{N}(\theta,1)$ et $\theta \sim \mathscr{N}(\mu,\tau)$. Puisque la moyenne a posteriori de θ est

$$\delta^\pi(x_1,\ldots,x_n) = \frac{\bar{x}\tau + \mu/n}{\tau + 1/n},$$

l'hyperparamètre τ^{-1} se comporte comme n, la taille de l'échantillon, et μ comme $\bar{x}$, la moyenne de l'échantillon. Ces hyperparamètres peuvent donc être approchés en comparant la quantité d'information apportée par (μ,τ) à celle apportée par un échantillon ; par exemple, en considérant que la moyenne (connue) μ est la moyenne d'un *échantillon virtuel de taille* $1/\tau$. ‖

D'un point de vue formel, il est possible de construire une distribution a priori de la même façon que pour les fonctions d'utilité dans le chapitre précédent, c'est-à-dire en déterminant une échelle des vraisemblances respectives des valeurs du paramètre θ. Quand cette échelle est *cohérente*, c'est-à-dire respecte les axiomes donnés ci-dessous, l'existence d'une distribution a priori peut être déduite. L'existence d'une distribution a priori subjective comme résultat d'un ordre des vraisemblances relatives est très important, car il nous permet d'échapper au cadre restrictif des justifications fréquentistes qui n'est pas toujours applicable à ce type de situations. Nous donnons dans la Note 3.8.1 les axiomes sur lesquels se fonde la preuve de l'existence d'une distribution a priori à partir d'un ordre des vraisemblances et renvoyons les lecteurs à DeGroot (1970, Chapitre 6) pour un traitement plus approfondi (voir aussi Jeffreys, 1961 et Bernardo et Smith, 1994).

Il arrive souvent que la détermination subjective d'une distribution a priori conduise à des incohérences dans l'ordre des vraisemblances, pour des raisons psychologiques, mais aussi parce que la capacité des individus à identifier des petites probabilités est assez limitée. À ce sujet, ainsi que sur la construction pratique d'une distribution de probabilité et l'évaluation de prévisionnistes, nous renvoyons les lecteurs à DeGroot et Fienberg (1983), Dawid (1984), Lindley (1985) et Smith (1988).

3.2.2 Approximations de la loi a priori

Quand l'espace des paramètres Θ est *fini*, il est souvent possible d'obtenir une évaluation subjective des probabilités des différentes valeurs de θ. Parfois, on peut utiliser des expériences précédentes du même type, mais ce n'est pas toujours le cas. Pensons, par exemple, à l'obtention de la loi d'un incident nucléaire majeur ! Plus fondamentalement, cette approche fréquentiste mène à se poser la question conceptuelle de la répétabilité des expériences *(Les cadres expérimentaux sont-ilstoujours les mêmes ? Une expérience peut-elle n'avoir aucun effet sur l'expérience suivante ?)*. Jeffreys (1961) fournit une critique détaillée de cette hypothèse.

Quand l'espace des paramètres Θ n'est pas dénombrable, par exemple, lorsqu'il s'agit d'un intervalle, la détermination subjective de la loi a priori π est évidemment beaucoup plus compliquée. En général, une première approximation de π est obtenue par le partitionnement de Θ en différents ensembles (par exemple des intervalles) et la détermination de la probabilité de chaque ensemble ; $\pi(\theta)$ est alors approchée par un *histogramme*. Une autre démarche consiste à sélectionner des éléments significatifs de Θ, à évaluer leurs vraisemblances respectives et à en déduire une courbe de vraisemblance proportionnelle à π. Dans les deux cas, une difficulté majeure se présente lorsque Θ n'est pas *borné*. En effet, il est alors nécessaire de construire les *queues* de la distribution et il est assez difficile d'évaluer subjectivement les probabilités des régions extrêmes de l'espace des paramètres ; c'est d'autant plus gênant que la forme et les propriétés des estimateurs résultants dépendent fortement de ces queues (voir l'Exemple 3.5).

Quand aucune information directe n'est disponible sur θ, une alternative est de recourir à la *distribution marginale* de x,

$$m(x) = \int_{\Theta} f(x|\theta)\pi(\theta)\, d\theta$$

afin d'obtenir de l'information sur π. Plusieurs techniques ont été proposées dans la littérature (voir Berger, 1985b, Section 3.5) ; en plus de la méthode des moments, nous pouvons citer l'entropie maximale et les méthodes *ML-II* (Good, 1983). Le principe de cette construction est que le phénomène aléatoire observé peut dans certains cas être incorporé dans une classe plus large (ou *méta modèle*) pour laquelle une information est disponible. Par exemple, si θ est la moyenne journalière de production de lait pour une vache laitière donnée, une information sur θ peut être obtenue à partir de la production du troupeau auquel appartient la vache, bien que ces observations proviennent de la distribution marginale. Cette perspective est au cœur des *modèles hiérarchiques* (Chapitre 10) et elle permet de résoudre la difficulté de la répétabilité des expériences mentionnée ci-dessus.

3.2.3 Lois a priori d'entropie maximale

Si certaines caractéristiques de la loi a priori sont connues (moments, quantiles, etc.), en supposant qu'elles peuvent s'écrire comme des espérances a priori ($k = 1, \ldots, K$),

$$\mathbb{E}^{\pi}[g_k(\theta)] = \omega_k\,, \tag{3.1}$$

une façon de choisir un a priori qui satisfait ces contraintes est la méthode de l'*entropie maximale*, développée par Jaynes (1980, 1983).

Dans un cadre fini, l'*entropie* est définie comme

$$\mathfrak{E}(\pi) = -\sum_i \pi(\theta_i)\log\{\pi(\theta_i)\}\,.$$

Cette quantité a été introduite par Shannon (1948) comme une mesure de l'incertitude en théorie de l'information et en traitement du signal. L'a priori π qui maximise l'entropie minimise, dans ce sens théorico-informatif, l'information a priori apportée par π sur θ. La *distribution d'entropie maximale*, sous les contraintes de moments (3.1), est la distribution associée à la densité

$$\pi^*(\theta_i) = \frac{\exp\left\{\sum_1^K \lambda_k g_k(\theta_i)\right\}}{\sum_j \exp\left\{\sum_1^K \lambda_k g_k(\theta_j)\right\}},$$

les nombres λ_k étant obtenus à partir de (3.1) comme des multiplicateurs de Lagrange. Par exemple, sans contrainte sur π, la distribution d'entropie maximale est la distribution uniforme sur Θ. (Cette propriété révèle un problème de fond de la méthode, car les lois a priori d'entropie maximale ne sont pas invariantes par reparamétrisation ; voir la Section 3.5.1.)

L'extension au cas continu est plus délicate, car elle implique le choix d'une mesure de référence π_0, qui peut être caractérisée comme la distribution complètement non informative. Il s'agit en effet de l'a priori d'entropie maximale en l'absence de contrainte. Cette mesure de référence peut être obtenue de plusieurs façons (voir la Section 3.5) et la distribution d'entropie maximale dépend de ce choix. Quand une structure de groupe est disponible pour le problème d'intérêt (et acceptée comme une partie de l'information a priori), on convient généralement que la mesure de Haar invariante à droite associée à ce groupe est un choix acceptable pour π_0. (Les justifications pour un tel choix sont données dans le Chapitre 9.) Une fois la mesure de référence π_0 choisie, l'entropie de π est définie par

$$\mathfrak{E}(\pi) = \mathbb{E}^{\pi_0}\left[\log\left(\frac{\pi(\theta)}{\pi_0(\theta)}\right)\right] = \int \log\left(\frac{\pi(\theta)}{\pi_0(\theta)}\right)\pi_0(d\theta),$$

qui est aussi la distance de Kullback-Leibler entre π et π_0. Dans ce cas, la distribution d'entropie maximale sous (3.1) est donnée par la densité

$$\pi^*(\theta) = \frac{\exp\left\{\sum_1^K \lambda_k g_k(\theta)\right\}\pi_0(\theta)}{\int \exp\left\{\sum_1^K \lambda_k g_k(\eta)\right\}\pi_0(d\eta)}, \tag{3.2}$$

ce qui prouve l'importance de π_0. Notons que les distributions π^* ci-dessus appartiennent formellement à une famille exponentielle (voir la Section 3.3.3).

En plus de la dépendance à π_0 exhibée par (3.2) et du manque d'invariance par reparamétrisation, un autre inconvénient de la méthode d'entropie maximale est que les contraintes (3.1) ne sont pas toujours suffisantes pour obtenir une distribution sur θ. Signalons que c'est souvent le cas quand les caractéristiques (3.1) sont liées aux *quantiles*, car les fonctions $g_k(\theta)$ sont alors de la forme $\mathbb{I}_{(-\infty,a_k]}(\theta)$ ou $\mathbb{I}_{(b_k,\infty]}(\theta)$.

Exemple 3.3. Soit θ, un paramètre réel tel que $\mathbb{E}^\pi[\theta] = \mu$. Si la mesure de référence π_0 est la mesure de Lebesgue sur $\mathbb{R}$, l'a priori d'entropie maximale satisfait $\pi^*(\theta) \propto e^{\lambda\theta}$ et ne peut pas être normalisé comme une distribution de probabilité. En revanche, si on sait de plus que $\text{var}(\theta) = \sigma^2$, la loi a priori d'entropie maximale correspondante est

$$\pi^*(\theta) \propto \exp\{\theta\lambda_1 + \theta^2\lambda_2\},$$

soit donc la distribution normale $\mathcal{N}(\mu, \sigma^2)$. ‖

Seidenfeld (1987) et Kass et Wasserman (1996) avancent des critiques supplémentaires sur l'approche par entropie maximale (Exercice 3.2).

3.2.4 Approximations paramétriques

Une alternative fréquemment utilisée pour construire un a priori continu consiste à restreindre arbitrairement le choix de π à une famille de densités *paramétrées* et à déterminer les paramètres correspondants via les *moments* ou via les *quantiles*, cette seconde option étant plus robuste. Par exemple, des évaluations subjectives de la médiane et du quantile à 75% sont suffisantes pour identifier les deux paramètres d'une distribution normale. (Voir aussi l'Exemple 3.1.)

Exemple 3.4. Soit $X_i \sim \mathcal{B}(n_i, p_i)$ le nombre d'étudiants réussissant l'examen d'introduction à l'analyse, dans une classe de n_i étudiants. Les années précédentes, la moyenne des p_i a été de 0.70, avec une variance de 0.1. Si nous supposons que les p_i sont tous générés selon la même distribution bêta, $\mathcal{B}e(\alpha, \beta)$, les paramètres α et β sont estimés par

$$\frac{\alpha}{\alpha+\beta} = 0.7, \qquad \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = 0.1,$$

soit $\alpha = 0.77$ et $\beta = 0.33$, ce qui conduit à la distribution a priori

$$p \sim \mathscr{B}e(0.77, 0.33).$$

Dans ce cas, le choix de la distribution bêta est motivé par son caractère conjugué (voir la Section 3.3). ‖

La méthode des moments est souvent difficilement applicable et engendre parfois des valeurs impossibles des paramètres, comme par exemple des variances négatives. Cependant, un inconvénient plus grave de la plupart des approches paramétriques est que la sélection de la famille paramétrée est fondée sur la simplicité du traitement mathématique et non sur des bases subjectives comme un histogramme préliminaire approchant π. Cette approche peut même provoquer un rejet partiel de l'information disponible, parce qu'elle n'est pas compatible avec la distribution paramétrée. Ainsi, dans les Exemples 3.1 et 3.4, la connaissance a priori supplémentaire de la médiane peut empêcher l'utilisation d'une distribution bêta. En réalité, la construction d'une distribution à partir d'un histogramme peut aussi être trompeuse, car différentes familles peuvent correspondre au même histogramme et mener malgré tout à des inférences assez différentes. (Néanmoins, nous étudierons dans la prochaine section une méthode particulière de détermination de loi a priori paramétrée, car les cas où l'information est limitée nécessitent une telle approche.)

Exemple 3.5. (Berger, 1985b) Soit $x \sim \mathscr{N}(\theta, 1)$. Supposons que la médiane a priori de θ soit 0, que le premier quartile a priori soit -1, et que le troisième quartile a priori soit $+1$. Alors, si la distribution a priori sur θ est de la forme $\mathscr{N}(\mu, \tau)$, nous devons avoir $\theta \sim \mathscr{N}(0, 2.19)$. En revanche, le choix d'une distribution de Cauchy mène à $\theta \sim \mathscr{C}(0, 1)$. Sous une perte quadratique, l'estimateur de Bayes devrait être, dans le premier cas,

$$\delta_1^\pi(x) = x - \frac{x}{3.19}$$

et

$$\delta_2^\pi(x) \approx x - \frac{x}{1 + x^2}$$

dans le deuxième cas pour $|x| \geq 4$ (voir Berger et Srinivasan, 1978). Par conséquent, pour $x = 4$, qui est une observation assez compatible avec l'information a priori dans les deux cas, les deux estimations devraient être $\delta_1^\pi(4) = 2.75$ et $\delta_2^\pi(4) = 3.76$! La Figure 3.1 compare les deux estimateurs pour une série de valeurs de x, le calcul de δ_2^π étant fait par la méthode de Monte Carlo (voir le Chapitre 6). ‖

Ces différences de résultats démontrent la nécessité de conduire des tests sur la validité (ou *robustesse*) des lois a priori choisies, tests dépendants des observations, afin d'évaluer à quel point un léger changement dans la distribution a priori influe sur l'inférence sur les paramètres d'intérêt. (La Section

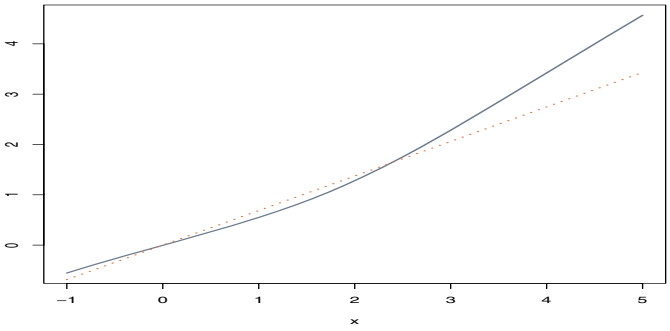

Fig. 3.1. Comparaison des estimateurs $\delta_1^\pi(x)$ *(pointillés)* et $\delta_2^\pi(x)$ *(traits pleins).*

3.5 traite de cette évaluation.) L'exemple ci-dessous illustre à nouveau le fait qu'une information trop vague peut mener à des conclusions très différentes, selon la façon dont elle est interprétée.

Tab. 3.3. Étendue des valeurs des moments a posteriori pour des moments a priori $\mu_1 = 0$ et μ_2 fixés. (*Source :* Goutis, 1990.)

μ_2	x	Moyenne minimale	Moyenne maximale	Variance maximale
3	0	-1.05	1.05	3.00
3	1	-0.70	1.69	3.63
3	2	-0.50	2.85	5.78
1.5	0	-0.59	0.59	1.50
1.5	1	-0.37	1.05	1.97
1.5	2	-0.27	2.08	3.80

Exemple 3.6. (Goutis, 1990, 1994) Soit $x \sim f(x|\theta)$, avec $\theta \in \mathbb{R}$, et supposons que la moyenne a priori de θ, μ_1, soit connue. Trop de distributions a priori s'accordent avec cette information, car

$$\inf_\pi \mathbb{E}^\pi[\theta|x] = -\infty \qquad \text{et} \qquad \sup_\pi \mathbb{E}^\pi[\theta|x] = +\infty$$

et aucune inférence utile ne peut être menée à partir de cette seule information ; notons que dans ce cas il est aussi impossible de construire une distribution d'entropie maximale (voir l'Exemple 3.3). Si, de plus, la variance a priori μ_2 est fixée, la variabilité des réponses a posteriori est plus restreinte, car

$$-\infty < \inf_\pi \mathbb{E}^\pi[\theta|x] \leq \sup_\pi \mathbb{E}^\pi[\theta|x] < +\infty, \tag{3.3}$$

tant que $f(x|\theta)$ est positive dans un voisinage de μ_1 et bornée quand $|\theta - \mu_1|$ est grand. Sous les mêmes hypothèses, nous avons de plus

$$0 = \inf_{\pi} \mathrm{Var}^{\pi}[\theta|x] \leq \sup_{\pi} \mathrm{Var}^{\pi}[\theta|x] < +\infty\,. \tag{3.4}$$

Le Tableau 3.3 donne l'étendue exacte des bornes (3.3) et (3.4) pour une distribution normale $\mathscr{N}(\theta, 1)$ et $\mu_1 = 0$. ‖

3.2.5 Autres techniques

Les techniques de Bayes dites *empiriques* et *hiérarchiques* sont deux approches relativement opposées qui intègrent l'incertitude sur la distribution a priori d'une façon naturelle et qui seront traitées en détail dans le Chapitre 10 (voir aussi Carlin et Louis, 2000a). L'approche bayésienne empirique se fonde sur les observations (et la distribution marginale) pour estimer les paramètres de la distribution a priori ; elle est utilisée plus souvent par les fréquentistes que par les bayésiens, car elle n'obéit pas au paradigme bayésien. Formellement, il semble paradoxal de choisir *a posteriori* une distribution *a priori* ! Plus fondamentalement, le choix de π dépendant de x, les estimateurs obtenus ne bénéficient pas des propriétés d'optimalité des vrais estimateurs de Bayes. Une dernière critique est qu'il existe de trop nombreuses possibilités pour les techniques d'estimations utilisées dans la construction des distributions a priori, ce qui donne par conséquent un caractère fortement arbitraire à la sélection d'un a priori.

L'approche hiérarchique bayésienne modélise le manque d'information sur les paramètres d'une distribution a priori en recourant au paradigme de Bayes, c'est-à-dire en spécifiant une autre distribution sur ces paramètres (les paramètres de cette distribution sont appelés *hyperparamètres* et ces nouveaux a priori, des lois *hyper a priori*). Bien que ce choix puisse paraître conceptuellement trop abstrait, les bayésiens préfèrent généralement cette approche à l'alternative empirique, car, dans un sens pratique et théorique, celle-ci fournit de meilleurs estimateurs. (Le Chapitre 10 présente et compare ces deux techniques.)

3.3 Lois a priori conjuguées

3.3.1 Introduction

Quand l'information a priori sur le modèle est trop vague ou peu fiable, une construction subjective complète de la distribution a priori est évidemment impossible. D'autres raisons (retards, coûts à respecter, manque de communication entre statisticiens et décideurs, etc.) peuvent expliquer l'absence

de distributions correctement définies. De plus, des exigences d'objectivité peuvent forcer le statisticien à fournir une réponse aussi neutre que possible, afin de fonder l'inférence sur le modèle d'échantillonnage uniquement. De tels cas semblent justifier le recours à des solutions non bayésiennes (estimateurs du maximum de vraisemblance, estimateurs sans biais optimaux, etc.). Cependant, tout en gardant à l'esprit les fondements bayésiens des critères fréquentistes d'optimalité (voir les Chapitres 2, 8 et 9), il paraît préférable de suivre l'approche bayésienne, en utilisant un a priori dit objectif, c'est-à-dire construit à partir du modèle d'échantillonnage, comme un outil technique. Lorsque aucune information a priori n'est disponible, ces a priori sont dits *non informatifs* et sont traités dans la Section 3.5.

D'abord, nous étudierons dans cette section une approche paramétrique classique qui implique un apport d'information subjective le plus limité possible et qui est à la base des deux techniques bayésiennes, hiérarchique et empirique, du Chapitre 10. En dehors de l'exigence d'une contribution subjective minimale, les lois a priori conjuguées peuvent être considérées comme un point de départ pour l'élaboration de distributions a priori fondées sur une information a priori limitée, dont l'imprécision peut être déterminée grâce à des distributions a priori supplémentaires. Cependant, il faut garder à l'esprit le fait que l'impression commune que les lois conjuguées sont non informatives est fausse : le choix d'un a priori conjugué, bien qu'il soit défendable comme on le verra ci-dessous, est toujours un choix particulier et influence donc, dans une certaine mesure, l'inférence résultante. De plus, il peut obliger à ignorer une partie de l'information a priori si cette dernière n'est pas complètement compatible avec la structure de la loi a priori conjuguée. Enfin il existe d'autres lois a priori fondées sur la même information subjective limitée, mais avec une influence plus limitée sur l'inférence résultante (voir la Section 3.6).

Définition 3.7. *Une famille $\mathscr{F}$ de distributions de probabilité sur Θ est dite* conjuguée *(ou* fermée par échantillonnage*) par une fonction de vraisemblance $f(x|\theta)$ si, pour tout $\pi \in \mathscr{F}$, la distribution a posteriori $\pi(\cdot|x)$ appartient également à $\mathscr{F}$.*

Un exemple trivial d'une famille conjuguée est l'ensemble $\mathscr{F}_0$ de toutes les lois de probabilité sur Θ, qui est bien entendu inutile pour le choix d'une loi a priori. L'intérêt principal du caractère conjugué devient plus évident quand $\mathscr{F}$ est *paramétrée*. Effectivement, le passage de distribution a priori à distribution a posteriori se réduit dans ce cas à une mise à jour des paramètres correspondants. Cette seule propriété peut expliquer pourquoi les lois a priori conjuguées sont si populaires, car les distributions a posteriori sont toujours calculables (au moins jusqu'à un certain degré). En revanche, une telle justification est plutôt faible d'un point de vue subjectif et d'autres familles pourraient aussi bien convenir. Notons que l'objectif d'obtenir la *famille conjuguée minimale* comme l'intersection de toutes les familles conjuguées est malheureusement voué à l'échec, car cette intersection est vide (Exercice 3.13).

3.3.2 Justifications

L'approche a priori conjuguée, introduite par Raiffa et Schlaifer (1961), peut être justifiée partiellement par un raisonnement d'*invariance*. En fait, quand l'observation de $x \sim f(x|\theta)$ modifie $\pi(\theta)$ en $\pi(\theta|x)$, l'information transmise par x sur θ est évidemment limitée ; par conséquent, elle ne devrait pas entraîner une modification de toute la *structure* de $\pi(\theta)$, mais simplement de ses *paramètres*. En d'autres termes, la modification résultant de l'observation de x devrait être de dimension finie. Un changement plus radical de π est alors inacceptable et le choix des lois a priori devrait toujours être fait parmi les lois conjuguées, quelle que soit l'information a priori. D'une certaine façon, de Finetti (1974) avait un avis similaire parce qu'il considérait que l'information a priori pouvait être interprétée comme des observations passées virtuelles, comme dans l'Exemple 3.2, ce qui mène forcément à des lois a priori conjuguées pour des familles exponentielles (voir ci-dessous). Malheureusement, cette condition devient paradoxale dans les cas extrêmes où toute la distribution a priori est déjà disponible ! Mais les lois a priori conjuguées sont surtout utilisées dans des environnements où l'information est limitée, car elles ne nécessitent la détermination que de quelques paramètres. Une autre justification pour utiliser les lois a priori conjuguées est que certains estimateurs de Bayes sont alors linéaires, comme l'ont montré Diaconis et Ylvisaker (1979) (voir la Proposition 3.19 ci-dessous). Néanmoins, nous devons reconnaître que la principale motivation pour utiliser les lois a priori conjuguées reste la commodité de traitement.

Cette modélisation particulière par une famille paramétrée de lois a priori est effectivement très tentante, car elle autorise des manipulations explicites des lois a posteriori. Ces lois a priori conjuguées sont parfois appelées *objectives* parce que le modèle d'échantillonnage, $f(x|\theta)$, détermine entièrement la classe des lois a priori, mais toute méthode qui produit de façon automatique des lois a priori à partir de la distribution d'échantillonnage serait tout aussi objective. A contrario, leur utilisation est fortement critiquée par certains bayésiens, car elle obéit à des contraintes techniques plutôt qu'à des impératifs d'adéquation à l'information a priori disponible. Le rôle des lois a priori conjuguées est alors de fournir une première approximation de la distribution a priori adéquate, qui devrait être suivie d'une analyse de robustesse (voir la Section 3.5). Nous verrons dans la Section 3.4 qu'elles sont plus justifiées si on les traite comme une base (dans un sens *fonctionnel*) pour la modélisation de l'information a priori.

3.3.3 Familles exponentielles

Les lois a priori conjuguées sont généralement associées à un type particulier de lois d'échantillonnage qui permet toujours leur obtention ; il est même caractéristique des lois a priori conjuguées comme nous le verrons ci-

dessous. Ces lois constituent ce qu'on appelle des *familles exponentielles* et sont étudiées en détail dans Brown (1986b).

Définition 3.8. *Soient μ une mesure σ-finie sur $\mathcal{X}$, Θ l'espace des paramètres, C et h des fonctions respectivement de $\mathcal{X}$ et Θ dans $\mathbb{R}_+$, et R et T des fonctions de Θ et $\mathcal{X}$ dans $\mathbb{R}^k$. La famille des distributions de densité (par rapport à μ)*

$$f(x|\theta) = C(\theta)h(x)\exp\{R(\theta)\cdot T(x)\} \tag{3.5}$$

est dite famille exponentielle de dimension k. *Dans le cas particulier où $\Theta \subset \mathbb{R}^k$, $\mathcal{X} \subset \mathbb{R}^k$ et*

$$f(x|\theta) = C(\theta)h(x)\exp\{\theta \cdot x\}, \tag{3.6}$$

la famille est dite naturelle.

Notons qu'un changement de variable de x en $z = T(x)$ et une reparamétrisation de θ en $\eta = R(\theta)$ nous permettent de considérer principalement la forme naturelle (3.6), bien que les espaces $T(\mathcal{X})$ et $R(\Theta)$ puissent être difficiles à décrire et à utiliser.

D'un point de vue analytique, les familles exponentielles ont certaines caractéristiques intéressantes (voir Brown, 1986b). En particulier, elles sont telles que, pour tout échantillon de (3.5), il existe une statistique exhaustive de dimension constante. En effet, si $x_1, \ldots, x_n \sim f(x|\theta)$, avec f satisfaisant (3.6),

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \in \mathbb{R}^k$$

est exhaustive pour tout n. La réciproque de ce résultat a été aussi établie par Koopman (1936) et Pitman (1936) (voir aussi Jeffreys, 1961, Section 3.7.1 pour une preuve).

Théorème 3.9. (Lemme de Pitman-Koopman) *Si une famille de lois $f(\cdot|\theta)$ à support constant est telle que, à partir d'une taille d'échantillon suffisamment grande, il existe une statistique exhaustive de taille fixe, la famille est exponentielle.*

La restriction sur le support de $f(\cdot|\theta)$ est une condition nécessaire pour le lemme parce que les distributions uniforme $\mathcal{U}([-\theta,\theta])$ et de Pareto $\mathcal{P}(\alpha,\theta)$ satisfont aussi cette propriété (voir l'Exemple 3.16). En réalité, ces distributions pourraient être appelées *familles quasi exponentielles*, car elles héritent de plusieurs des propriétés intéressantes des familles exponentielles, incluant l'existence de statistiques suffisantes de dimension constante et de lois conjuguées (Exercice 3.15).

De nombreuses distributions usuelles continues et discrètes appartiennent à des familles exponentielles.

Exemple 3.10. Si $\mathscr{S}_k$ est le simplexe de $\mathbb{R}^k$,

$$\mathscr{S}_k = \left\{ \omega = (\omega_1, \ldots, \omega_k); \sum_{i=1}^{k} \omega_i = 1,\, \omega_i > 0 \right\},$$

la *loi de Dirichlet* sur $\mathscr{S}_k$, $\mathscr{D}_k(\alpha_1, \ldots, \alpha_k)$, est une extension de la distribution bêta, définie comme

$$f(p|\alpha) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} \prod_{i=1}^{k} p_i^{\alpha_i - 1} \mathbb{I}_{\mathscr{S}_k}(p),$$

où $p = (p_1, \ldots, p_k)$. Puisque

$$f(p|\alpha) = C(\alpha) h(p) \exp\big(\sum_{i=1}^{k} \alpha_i \log(p_i) \big),$$

la loi de Dirichlet constitue une famille naturelle exponentielle pour $T(p) = (\log(p_1), \ldots, \log(p_k))$. ‖

Exemple 3.11. Soit $x \sim \mathscr{N}_p(\theta, \sigma^2 I_p)$. Alors

$$\begin{aligned} f(x|\theta) &= \frac{1}{\sigma^p} \frac{1}{(2\pi)^{p/2}} \exp\left\{ -\sum_{i=1}^{p} (x_i - \theta_i)^2 / 2\sigma^2 \right\} \\ &= C(\theta, \sigma) h(x) \exp\{x.(\theta/\sigma^2) + ||x||^2 (-1/2\sigma^2)\} \end{aligned}$$

et la distribution normale appartient à une famille exponentielle de paramètres naturels θ/σ^2 et $-1/2\sigma^2$. De la même façon, si $x_1, \ldots, x_n \sim \mathscr{N}_p(\theta, \sigma^2 I_p)$, la distribution jointe satisfait

$$\begin{aligned} f(x_1, \ldots, x_n) &= C'(\theta, \sigma) h'(x_1, \ldots, x_n) \\ &\quad \times \exp\left\{ n\bar{x} \cdot (\theta/\sigma^2) + \sum_{i=1}^{n} ||x_i - \bar{x}||^2 (-1/2\sigma^2) \right\} \end{aligned}$$

et la statistique $(\bar{x}, \sum_i ||x_i - \bar{x}||^2)$ est exhaustive pour tout $n \geq 2$. ‖

Dans l'exemple précédent, notons que l'espace des paramètres est de dimension $p + 1$, tandis que la dimension des observables, x, est p. Bien que la dimension d'une famille exponentielle ne soit pas fixée, car il est toujours possible d'ajouter des combinaisons convexes des paramètres originaux comme des paramètres supplémentaires (et évidemment inutiles), une dimension minimale intrinsèque est associée à cette famille.

Définition 3.12. *Soit* $f(x|\theta) = C(\theta)h(x)\exp(\theta.x)$, *une famille exponentielle naturelle. L'espace naturel des paramètres est*

$$N = \left\{\theta; \int_{\mathscr{X}} e^{\theta.x} h(x)\, d\mu(x) < +\infty\right\}.$$

La famille est dite régulière *si* N *est un ensemble ouvert et* minimale *si* $\dim(N) = \dim(K) = k$, *où* K *est la clôture de l'enveloppe convexe du support de* μ.

Il est toujours possible de réduire une famille exponentielle à une forme standard et minimale de dimension m, et cette dimension m ne dépend aucunement de la paramétrisation choisie (Brown, 1986b, p. 13-16). (Voir l'Exercice 3.23 pour l'exemple d'une famille exponentielle non régulière.)

Les familles exponentielles naturelles peuvent aussi être réécrites sous la forme

$$f(x|\theta) = h(x)\, e^{\theta.x - \psi(\theta)} \tag{3.7}$$

et $\psi(\theta)$ est dite *fonction cumulante des moments* pour la raison suivante, dont la démonstration est laissée aux lecteurs.

Lemme 3.13. *Si* $\theta \in \mathring{N}$, *intérieur de* N, *la fonction cumulante des moments* ψ *est* $\mathscr{C}^\infty$ *et*

$$\mathbb{E}_\theta[x] = \nabla\psi(\theta), \qquad \operatorname{cov}(x_i, x_j) = \frac{\partial^2\psi}{\partial\theta_i\partial\theta_j}(\theta),$$

où ∇ *désigne l'opérateur gradient.*

Exemple 3.14. Soit $x \sim \mathscr{P}(\lambda)$. Alors

$$f(x|\lambda) = e^{-\lambda}\,\frac{\lambda^x}{x!} = \frac{1}{x!}\, e^{\theta.x - e^\theta}$$

et $\psi(\theta) = \exp(\theta)$ pour le paramètre naturel $\theta = \log\lambda$. Par conséquent, $\mathbb{E}_\lambda[x] = e^\theta = \lambda$ et $\operatorname{var}(x) = \lambda$. ‖

La structure régulière des familles exponentielles permet de nombreuses applications statistiques, comme en témoigne la vaste littérature sur ce sujet. (Voir, par exemple, la classification des familles exponentielles selon le type de fonction de variance : Morris, 1982, Letac et Mora, 1990, et Exercices 3.24 et 10.33.) Nous verrons dans la Section 3.3.4 qu'elles autorisent également une construction simple des lois a priori conjuguées.

Exemple 3.15. Si $x \sim \mathscr{N}(\theta, \theta^2)$ dans un modèle multiplicatif, la loi a priori conjuguée n'est pas la loi normale. La vraisemblance est proportionnelle à

$$\frac{1}{|\,\theta\,|}\exp\left\{\frac{x}{\theta}-\frac{x^2}{2\theta^2}\right\}$$

et la distribution induit une famille exponentielle de dimension 2. Par conséquent, les *lois normales inverses généralisées* $\mathscr{IN}(\alpha,\mu,\tau)$, de densité

$$\pi(\theta)\propto|\theta|^{-\alpha}\exp\left\{-\left(\frac{1}{\theta}-\mu\right)^2\Big/2\tau^2\right\}$$

constituent pour ce modèle une famille conjuguée. Cette famille de lois, qui forme une famille exponentielle, généralise la loi de l'inverse d'une observation normale (qui correspond au cas $\alpha=2$). (Voir l'Exercice 3.33 pour plus de détails.) ‖

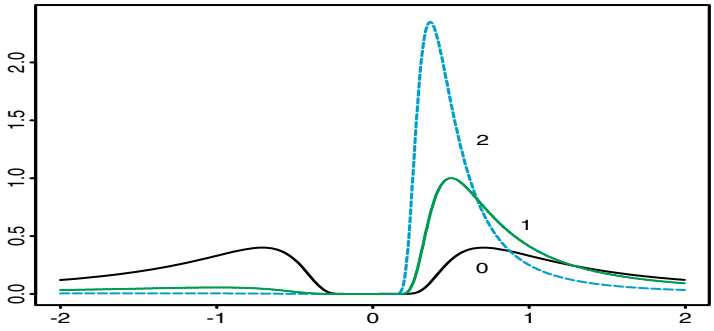

Fig. 3.2. Densités $\mathscr{IN}(\alpha,\mu,\tau)$ pour $\alpha=2$, $\tau=1$ et $\mu=0,1,2$.

Évidemment, la plupart des lois n'appartiennent pas à une famille exponentielle ! Par exemple, la loi de Student, $\mathscr{T}_p(\nu,\theta,\sigma^2)$, ne peut pas s'exprimer sous la forme (3.5). La Définition 3.8 exclut aussi toutes les lois avec un support non constant, alors que certaines d'entre elles admettent des lois a priori conjuguées avec un nombre fini de paramètres (ou plus exactement, d'*hyperparamètres*).

Exemple 3.16. Les lois de Pareto, $\mathscr{P}(\alpha,\theta)$, de densité

$$f(x|\alpha,\theta)=\alpha\frac{\theta^\alpha}{x^{\alpha+1}}\mathbb{I}_{[\theta,+\infty[}(x)\qquad(\theta>0),$$

sont de telles lois puisque, bien qu'en dehors du cadre des familles exponentielles, elles admettent des lois conjuguées simples sur θ, qui sont des lois de Pareto pour $1/\theta$. ‖

D'autres exemples de familles pour lesquelles des lois conjuguées sont disponibles sont les distributions $\mathcal{U}_{[-\theta,\theta]}$ et $\mathcal{U}_{[0,\theta]}$; ces lois sont aussi quasi exponentielles, car elles admettent des statistiques exhaustives de dimension constante. Par exemple, si $x_1, \ldots, x_n \sim \mathcal{U}_{[-\theta,\theta]}$, une statistique exhaustive est la *statistique d'ordre* $(x_{(1)}, x_{(n)})$, où $x_{(1)}$ est la valeur la plus petite de l'échantillon et $x_{(n)}$ la plus grande.

Notons que, dans l'Exemple 3.15, la loi a priori conjuguée sur θ dépend de trois hyperparamètres, α, μ, et τ^2 ; par conséquent, leur utilisation introduit une plus grande complexité dans la loi du modèle. Ce type de phénomène, c'est-à-dire le fait que la structure du modèle exige un nombre plus grand d'hyperparamètres, est souvent rencontré dans les *familles exponentielles courbes*, par exemple quand une reparamétrisation naturelle par $\eta = R(\theta)$ n'est pas utile à cause des contraintes portant sur les paramètres naturels. Il s'agit évidemment d'un inconvénient, car les valeurs de ces hyperparamètres doivent être déterminées pour obtenir une inférence sur θ utilisant des lois conjuguées.

Quand une distribution n'admet pas de famille conjuguée, sauf le cas trivial $\mathcal{F}_0$, il est parfois possible d'exprimer cette distribution comme un *mélange* de distributions de familles exponentielles ; f est appelée *mélange caché*, car cette représentation est sans pertinence pour le problème inférentiel, mais est utile pour le calcul pratique de la loi a posteriori et des estimateurs de Bayes, comme nous le verrons dans le Chapitre 6.

Exemple 3.17. (Dickey, 1968) Pour la loi de Student, une représentation de mélange caché existe, fondée sur la distribution normale, car $f(x|\theta)$ est le mélange d'une distribution normale par l'inverse d'une distribution gamma : si $x \sim \mathcal{T}_1(p, \theta, \sigma^2)$,

$$x|z \sim \mathcal{N}(\theta, z\sigma^2), \qquad z^{-1} \sim \mathcal{G}(p/2, p/2)\,.$$

Une loi a priori techniquement intéressante sur θ est alors $\mathcal{N}(\mu, \tau^2)$ et la plupart des calculs peuvent être faits conditionnellement à z. Cette décomposition est plus utile quand x est multidimensionnel, car certaines intégrales deviennent alors unidimensionnelles. ‖

Exemple 3.18. Plusieurs lois non centrées peuvent s'écrire comme un mélange (caché) des lois centrées correspondantes par la loi de Poisson, de par une propriété d'*infinie divisibilité* (voir Feller, 1971, Chapitre 9). Par exemple, tel est le cas de la loi du khi deux décentré, $\chi^2_p(\lambda)$: Lorsque $x \sim \chi^2_p(\lambda)$, la génération de x peut être aussi décomposée comme

$$x|z \sim \chi^2_{p+2z}, \qquad z \sim \mathcal{P}(\lambda/2).$$

Cette décomposition est utilisée par James et Stein (1961) pour exprimer le risque de leur estimateur et obtenir une condition suffisante de domination de l'estimateur de maximum de vraisemblance (voir la Note 2.8.2). ‖

Cette extension du champ d'application des lois conjuguées est cependant discutable, car la représentation par mélange caché n'est pas unique et le choix du mélange détermine celui de la loi a priori.

3.3.4 Lois conjuguées des familles exponentielles

Soit $f(x|\theta) = h(x)e^{\theta.x-\psi(\theta)}$, loi générique d'une famille exponentielle. Cette loi admet alors une famille conjuguée, comme le démontre le résultat suivant (dont la démonstration est directe).

Proposition 3.19. *Une famille conjuguée pour $f(x|\theta)$ est donnée par*

$$\pi(\theta|\mu,\lambda) = K(\mu,\lambda)\, e^{\theta.\mu-\lambda\psi(\theta)}, \tag{3.8}$$

où $K(\mu,\lambda)$ est la constante de normalisation de la densité. La loi a posteriori correspondante est $\pi(\theta|\mu+x,\lambda+1)$.

La mesure définie par (3.8) est σ-finie ; elle génère une loi de probabilité sur Θ si et seulement si

$$\lambda > 0 \qquad \text{et} \qquad \frac{\mu}{\lambda} \in \overset{\circ}{N} \tag{3.9}$$

(Exercice 3.35) : c'est uniquement quand (3.9) est vérifié que $K(\mu,\lambda)$ est bien défini. Par conséquent, une loi conjuguée pour $f(x|\theta)$ peut être obtenue de façon automatique ; c'est pourquoi (3.8) est souvent appelée *loi conjuguée naturelle* de f. Le Tableau 3.4 présente les lois conjuguées pour certaines lois usuelles appartenant à une famille exponentielle[24]. Évidemment, l'inférence bayésienne ne peut être menée que si les hyperparamètres μ et λ sont connus. L'aspect automatique des lois conjuguées a priori est ainsi trompeur, car un apport subjectif via la détermination de ces valeurs demeure nécessaire. Notons aussi que (3.8) requiert un paramètre additionnel, relativement à $f(x|\theta)$.

Pour des familles exponentielles naturelles, les lois a priori conjuguées ont un attrait supplémentaire, comme le montrent Diaconis et Ylvisaker (1979) : si $\xi(\theta)$ est l'espérance de $x \sim f(x|\theta)$, l'espérance a posteriori de $\xi(\theta)$ est linéaire en x pour une loi a priori conjuguée.

Proposition 3.20. *Si Θ est un ensemble ouvert dans $\mathbb{R}^k$ et θ a pour loi a priori*

$$\pi_{\lambda,x_0}(\theta) \propto e^{\theta\cdot x_0-\lambda\psi(\theta)}$$

avec $x_0 \in \mathcal{X}$, alors

$$\mathbb{E}^{\pi}[\xi(\theta)] = \mathbb{E}^{\pi}[\nabla\psi(\theta)] = \frac{x_0}{\lambda}.$$

[24] Puisque les lois conjuguées viennent aussi d'une famille exponentielle, Bar-Lev *et al.* (1994) ont étudié le problème réciproque, à savoir la détermination des distributions $\pi(\theta)$ pour lesquelles une famille exponentielle admet $\pi(\theta)$ comme loi conjuguée.

Tab. 3.4. Lois a priori conjuguées naturelles pour quelques familles exponentielles usuelles.

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Normale $\mathcal{N}(\theta,\sigma^2)$	Normale $\mathcal{N}(\mu,\tau^2)$	$\mathcal{N}(\varrho(\sigma^2\mu+\tau^2x),\varrho\sigma^2\tau^2)$ $\varrho^{-1}=\sigma^2+\tau^2$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha,\beta)$	$\mathcal{G}(\alpha+x,\beta+1)$
Gamma $\mathcal{G}(\nu,\theta)$	Gamma $\mathcal{G}(\alpha,\beta)$	$\mathcal{G}(\alpha+\nu,\beta+x)$
Binomiale $\mathcal{B}(n,\theta)$	Bêta $\mathcal{B}e(\alpha,\beta)$	$\mathcal{B}e(\alpha+x,\beta+n-x)$
Binomiale Négative $\mathcal{N}eg(m,\theta)$	Bêta $\mathcal{B}e(\alpha,\beta)$	$\mathcal{B}e(\alpha+m,\beta+x)$
Multinomiale $\mathcal{M}_k(\theta_1,\ldots,\theta_k)$	Dirichlet $\mathcal{D}(\alpha_1,\ldots,\alpha_k)$	$\mathcal{D}(\alpha_1+x_1,\ldots,\alpha_k+x_k)$
Normale $\mathcal{N}(\mu,1/\theta)$	Gamma $\mathcal{G}a(\alpha,\beta)$	$\mathcal{G}(\alpha+0.5,\beta+(\mu-x)^2/2)$

Par conséquent, si $x_1,\ldots,x_n$ *sont i.i.d.* $f(x|\theta)$,

$$\mathbb{E}^\pi[\xi(\theta)|x_1,\ldots,x_n]=\frac{x_0+n\bar{x}}{\lambda+n}. \tag{3.10}$$

Ce résultat est bien connu pour les distributions normales (Exemple 3.2) et peut ainsi être généralisé à toutes les familles exponentielles. L'équation (3.10) montre de nouveau que le paramètre λ est comparable à la taille de l'échantillon n. Par conséquent, sa détermination peut être réalisée, si nécessaire, en considérant que l'information a priori sur x_0 provient d'un échantillon virtuel de taille λ. Brown (1986b) établit que la Proposition 3.20 peut s'étendre au cas où π_{λ,x_0} est impropre, par exemple quand $\lambda=0$ et $x_0=0$. Dans ce cas, l'espérance a posteriori est $\bar{x}$, qui est aussi l'estimateur du maximum de vraisemblance de $\xi(\theta)$.

Diaconis et Ylvisaker (1979) ont montré, de surcroît, une réciproque de cette proposition, à savoir que, si la mesure de référence est continue par rapport à la mesure de Lebesgue, la linéarité de $\mathbb{E}^\pi[\xi(\theta)|x]$ comme dans (3.10) entraîne que la loi a priori est de la forme (3.6). Les extensions aux cas discrets sont plus délicates.

Bien que les familles exponentielles permettent généralement un traitement plus aisé et, particulièrement, l'utilisation commode de lois a priori conjuguées et le calcul analytique des espérances a posteriori, comme dans la Proposition 3.20, ce n'est pas toujours le cas. Par exemple, quand $x\sim\mathcal{B}e(\alpha,\theta)$ avec α connu, la distribution appartient à une famille exponentielle, car

$$f(x|\theta)\propto\frac{\Gamma(\alpha+\theta)(1-x)^\theta}{\Gamma(\theta)},$$

mais les lois conjuguées ne sont pas faciles à utiliser, car

$$\pi(\theta|x_0,\lambda) \propto \left(\frac{\Gamma(\alpha+\theta)}{\Gamma(\theta)}\right)^{\lambda} (1-x_0)^{\theta}$$

dépend de la fonction gamma $\Gamma(\theta)$, qui n'a pas d'expression explicite.

Exemple 3.21. La *régression logistique* est utilisée pour décrire des modèles qualitatifs comme dans l'Exemple 1.1. Soit une variable indicatrice y, prenant ses valeurs dans $\{0,1\}$, et des variables explicatives $x \in \mathbb{R}^k$, telles que la distribution de y conditionnelle à x soit

$$P_\alpha(y=1) = 1 - P_\alpha(y=0) = \frac{\exp(\alpha^t x)}{1+\exp(\alpha^t x)}\,. \tag{3.11}$$

Ce modèle permet l'extension du très utile modèle de régression linéaire à des cadres plus qualitatifs. Pour un échantillon $(y_1,x_1),\ldots,(y_n,x_n)$ de (3.11), le modèle est bien sûr exponentiel *conditionnellement aux* x_i, puisque

$$f(y_1,\ldots,y_n|x_1,\ldots,x_n,\alpha) = \exp\left(\alpha^t \sum_{i=1}^n y_i x_i\right) \prod_{i=1}^n (1+e^{\alpha^t x_i})^{-1},$$

qui dépend uniquement de la statistique exhaustive $\sum_{i=1}^n y_i x_i$. Dans la pratique, les lois a priori conjuguées pour ce modèle sont plutôt difficiles à utiliser, car elles sont de la forme

$$\pi(\alpha|y_0,\lambda) \propto e^{\alpha^t y_0} \prod_{i=1}^n (1+e^{\alpha^t x_i})^{-\lambda}.$$

La constante de normalisation pour $\pi(\alpha|y_0,\lambda)$ est inconnue et les approximations des quantités a posteriori comme l'espérance et la médiane a posteriori ne peuvent être obtenues qu'à travers des techniques de simulation présentées dans le Chapitre 6. ‖

3.4 Critiques et extensions

Comme nous l'avons déjà vu ci-dessus, le caractère automatique des lois conjuguées est à la fois un avantage et un inconvénient. En sus des arguments d'invariance et de linéarité, on peut argumenter qu'il s'agit d'une approche objective, où l'apport subjectif est réduit à la détermination des hyperparamètres. Mis à part le fait que l'objectivité est un concept difficile à définir, on peut répliquer que toute autre loi a priori avec le même nombre d'hyperparamètres pourrait paraître tout aussi objective. De plus, les

lois a priori conjuguées ne sont pas forcément les lois a priori les plus robustes (voir la Section 3.5) et, de ce point de vue, d'autres lois peuvent être préférées, si l'impératif est de minimiser l'influence de l'a priori sur le résultat de l'inférence. L'exemple suivant montre comment le choix d'une loi a priori peut modifier la distribution a posteriori pour des échantillons de petite taille.

Exemple 3.22. (Diaconis et Ylvisaker, 1985) Lorsqu'on fait tourner une pièce sur la tranche, plutôt que de la lancer dans l'air, la proportion de *piles* est rarement proche de 1/2, mais se stabilise plutôt autour de 1/3 ou 2/3, du fait d'irrégularités de fabrication qui biaisent le résultat en faveur d'un côté ou de l'autre. On observe le nombre de *piles*, $x \sim \mathcal{B}(n,p)$ pour une pièce donnée qu'on fait tourner n fois sur sa tranche. La loi a priori sur p semble être bimodale, ce que ne peut refléter une loi a priori conjuguée π_1 comme $\mathcal{B}e(1,1)$. Un mélange de lois a priori π_2 tel que

$$\frac{1}{2}\left[\mathcal{B}e(10,20) + \mathcal{B}e(20,10)\right]$$

est donc plus approprié. Il peut arriver aussi que des expériences précédentes avec la même pièce indiquent un biais vers *pile* et mènent à l'a priori alternatif suivant, π_3 :

$$0.5\,\mathcal{B}e(10,20) + 0.2\,\mathcal{B}e(15,15) + 0.3\,\mathcal{B}e(20,10).$$

La Figure 3.3 fournit le graphe des deux densités a priori ci-dessus et de l'a priori neutre $\mathcal{B}e(1,1)$, les différences entre les trois modèles a priori étant effectivement assez importantes. Si, pour $n = 10$, nous observons $x = 3$, les lois a posteriori correspondantes sont :

(i) $\mathcal{B}e(1+x, 1+n-x)$, soit $\mathcal{B}e(4,8)$;

(ii) $0.84\,\mathcal{B}e(13,27) + 0.16\,\mathcal{B}e(23,17)$; et

(iii) $0.77\,\mathcal{B}e(13,27) + 0.16\,\mathcal{B}e(18,22) + 0.07\,\mathcal{B}e(23,17)$.

En (ii), les pondérations de probabilités a posteriori sont obtenues comme étant proportionnelles à

$$\frac{1}{2}\frac{B(13,27)}{B(10,20)} \qquad \text{et} \qquad \frac{1}{2}\frac{B(23,17)}{B(20,10)}$$

et, pour (iii),

$$0.5\,\frac{B(13,27)}{B(10,20)}, \qquad 0.2\,\frac{B(18,22)}{B(15,15)}, \qquad \text{et} \qquad 0.3\,\frac{B(23,17)}{B(20,10)},$$

où

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

est l'inverse du terme de normalisation de la densité bêta (définie dans l'Appendice A), qui peut être approchée numériquement (ou calculée exactement dans le cas de coefficients entiers).

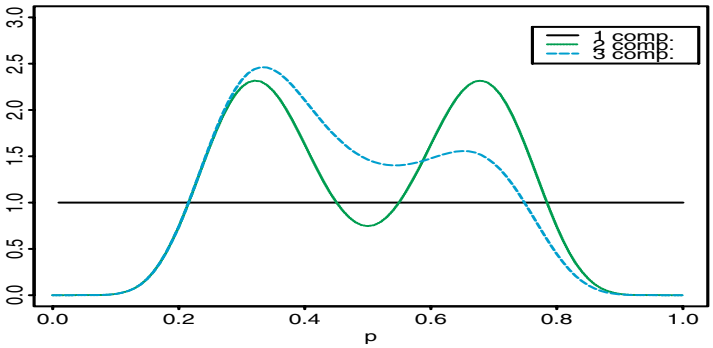

Fig. 3.3. Trois lois a priori pour une expérience de pile ou face.

Par conséquent, pour cet échantillon, les trois moyennes a posteriori, 1/3, 0.365 et 0.362 respectivement, sont assez proches mais les formes des lois a posteriori différent malgré tout (voir la Figure 3.4). Considérons maintenant un échantillon de taille $n = 50$ avec $x = 36$. Les lois a posteriori sont :

(i) $\mathcal{B}e(15, 37)$;

(ii) $0.997\,\mathcal{B}e(24, 56) + 0.003\,\mathcal{B}e(34, 46)$; et

(iii) $0.95\,\mathcal{B}e(24, 56) + 0.047\,\mathcal{B}e(29, 51) + 0.003\,\mathcal{B}e(34, 46)$.

Elles sont alors plus proches les unes des autres que pour $n = 10$, comme le montre la Figure 3.5. ‖

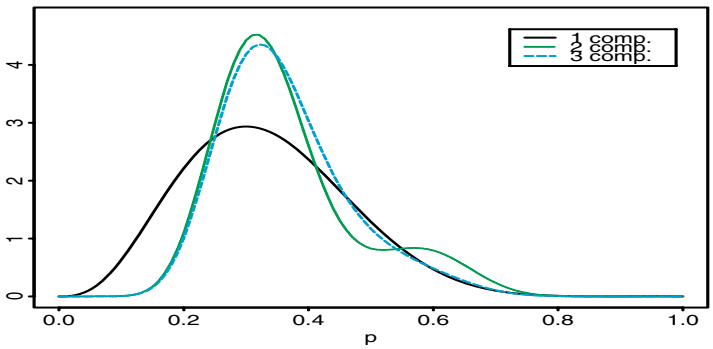

Fig. 3.4. Lois a posteriori pour un modèle de pile ou face pour dix observations.

Deux remarques découlent logiquement de cet exemple. D'abord, il montre qu'un modèle a priori est certainement important pour de petits échantillons, mais aussi qu'il l'est de moins en moins à mesure que la taille de l'échantillon

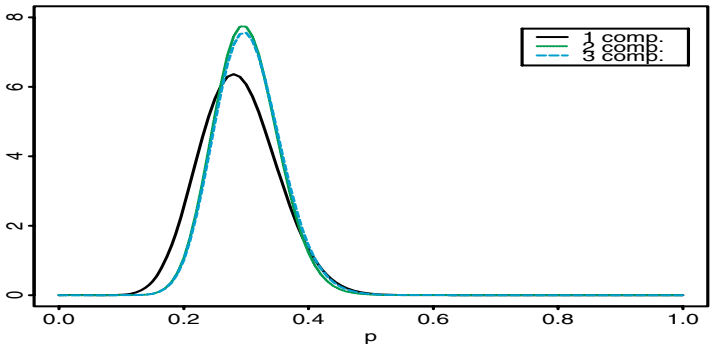

Fig. 3.5. Lois a posteriori pour cinquante observations.

augmente. Quand la taille de l'échantillon tend vers l'infini, la plupart des lois a priori mèneront à la même inférence, qui sera équivalente à celle fondée seulement sur la fonction de vraisemblance, comme remarqué dans la Note 1.8.4. De plus, cet exemple montre que les *mélanges* de lois a priori conjuguées sont aussi faciles à manipuler que les lois a priori habituelles, tout en permettant une plus grande liberté dans la modélisation de l'information a priori. En effet, les mélanges de lois conjuguées forment aussi des familles conjuguées.

Lemme 3.23. *Soit $\mathscr{F}$ la famille conjuguée naturelle d'une famille exponentielle* (3.6)*. Alors l'ensemble des mélanges de N lois conjuguées,*

$$\tilde{\mathscr{F}}_N = \left\{ \sum_{i=1}^{N} \omega_i \pi(\theta|\lambda_i, \mu_i); \sum_{i=1}^{N} \omega_i = 1,\ \omega_i > 0 \right\},$$

est aussi une famille conjuguée. De plus, si

$$\pi(\theta) = \sum_{i=1}^{N} \omega_i \pi(\theta|\lambda_i, \mu_i),$$

la loi a posteriori est un mélange

$$\pi(\theta|x) = \sum_{i=1}^{N} \omega_i'(x) \pi(\theta|\lambda_i + 1, \mu_i + x),$$

avec

$$\omega_i'(x) = \frac{\omega_i K(\mu_i, \lambda_i)/K(\mu_i + x, \lambda_i + 1)}{\sum_{j=1}^{N} \omega_j K(\mu_j, \lambda_j)/K(\mu_j + x, \lambda_j + 1)}.$$

Les mélanges peuvent alors être utilisés comme base pour approcher une loi a priori quelconque, au sens où la distance de Prohorov entre une loi

et sa représentation par un mélange peut être rendue arbitrairement petite. Rappelons que la *distance de Prohorov* entre deux mesures π et $\tilde{\pi}$, $d^P(\pi, \tilde{\pi})$ est définie comme

$$d^P(\pi, \tilde{\pi}) = \inf_A \{\epsilon\,;\, \pi(A) \leq \tilde{\pi}(A^\epsilon) + \epsilon\}\,,$$

où l'infimum est pris sur les ensembles boréliens et où A^ϵ indique l'ensemble des points distants de A d'au plus ϵ (Le Cam, 1986).

Théorème 3.24. *Si Θ est l'espace naturel des paramètres pour la famille exponentielle $f(x|\theta)$ et π est une loi a priori sur Θ, alors, pour tout $\epsilon > 0$, on peut trouver N et $\tilde{\pi} \in \tilde{\mathscr{F}}_N$ tels que $d^P(\pi, \tilde{\pi}) < \epsilon$.*

La démonstration de ce théorème peut être reliée au fait que les mélanges finis de mesures de Dirac sont denses dans la topologie de Prohorov et que les masses de Dirac peuvent s'approcher par des mélanges de lois a priori conjuguées. (Pour plus de détails, voir Brown, 1986b, p. 254-267.) Ce résultat justifie beaucoup plus fortement l'utilisation de lois conjuguées que l'invariance, la linéarité ou les arguments de simplicité de la section précédente. Quelle que soit l'information a priori disponible, celle-ci peut toujours être modélisée par un mélange de $\tilde{\mathscr{F}}_N$ avec N aussi petit que possible. Cependant, ce résultat d'approximation est aussi incomplet, car il ne montre pas comment l'approximation s'étend aux quantités a posteriori, alors que l'inférence bayésienne ne s'intéresse qu'à celles-ci. Berger (1985b) illustre cette différence à travers l'exemple suivant.

Exemple 3.25. Soit $x \sim \mathscr{N}(\theta, 1)$ et prenons pour a priori associé π_0 une loi de Cauchy, $\mathscr{C}(0, 1)$. Les lois conjuguées naturelles étant $\mathscr{N}(\mu, A)$, π_0 peut s'approcher par

$$\tilde{\pi} = \sum_{i=1}^{N} \lambda_i \pi_i,$$

où π_i est $\mathscr{N}(\mu_i, A_i)$, selon le Théorème 3.24. Lorsque x tend vers $+\infty$, $\pi_0(\theta|x)$ tend vers $\mathscr{N}(x, 1)$ tandis que $\tilde{\pi}(\theta|x)$ est approximativement $\mathscr{N}(\mu(x), \varrho)$, avec

$$\varrho = \frac{A^*}{1 + A^*}, \quad \mu(x) = \varrho x + (1 - \varrho)\mu^*, \quad A^* = \max_i\{A_i\}, \quad \mu^* = \max_{A_i = A^*} \mu_i.$$

Par conséquent, $\pi_0(\theta|x)$ et $\tilde{\pi}(\theta|x)$ vont nettement différer pour de grandes valeurs de x. On peut remarquer que ces valeurs ne sont pas compatibles avec l'information a priori et devraient conduire à une modification de la modélisation a priori. Mais ces différences démontrent malgré tout que l'approximation a priori n'est pas uniformément valide a posteriori. ‖

L'Exemple 3.25 illustre avec force le point suivant : les lois à queues lourdes seront mal approchées par des distributions à queue moins lourde. Cette difficulté et, plus généralement, le problème d'approximation de lois a posteriori disparaissent d'une certaine façon dans la généralisation de Dalal et Hall

(1983), qui considèrent des mélanges continus (dans un cas continu). Nous décrirons brièvement leur approche dans la Note 3.8.3, mais nous tenons à remarquer que leur approximation via des mélanges continus n'a pas l'attrait des approximations précédentes, car elle requiert souvent une résolution numérique ou de Monte Carlo.

3.5 Lois a priori non informatives

La section précédente a montré que les lois conjuguées peuvent être utiles en tant qu'approximations des véritables lois a priori. En revanche, lorsque aucune information a priori n'est disponible, leur unique justification est analytique, puisqu'elles donnent des expressions exactes pour quelques quantités a posteriori. Dans de telles situations, il est impossible de justifier le choix d'une loi a priori sur des bases subjectives et les hyperparamètres des lois conjuguées ne peuvent être déterminés qu'arbitrairement. Plutôt que de revenir aux alternatives classiques, comme l'estimation par maximum de vraisemblance, ou d'utiliser les données pour approcher ces hyperparamètres, comme dans une analyse bayésienne empirique, il est préférable de faire appel à des techniques bayésiennes, ne serait-ce que parce qu'elles sont à la base des critères classiques d'optimalité (voir les Chapitres 2, 8 et 9). Dans un tel cas, ces lois a priori particulières doivent être construites à partir de la distribution d'échantillonnage, puisque c'est la seule information disponible. Pour des raisons évidentes, de telles lois sont dites *non informatives*. Nous décrivons plus loin quelques-unes des techniques les plus importantes de construction de lois non informatives, en demandant aux lecteurs de se référer à Kass et Wasserman (1996) pour un traitement plus approfondi de ces notions et une bibliographie commentée. Le point principal mérite d'être reproduit ici, avant que nous entamions cette description : on ne peut attendre des lois non informatives qu'elles représentent exactement une ignorance *totale* sur le problème considéré. Celles-ci doivent plutôt être comprises comme des lois de référence ou des lois choisies par défaut, auxquelles chacun pourrait avoir recours quand toute information a priori est absente. À cet égard, certaines lois non informatives sont plus utiles ou plus efficaces que d'autres, mais ne peuvent être pour autant perçues comme moins informatives que d'autres.

3.5.1 Les lois a priori de Laplace

Historiquement, Laplace fut le premier à utiliser des techniques non informatives puisque, bien que ne disposant pas d'information sur le nombre de boules blanches dans l'urne ou sur la proportion de naissances mâles (Exemples 1.9 et 1.11), il munit ces paramètres d'une loi a priori qui prend en compte son ignorance en donnant la même vraisemblance à chaque valeur du paramètre, soit donc en utilisant une *loi uniforme*. Son raisonnement, appelé

plus tard *principe de la raison insuffisante*, se fondait sur l'*équiprobabilité* des événements élémentaires.

Trois critiques ont été plus tard avancées sur ce choix. Premièrement, les lois résultantes sont impropres quand l'espace des paramètres n'est pas compact et certains statisticiens se refusent à utiliser de telles lois, car elles mènent à des difficultés comme le *paradoxe de marginalisation* (voir les Exercices 3.45-3.51). De telles inquiétudes ne sont pas justifiées, puisqu'en réalité il est possible de travailler avec des lois impropres, comme nous l'avons vu dans la Section 1.5, du moment que nous n'essayons pas de les interpréter comme des lois de probabilité (voir aussi Stone, 1976). Comme cela est mentionné dans la Section 3.2, il peut être avancé que, au contraire, une détermination subjective d'une loi a priori *devrait* conduire à une loi impropre.

Deuxièmement, le principe des événements équiprobables de Laplace n'est pas cohérent en termes de partitionnement : si $\Theta = \{\theta_1, \theta_2\}$, la règle de Laplace donne $\pi(\theta_1) = \pi(\theta_2) = 1/2$ mais, si la définition de Θ est plus détaillée, avec $\Theta = \{\theta_1, \omega_1, \omega_2\}$, la règle de Laplace mène à $\pi(\theta_1) = 1/3$, ce qui évidemment n'est pas cohérent avec la première formulation. Comme cela est discuté dans Kass et Wasserman (1996), cette cohérence n'est pas un problème important : il peut être évacué en argumentant que le niveau de partitionnement doit être fixé à un certain stade de l'analyse et que l'introduction d'un degré plus fin dans le partitionnement modifie le problème d'inférence.

La troisième critique est plus fondamentale, car elle concerne le problème de l'*invariance par reparamétrisation*. Si on passe de $\theta \in \Theta$ à $\eta = g(\theta)$ par une transformation bijective g, l'information a priori reste totalement inexistante et ne devrait pas être modifiée. Cependant, si $\pi(\theta) = 1$, la loi a priori sur η est

$$\pi^*(\eta) = \left| \frac{d}{d\eta} g^{-1}(\eta) \right|$$

par la formule du changement de variable. Donc $\pi(\eta)$ est le plus souvent non constante.

Exemple 3.26. Si p, la proportion de naissances mâles, suit une loi uniforme sur [0,1], le paramètre de rapport des chances $\varrho = \frac{p}{1-p}$ suit une loi a priori de densité $1/(1+\varrho)^2$, qui est donc non constante. ‖

Bien entendu, on peut parfois soutenir qu'il existe un paramètre naturel d'intérêt et par conséquent que le choix d'une loi uniforme pour ce paramètre d'intérêt n'a pas besoin d'être invariant par reparamétrisation. Mais cet argument ne tient pas si plus d'une inférence sur θ doit être menée ; par exemple, nous pourrions avoir besoin de calculer les deux premiers moments de θ, mais ce dernier est aussi l'espérance de θ^2. Ou, dans l'Exemple 3.26, la probabilité θ *et* le rapport des risques ϱ peuvent être d'intérêt. Par conséquent, il semble qu'une notion plus intrinsèque et plus acceptable de la loi non informative devrait satisfaire l'*invariance par reparamétrisation.*

3.5.2 Lois invariantes

Une première solution est de tirer profit des caractéristiques d'invariance du problème, c'est-à-dire d'utiliser les groupes $\mathscr{G}$ agissant sur $\mathscr{X}$ qui induisent des groupes $\mathscr{G}^*$ agissant sur Θ (au sens où seuls les paramètres de la distribution de x changent dans une transformation de x par des éléments de $\mathscr{G}$). Le Chapitre 9 détaille les liens entre structures d'invariance et approche bayésienne, ces structures permettant d'obtenir une certaine loi non informative compatible avec les exigences d'invariance, à savoir, la mesure de Haar à droite sur $\mathscr{G}^*$; voir Kass et Wasserman (1996) pour plusieurs arguments en faveur de la mesure de Haar à droite.

Deux exemples introductifs sont présentés ci-dessous.

Exemple 3.27. La famille de lois $f(x-\theta)$ est *invariante par translation*, car $y = x - x_0$ a une loi de la même famille pour tout x_0, $f(y-(\theta-x_0))$; θ est alors dit *paramètre de position* et une exigence d'invariance est que la loi a priori soit invariante par translation, donc satisfasse

$$\pi(\theta) = \pi(\theta - \theta_0)$$

pour tout θ_0. La solution est $\pi(\theta) = c$, la loi uniforme sur Θ. ‖

Exemple 3.28. Si la famille de lois est paramétrée par un *paramètre d'échelle*, c'est-à-dire est de la forme $1/\sigma f(x/\sigma)$ $(\sigma > 0)$, elle est *invariante par changement d'échelle*, $y = x/\sigma \sim f(y)$. La loi a priori invariante par changement d'échelle π satisfait $\pi(A) = \pi(A/c)$ pour tout ensemble mesurable A dans $(0, +\infty)$ et $c > 0$, soit

$$\pi(\sigma) = \frac{1}{c}\pi(\frac{\sigma}{c}).$$

Ceci implique $\pi(\sigma) = \alpha/\sigma$, où α est une constante. Donc la mesure invariante n'est plus constante. ‖

L'approche invariante n'est que partiellement satisfaisante, car elle implique la référence à une structure d'invariance, qui peut être parfois choisie de plusieurs manières, ne pas exister (voir le Chapitre 9), ou être sans intérêt pour le décideur.

3.5.3 La loi a priori de Jeffreys

Jeffreys (1946, 1961) propose une approche intrinsèque qui évite effectivement le besoin de prendre en compte une structure d'invariance potentielle, tout en étant souvent compatible lorsque cette structure existe. Les *lois a priori non informatives de Jeffreys* sont fondées sur l'*information de Fisher*, donnée par

$$I(\theta) = \mathbb{E}_\theta \left[\left(\frac{\partial \log f(X \mid \theta)}{\partial \theta} \right)^2 \right]$$

dans le cas unidimensionnel. Sous certaines conditions de régularité, cette information est aussi égale à

$$I(\theta) = -\mathbb{E}_\theta \left[\frac{\partial^2 \log f(X \mid \theta)}{\partial \theta^2} \right]. \tag{3.12}$$

La loi a priori de Jeffreys est

$$\pi^*(\theta) \propto I^{1/2}(\theta),$$

définie à un coefficient de normalisation près quand π^* est propre. Elle vérifie effectivement l'exigence d'invariance par reparamétrisation, puisque, pour une transformation bijective h donnée, nous avons la transformation (jacobienne)

$$I(\theta) = I(h(\theta))(h'(\theta))^2$$

(qui explique l'exposant 1/2). De plus, elle correspond aux lois invariantes obtenues dans les Exemples 3.27 et 3.28. Plus fondamentalement, le choix d'une loi a priori dépendant de l'information de Fisher se justifie par le fait que $I(\theta)$ est largement accepté comme un indicateur de la quantité d'information apportée par le modèle (ou l'observation) sur θ (Fisher, 1956). Par conséquent, au moins à un niveau qualitatif, il paraît intuitivement justifié que les valeurs de θ pour lesquelles $I(\theta)$ est plus grande doivent être plus probables a priori. En d'autres termes, $I(\theta)$ mesure la capacité du modèle à discriminer entre θ et $\theta \pm d\theta$ via la pente moyenne de $\log f(x|\theta)$. Favoriser les valeurs de θ pour lesquelles $I(\theta)$ est plus grande équivaut à minimiser l'influence de la loi a priori et est donc aussi non informatif que possible. En fait, la loi de Jeffreys est fréquemment impropre mais les développements de la Section 1.5 montrent comment conduire une analyse bayésienne dans ce cas.

Exemple 3.29. (Suite de l'Exemple 3.26) Si $x \sim B(n, p)$,

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x},$$

$$\frac{\partial^2 \log f(x|p)}{\partial p^2} = \frac{x}{p^2} + \frac{n-x}{(1-p)^2},$$

et

$$I(p) = n \left[\frac{1}{p} + \frac{1}{1-p} \right] = \frac{n}{p(1-p)}.$$

Donc la loi de Jeffreys pour ce modèle est

$$\pi^*(p) \propto [p(1-p)]^{-1/2}$$

et est alors propre, car il s'agit de la distribution $\mathscr{B}e(1/2, 1/2)$. ‖

Dans le cas où θ est un paramètre multidimensionnel, on définit la matrice d'information de Fisher par généralisation de (3.12). Pour $\theta \in \mathbb{R}^k$, $I(\theta)$ a les éléments suivants :

$$I_{ij}(\theta) = -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f(x|\theta) \right] \qquad (i, j = 1, \ldots, k),$$

et la loi non informative de Jeffreys est alors définie par

$$\pi^*(\theta) \propto [\det(I(\theta))]^{1/2}.$$

Elle est encore invariante par reparamétrisation. Notons que, si $f(x|\theta)$ appartient à une famille exponentielle,

$$f(x|\theta) = h(x) \exp(\theta \cdot x - \psi(\theta)),$$

la matrice d'information de Fisher est donnée par $I(\theta) = \nabla\nabla^t\psi(\theta)$ et

$$\pi^*(\theta) \propto \left(\prod_{i=1}^{k} \psi''_{ii}(\theta) \right)^{1/2}, \tag{3.13}$$

où $\psi''_{ii}(\theta) = \frac{\partial^2}{\partial\theta_i^2}\psi(\theta)$.

Dans un cas multidimensionnel, l'approche non informative de Jeffreys peut conduire à des incohérences ou même à des paradoxes (voir les Exemples 3.31 et 3.34) et nous notons que Jeffreys (1961) a surtout insisté sur l'utilisation de cette loi dans des cas unidimensionnels (voir Berger et Bernardo, 1992a). Cependant, sa méthode fournit une des meilleures techniques automatiques pour obtenir les lois non informatives. De plus, elle permet bien souvent de retrouver les estimateurs classiques.

Exemple 3.30. Soit $x \sim \mathcal{N}(\theta, I_p)$. Comme il s'agit d'une famille de position, la loi de Jeffreys est constante. L'estimateur de Bayes généralisé est donné par

$$\delta^{\pi^*}(x) = \frac{\int_{\mathbb{R}^p} \theta \exp(-||x-\theta||^2/2)\, d\theta}{\int_{\mathbb{R}^p} \exp(-||x-\theta||^2/2)\, d\theta} = x.$$

Il est minimax pour tout p et admissible pour $p \leq 2$. Notons que cet estimateur est aussi le meilleur estimateur équivariant pour des paramètres de position (voir le Chapitre 9). ‖

Exemple 3.31. Soit $x \sim \mathcal{N}(\mu, \sigma^2)$ avec $\theta = (\mu, \sigma)$ inconnu. Dans ce cas,

$$\begin{aligned} I(\theta) &= \mathbb{E}_\theta \left[\begin{pmatrix} 1/\sigma^2 & 2(x-\mu)/\sigma^3 \\ 2(x-\mu)/\sigma^3 & 3(\mu-x)^2/\sigma^4 - 1/\sigma^2 \end{pmatrix} \right] \\ &= \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix} \end{aligned}$$

et la loi non informative associée est $\pi(\theta) \propto 1/\sigma^2$. Si, en revanche, on suppose μ et σ indépendants, la loi non informative correspondante est $\pi(\mu, \sigma) = \sigma^{-1}$, qui est aussi la mesure invariante de Haar pour ce modèle de position-échelle (voir l'Exemple 3.28 et le Chapitre 9). ‖

Cette approche est critiquée par certains bayésiens comme étant un outil sans justification subjective en termes d'information a priori. Cependant, la seule alternative à une approche automatique est d'exiger que l'information a priori soit toujours disponible, ce qui n'est pas possible dans tous les cadres. Une autre critique de la méthode de Jeffreys est que, bien qu'elle réponde aux exigences d'*invariance par reparamétrisation*, elle ne satisfait pas au principe de vraisemblance. En effet, l'information de Fisher peut différer pour deux expériences fournissant des vraisemblances proportionnelles, comme le montre l'exemple ci-dessous.

Exemple 3.32. Nous avons vu dans l'Exemple 1.16 que les modèles binomial et binomial négatif conduisent à la même vraisemblance. Cependant, si $x \sim \mathscr{B}(n, \theta)$, la loi non informative $\pi_1(\theta)$ est $\mathscr{B}e(1/2, 1/2)$ (Exemple 3.26) et, si $n \sim \mathscr{N}eg(x, \theta)$, la loi de Jeffreys est

$$\begin{aligned}\pi_2(\theta) &= -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)\right]\\ &= \mathbb{E}_\theta\left[\frac{x}{\theta^2}+\frac{n-x}{(1-\theta)^2}\right] = \frac{x}{\theta^2(1-\theta)},\end{aligned}$$

soit donc $\pi_2(\theta) \propto \theta^{-1}(1-\theta)^{-1/2}$, qui est impropre et, fait plus important, diffère de π_1. ‖

Comme le montre l'exemple suivant, il arrive souvent que la loi non informative de Jeffreys soit limite de lois conjuguées.

Exemple 3.33. Si $x \sim \mathscr{U}([0, \theta])$, une loi conjuguée est la loi de Pareto, $\mathscr{P}a(\theta_0, \alpha)$,

$$\pi(\theta) = \alpha\,\theta_0^\alpha\,\theta^{-\alpha-1}\mathbb{I}_{[\theta_0,+\infty[}(\theta),$$

qui donne la loi a posteriori $\mathscr{P}a(\max(\theta_0, x), \alpha+1)$. Sous le coût invariant

$$L(\theta, \delta) = \frac{(\theta-\delta)^2}{\theta^2},$$

l'estimateur de Bayes est, si $\theta_0 \vee x = \max(\theta_0, x)$,

$$\delta^\pi(x) = \frac{\int_{\theta_0\vee x}^{+\infty}\theta^{-1}(\alpha+1)\,\theta_0^{\alpha+1}\theta^{-\alpha-2}d\theta}{\int_{\theta_0\vee x}^{+\infty}\theta^{-2}(\alpha+1)\,\theta_0^{\alpha+1}\theta^{-\alpha-2}d\theta} = \frac{\alpha+3}{\alpha+2}\,(\theta\vee x),$$

qui tend vers l'estimateur minimax, $\delta_0(x) = (3/2)x$, quand α et θ_0 tendent vers 0. Comme θ est un paramètre d'échelle, la loi non informative est $\pi(\theta) = 1/\theta$,

qui est aussi une loi de Jeffreys pour ce modèle. Cette loi correspond à $\theta_0 = 0$ et $\alpha = 0$ pour une loi de Pareto non normalisée (c'est-à-dire sans le facteur d'échelle $\alpha\theta_0^\alpha$). Cette représentation permet par ailleurs de prouver que δ_0 est admissible, en utilisant la *condition d'admissibilité suffisante de Stein* (voir le Chapitre 8). ‖

Un inconvénient plus important de la loi non informative de Jeffreys est qu'elle ne donne pas des résultats satisfaisants pour tous les buts inférentiels, en particulier lorsqu'on considère des sous-vecteurs d'intérêt. Le problème ci-dessous a été mis en évidence par Stein (1959) (voir aussi Tibshirani, 1989).

Exemple 3.34. Si $x \sim \mathcal{N}_p(\theta, I_p)$, la loi non informative est $\pi(\theta) = 1$. L'estimateur résultant de θ, x, est assez raisonnable, comme le montre l'Exemple 3.30. Cependant, comme $\theta|x \sim \mathcal{N}_p(x, I_p)$, la loi a posteriori de $\eta = ||\theta||^2$ est $\chi^2_p(||x||^2)$, la loi du khi deux décentré. Quand η est le paramètre d'intérêt, l'espérance a posteriori de η est

$$\delta^\pi(x) = \mathbb{E}^\pi[\eta|x] = ||x||^2 + p\,.$$

Cependant, le meilleur estimateur parmi les estimateurs de la forme $||x||^2 + c$ (pour le coût quadratique) est $||x||^2 - p$, qui domine uniformément l'estimateur généralisé de Bayes, δ^π (voir l'Exercice 2.35). Par conséquent, la loi marginale sur η déduite de la loi non informative de Jeffreys sur θ est véritablement sous-optimale. De plus, la loi non informative de Jeffreys obtenue à partir de l'observation réduite $z = ||x||^2$ est totalement différente de $\chi^2_p(||x||^2)$ et conduit à un estimateur de η aux performances beaucoup plus acceptables (voir l'Exercice 3.53). ‖

L'Exercice 4.48 montre aussi que la loi de Jeffreys a priori peut être inconsistante dans le cadre d'une calibration linéaire et que ce problème peut être résolu par la méthode des lois a priori de référence.

3.5.4 Lois de référence

Le type de problème évoqué à la fin de la section précédente a été pris en compte par Bernardo (1979), qui propose une modification de l'approche de Jeffreys appelée *approche de la loi de référence*. Une différence majeure est que cette méthode fait la distinction entre paramètres d'intérêt et paramètres de nuisance (par exemple, $||\theta||^2$ et $\theta/||\theta||$ dans l'Exemple 3.34). Par conséquent, la loi a priori résultante ne dépend pas seulement de la loi d'échantillonnage, mais aussi du problème inférentiel considéré. Le reste de cette section présente brièvement la construction des lois de référence. Pour une étude détaillée, voir Berger et Bernardo (1989, 1992b,a) et Kass et Wasserman (1996).

Quand $x \sim f(x|\theta)$ et $\theta = (\theta_1, \theta_2)$, où θ_1 est le paramètre d'intérêt, la loi de référence est obtenue en définissant d'abord $\pi(\theta_2|\theta_1)$ comme la loi de Jeffreys associée à $f(x|\theta)$ pour θ_1 fixé, puis en calculant la loi marginale

$$\tilde{f}(x|\theta_1) = \int f(x|\theta_1, \theta_2)\pi(\theta_2|\theta_1)d\theta_2 \tag{3.14}$$

et la loi de Jeffreys $\pi(\theta_1)$ associée à $\tilde{f}(x|\theta_1)$. Le principe sous-jacent à la loi de référence est donc d'éliminer le paramètre de nuisance en utilisant la loi de Jeffreys correspondant au cas où le paramètre d'intérêt reste fixé. (Notons que l'intégrale dans (3.14) n'est pas forcément définie et il peut être nécessaire d'intégrer d'abord sur une suite d'ensembles compacts et de prendre la limite.)

Exemple 3.35. Le problème de Neyman-Scott (1948) est relié à l'observation de x_{ij} distribués selon $\mathcal{N}(\mu_i, \sigma^2)$, $i = 1, \ldots, n$, $j = 1, 2$. La loi de Jeffreys usuelle pour ce modèle est $\pi(\mu_1, \ldots, \mu_n, \sigma) = \sigma^{-n-1}$ et une inconsistance apparaît, car $\mathbb{E}[\sigma^2|x_{11}, \ldots, x_{n2}] = s^2/(2n-2)$, avec

$$s^2 = \sum_{i=1}^{n} \frac{(x_{i1} - x_{i2})^2}{2},$$

cette espérance a posteriori convergeant en n vers $\sigma^2/2$. (Notons qu'il s'agit d'un cas où le nombre de paramètres augmente avec le nombre d'observations.) La loi de référence associée à $\theta_1 = \sigma$ et $\theta_2 = (\mu_1, \ldots, \mu_n)$ donne une loi plate pour $\pi(\theta_2|\theta_1)$, car θ_2 est un paramètre de position. Alors

$$\tilde{f}(x|\theta_1) = \prod_{i=1}^{n} e^{-(x_{i1}-x_{i2})^2/4\sigma^2} \frac{1}{\sqrt{2\pi}2\sigma}$$

est une famille d'échelle et $\pi(\sigma) = 1/\sigma$. Par conséquent, $\mathbb{E}[\sigma^2|x_{11}, \ldots, x_{n2}] = s^2/(n-2)$, qui est convergent. ‖

La construction générale d'une loi de référence est la suivante : Soit $x \sim f(x|\theta)$, avec $\theta \in \Theta \subset \mathbb{R}^k$. Supposons que la matrice d'information de Fisher $I(\theta)$ existe et soit de plein rang. Notons $\mathbf{S} = I^{-1}(\theta)$. Les paramètres sont désormais séparés en m groupes correspondant à leur importance respective,

$$\theta_{(1)} = (\theta_1, \ldots, \theta_{n_1}), \quad \ldots \quad \theta_{(m)} = (\theta_{N_{m-1}+1}, \ldots, \theta_k), \tag{3.15}$$

avec $N_i = \sum_{j=1}^{i} n_j$ (après un possible changement d'indices des composants de θ). La méthode de la loi de référence construit une loi a priori sur $(\theta_{(1)}, \ldots, \theta_{(m)})$ qui prend en compte cette décomposition, c'est-à-dire qui fait vraiment la séparation entre paramètres de nuisance et paramètres d'intérêt. Elle permet même un niveau plus fin de séparation entre les niveaux d'importance respectifs de ces paramètres. Nous introduisons la notation suivante : pour $j = 1, \ldots, m$,

$$\theta_{[j]} = (\theta_{(1)}, \ldots, \theta_{(j)}) \qquad \text{et} \qquad \theta_{[\sim j]} = (\theta_{(j+1)}, \ldots, \theta_{(m)}).$$

La matrice $\mathbf{S}$ est décomposée selon la partition (3.15),

$$\mathbf{S} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{21}^t & \dots & \mathbf{A}_{m1}^t \\ \mathbf{A}_{21} & \mathbf{A}_{22} & & \mathbf{A}_{m2}^t \\ & & \dots & \\ \mathbf{A}_{m1} & & & \mathbf{A}_{mm} \end{pmatrix}$$

et S_j est le coin supérieur gauche (N_j, N_j) de $\mathbf{S}$; par exemple, $S_1 = \mathbf{A}_{11}$. Nous notons $\mathbf{H}_j = S_j^{-1}$ et $\mathbf{h}_j$ le coin en bas à droite (n_j, n_j) de $\mathbf{H}_j$; en particulier, $\mathbf{h}_1 = \mathbf{A}_{11}^{-1}$. La construction de la loi de référence continue comme suit :

ALGORITHME 3.1. **Loi de référence**

– Initialisation :

$$\pi_m(\theta_{(m)}|\theta_{[m-1]}) = \frac{|\mathbf{h}_m(\theta)|^{1/2}}{\int |\mathbf{h}_m(\theta)|^{1/2}\, d\theta_{(m)}}.$$

– Itération : For $j = m-1, \dots, 1$,

$$\pi_j(\theta_{[\sim j-1]}|\theta_{[j-1]}) = \frac{\pi_{j+1}(\theta_{[\sim j]}|\theta_{[j]}) \exp\{\frac{1}{2}\mathbb{E}_j[\log(|\mathbf{h}_j(\theta)|)|\theta_{[j]}]\}}{\int \exp\{\frac{1}{2}\mathbb{E}_j[\log(|\mathbf{h}_j(\theta)|)|\theta_{[j]}]\}\, d\theta_{(j)}},$$

où

$$\mathbb{E}_j[g(\theta)|\theta_{[j]}] = \int g(\theta)\pi_{j+1}(\theta_{[\sim j]}|\theta_{[j]})\, d\theta_{[\sim j]}.$$

– Conclusion : La loi de référence est $\pi(\theta) = \pi_1(\theta_{[\sim 0]}|\theta_{[0]})$.

Souvent, quelques-unes des intégrales apparaissant dans cet algorithme ne sont pas définies. Berger et Bernardo (1989) ont alors proposé de calculer la loi de référence pour des sous-ensembles compacts Θ_n de Θ et de considérer la limite de la suite de lois de référence correspondante (π_n) quand n tend vers l'infini et Θ_n tend vers Θ. En général, le résultat limite ne dépend pas du choix de la suite de compacts.

Exemple 3.36. (Suite de l'Exemple 3.34) Puisque $\eta = ||\theta||^2$ est le paramètre d'intérêt, θ peut s'écrire en coordonnées polaires $(\eta, \varphi_1, \dots, \varphi_{p-1})$, avec

$$\begin{aligned} \theta_1 &= \sqrt{\eta}\cos(\varphi_1), \\ \theta_2 &= \sqrt{\eta}\sin(\varphi_1)\cos(\varphi_2), \\ &\dots \\ \theta_{p-1} &= \sqrt{\eta}\sin(\varphi_1)\cdots\cos(\varphi_{p-1}), \\ \theta_p &= \sqrt{\eta}\sin(\varphi_1)\cdots\sin(\varphi_{p-1}). \end{aligned}$$

La matrice d'information de Fisher pour $(\eta, \varphi_1, \dots, \varphi_{p-1})$ est alors $\mathbf{H} = \mathbf{J}\mathbf{J}^t$, où $\mathbf{J}$ est la matrice jacobienne $\frac{D(\theta_1,\dots,\theta_p)}{D(\eta,\varphi_1,\dots,\varphi_{p-1})}$. On peut montrer que $\mathbf{J}$ est de la forme

$$\mathbf{J} = \begin{bmatrix} A^t/\sqrt{\eta} \\ \sqrt{\eta}\mathbf{B} \end{bmatrix},$$

avec des matrices $A \in \mathbb{R}^p$ et $\mathbf{B}$ $(p-1) \times p$. Alors, pour la partition de θ en $\theta_{(1)} = \eta$, $\theta_{(2)} = (\varphi_1, \ldots, \varphi_{p-1})$, nous avons

$$\pi_2(\varphi_1, \ldots, \varphi_{p-1}|\eta) \propto |\mathbf{H}_{22}|^{1/2},$$

qui ne dépend pas de η. La loi marginale de η est

$$\pi_1(\eta) \propto \exp\left\{\mathbb{E}\left[\log\left(\frac{1}{2}\frac{|\mathbf{H}|}{|\mathbf{H}_{22}|}\right)\middle|\eta\right]\right\}$$

et $\frac{|\mathbf{H}|}{|\mathbf{H}_{22}|} \propto (1/\eta)$. Par conséquent,

$$\pi_1(\eta) = 1/\sqrt{\eta},$$

qui mène à un estimateur de $||\theta||^2$ plus intéressant que $||x||^2+p$ (voir l'Exercice 3.53).

En réalité, le même problème de marginalisation apparaît pour l'estimation du maximum de vraisemblance. En effet, l'estimateur du maximum de vraisemblance fondé sur l'échantillon est $||x||^2$, qui est aussi dominé par $||x||^2 - p$. En revanche, l'estimateur du maximum de vraisemblance obtenu à partir de

$$z = ||x||^2 \sim \chi^2_p(||\theta||^2)$$

se conduit de la même façon que $(||x||^2 - p)^+$ (voir Saxena et Alam, 1982, Chow, 1987, et Chow et Hwang, 1990, et l'Exercice 3.53). ‖

Cet algorithme se justifie comme fournissant la loi a priori qui maximise l'information a posteriori (Bernardo, 1979, et Berger et Bernardo, 1992a). Plus précisément, si l'échantillon $(x_1, \ldots, x_n)$ est noté $x_{1:n}$ et si $K_n(\pi)$ est la divergence de Kullback-Leibler entre la loi a priori π et la loi a posteriori correspondante,

$$K_n(\pi) = \int \pi(\theta|x_{1:n}) \log\left(\pi(\theta|x_{1:n})/\pi(\theta)\right)\, d\theta,$$

l'idée de Bernardo (1979) est d'utiliser $\mathbb{E}[K_n(\pi)]$, où l'espérance est prise sur la loi marginale de $x_{1:n}$, comme mesure d'*information manquante*, et de définir la loi de référence comme la loi π maximisant

$$K^*(\pi) = \lim_{n\to\infty} \mathbb{E}[K_n(\pi)]\,.$$

Les difficultés techniques associées aux éventuelles intégrales infinies mises à part, la loi a priori résultante est la loi de Jeffreys pour des espaces continus des paramètres et la loi uniforme pour des espaces finis ; voir Ghosh et Mukerjee

(1992a), Clarke et Wasserman (1993) et Kass et Wasserman (1996) pour des motivations supplémentaires en termes d'optimalité asymptotique.

La loi de référence dépend aussi de la façon dont les paramètres ont été ordonnés (voir l'Exercice 3.60), un avantage comparé à la méthode de Jeffreys, car les paramètres de nuisance sont considérés différemment. Des paradoxes comme ceux de l'Exemple 3.34 sont alors évités. Il peut paraître excessif de modifier la loi a priori selon le problème d'intérêt, mais on doit se rendre compte que, mis à part la distribution de l'échantillon $f(x|\theta)$, ces problèmes inférentiels sont la seule information disponible[25]. Notons que l'invariance par reparamétrisation n'est maintenue que si les changements sont bijectifs et internes à chaque groupe dans (3.15). Cependant, l'exigence d'invariance est moins importante dans ce cadre parce que l'ordre (3.15) interdit d'une certaine manière une reparamétrisation entre les catégories, puisque les différents groupes ne sont pas du même type. Quand un tel ordre ne peut pas être proposé, Berger et Bernardo (1992b) suggèrent de considérer comme loi non informative la loi de référence correspondant au cas où chaque composante de θ est traitée séparément. (Par comparaison, la loi de Jeffreys traite θ comme un seul groupe.)

Exemple 3.37. (Berger et Bernardo, 1992b) Soit un modèle d'analyse de la variance

$$x_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, \ldots, p, \quad j = 1, \ldots, n,$$

avec $\alpha_i \sim \mathcal{N}(0, \tau^2)$, $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. Pour différents ordres des paramètres, $\mu,\ \tau^2,\ \sigma^2$, nous obtenons les lois de référence suivantes :

$$\begin{aligned}
\pi_1((\mu, \sigma^2, \tau^2)) &\propto \sigma^{-2}(n\tau^2 + \sigma^2)^{-3/2} \\
\pi_2(\mu, \sigma^2, \tau^2) &\propto \tau^{-C_n}\sigma^2 \left[(n-1) + (1 + n\tau^2/\sigma^2)^{-2}\right]^{1/2} \\
\pi_3(\mu, (\sigma^2, \tau^2)) &\propto \sigma^{-2}(n\tau^2 + \sigma^2)^{-1} \\
\pi_4((\mu, \sigma^2), \tau^2) &\propto \sigma^{-5/2}(n\tau^2 + \sigma^2)^{-1}
\end{aligned}$$

avec $C_n = \{1 - \sqrt{n-1}(\sqrt{n} + \sqrt{n-1})^{-3}\}$. ‖

3.5.5 Lois a priori coïncidantes

Une approche particulière, pour ne pas dire paradoxale, de la modélisation non informative est de s'intéresser aux propriétés fréquentistes de la loi a priori, c'est-à-dire en moyenne sur x plutôt que conditionnellement à x. Notons tout d'abord, comme cela est discuté dans les Chapitres 2 et 8, qu'il existe des

[25] Si une fonction de coût L est disponible, elle contient aussi quelque information sur θ et la dualité entre fonction de coût et loi a priori peut être utilisée pour obtenir une loi a priori adaptée à ce coût (voir Rubin, 1987). Mais très peu a été fait sur la construction de la loi a priori à partir d'une fonction de coût.

lois a priori donnant des estimateurs optimaux selon des critères fréquentistes comme la minimaxité ou l'admissibilité, et on peut souhaiter restreindre le choix de la loi a priori à ces distributions optimales. Cependant, une telle restriction réduit rarement le choix de la loi a priori à une distribution unique. Soit aucune loi ne vérifie cette condition, notamment en petite dimension pour l'estimation sous le coût quadratique (Note 2.8.2), soit une infinité de lois sont, par exemple, associées aux estimateurs minimax admissibles (Fourdrinier *et al.*, 1998). (Une exception se produit lorsque des structures d'invariance existent, auquel cas la mesure de Haar à droite est le choix approprié, comme cela est expliqué dans la Section 3.5.2.)

Une approche plus standard est d'imposer que certaines probabilités a posteriori coïncident, jusqu'à un certain degré d'approximation, avec la couverture fréquentiste correspondante; d'où l'appellation de *lois a priori coïncidantes* (traduction de *matching priors*), qu'on restreint souvent dans la littérature aux intervalles de confiance unilatéraux. Soit un ensemble de confiance a posteriori C_x donné sur $g(\theta)$,

$$\pi(g(\theta) \in C_x|x) = 1 - \alpha\,,$$

unilatéral ou bilatéral. Cet ensemble définit alors un ensemble de confiance au sens fréquentiste, de couverture

$$P_\theta(C_x \ni g(\theta)) = \int \mathbb{I}_{C_x}(g(\theta))\, f(x|\theta)\, dx\,,$$

qui diffère généralement de $1-\alpha$. Lorsque des quantités pivotales existent, comme dans le cas normal $\mathcal{N}(\theta, 1/n)$, la région de plus forte densité a posteriori (HPD) au niveau $1-\alpha$ (Chapitre 5) est donnée par

$$C_x = [\bar{x}_n - n^{-1/2} q_{\alpha/2}, \bar{x}_n + n^{-1/2} q_{\alpha/2}]\,,$$

où $q_{\alpha/2}$ est le quantile au niveau $1-\alpha/2$ d'une loi normale, et la couverture fréquentiste de C_x vaut aussi $1-\alpha$. Lindley (1957) généralise ce résultat à d'autres familles de position et démontre qu'il ne se vérifie que pour de telles familles. Dans un cadre général (unidimensionnel), Welch et Peers (1963) et Welch (1965) ont démontré que, lorsque $C_x = (-\infty, k_\alpha(x)]$,

$$P_\theta(\theta \leq k_\alpha(x)) = 1 - \alpha + O(n^{-1/2})\,,$$

et que, pour la loi a priori de Jeffreys,

$$P_\theta(\theta \leq k_\alpha(x)) = 1 - \alpha + O(n^{-1})\,,$$

ce qui améliore l'approximation d'un facteur $1/2$.

Les choses se compliquent en présence de paramètres de nuisance, c'est-à-dire lorsque l'inférence porte sur une composante unidimensionnelle θ_1 du paramètre. Des références sur des travaux dans ce domaine incluent Sweeting (1985), Severini (1991), Ghosh et Mukerjee (1992a,b, 1993), Mukerjee et

Dey (1993), DiCiccio et Stern (1993, 1994), Liseo (1993), et Datta et Ghosh (1995a,b). Nous nous concentrons ici sur certains des résultats obtenus par Rousseau (1997, 2000, 2001).

Le *développement d'Edgeworth* (voir Bhattacharya et Rao, 1986, Bickel et Ghosh, 1990, et DiCiccio et Stern, 1994) de la probabilité de couverture fréquentiste est donnée par

$$P_\theta(\theta_1 < k_n(\alpha)) = 1-\alpha + \frac{\varphi(\Phi^{-1}(1-\alpha))}{\sqrt{n}}\left(\frac{I'(\theta)\nabla\log\pi(\theta)}{I''(\theta)^{1/2}} - \nabla^t\frac{I'(\theta)}{I''(\theta)^{1/2}}\right) + O(n^{-1})\,,$$

dans le cas unilatéral, où φ et Φ sont respectivement la densité et la fonction de répartition d'une loi normale, et $I(\theta)$, $I'(\theta)$, et $I''(\theta)$ sont respectivement l'information de Fisher et ses dérivées première et seconde. Dans le cas d'une région HPD bilatérale de niveau $1-\alpha$, $C_x^{HPD}(\alpha)$, pour $\theta \in \mathbb{R}$, le développement correspondant est

$$P_\theta(\theta \in C_x^{HPD}) = 1-\alpha + n^{-1}q(\alpha)b(\pi,\theta) + O(n^{-3/2})\,,$$

où q correspond à une densité du χ^2 et

$$b(\pi,\theta) = \frac{\mu_3'-\mu_2''}{I(\theta)^2} + 2\frac{\mu_2'(\mu_3-\mu_2')}{I(\theta)^3} + \frac{\pi'(\theta)}{\pi(\theta)}\,\frac{\mu_3-\mu_2'}{I(\theta)^2} - \frac{\pi''(\theta)}{\pi(\theta)I(\theta)} - \frac{\mu_2'\pi'(\theta)}{\pi(\theta)I(\theta)^2}\,,$$

les μ_j étant définis par $(j=2,3)$

$$\mu_j = \mathbb{E}_\theta\left[\frac{\partial^j \log f(x|\theta)}{\partial\theta^j}\right].$$

La loi a priori coïncidante est alors obtenue par annulation du terme d'ordre un de ce développement, comme dans l'équation différentielle de Welch et Peers (1963) :

$$[I''(\theta)]^{-1/2}I'(\theta)\nabla\log\pi(\theta) + \nabla^t\{I'(\theta)[I''(\theta)]^{-1/2}\} = 0\,.$$

Cette équation différentielle peut ne pas avoir de solution. De plus, comme le montre la généralisation de Rousseau (2000) aux régions HPD, cette solution, lorsqu'elle existe, dépend du paramètre d'intérêt correspondant à ces régions HPD et diffère le plus souvent de la loi a priori de Jeffreys, même s'il existe toujours une paramétrisation permettant de retomber sur cette dernière.

Exemple 3.38. (Rousseau, 2000) Soit la loi $\mathscr{G}(k,\theta)$. Si θ est le paramètre d'intérêt, les lois a priori permettant d'annuler le terme de second ordre pour des régions HPD sont de la forme

$$\pi(\theta) = \frac{c_1+c_2\theta}{\theta}, \qquad c_1, c_2 > 0,$$

et incluent donc la loi a priori de Jeffreys comme cas particulier. Si $\eta = c_1\theta^{5/3} + c_2 \log(\theta)$ est la quantité d'intérêt, correspondant à la paramétrisation du χ^2, la loi a priori de coïncidence maximale est

$$\pi(\eta) = I(\eta)^{-1}\,,$$

et diffère de la loi de Jeffreys, $I(\eta)^{1/2}$. Enfin, considérons la paramétrisation de la moyenne, $\mu = k/\theta$. Les lois a priori coïncidantes sont alors de la forme

$$\pi(\mu) = c_1\mu^2 + c_2/\mu\,, \qquad c_1, c_2 > 0,$$

et, de nouveau, n'incluent pas la loi de Jeffreys. ‖

On peut aussi consulter Rousseau (1997) pour une extension au cadre discret où une coïncidence ne peut pas être obtenue pour des ordres supérieurs à $n^{1/2}$ et où une randomisation est nécessaire pour atteindre de tels ordres.

Exemple 3.39. (Ghosh *et al.*, 1995) Une version simple du *modèle de calibration linéaire* est ($i = 1, \ldots, n$, $j = 1, \ldots, k$),

$$y_i = \alpha + \beta x_i + \varepsilon_i, \qquad y_{0j} = \alpha + \beta x_0 + \varepsilon_{0j}\,, \tag{3.16}$$

où x_0, inconnue, est la quantité d'intérêt (voir l'Exercice 4.48 pour plus de détails sur ce modèle). Pour des intervalles de confiance unilatéraux, l'équation différentielle associée à (3.16) est alors

$$\begin{aligned}&|\beta|^{-1}s^{-1/2}\frac{\partial}{\partial x_0}\{e(x_0)\pi(\theta)\} - e^{-1/2}(x_0)\mathrm{sgn}(\beta)n^{-1}s^{1/2}\frac{\partial\pi(\theta)}{\partial x_0}\\ &-e^{-1/2}(x_0)(x_0-\bar{x})s^{-1/2}\frac{\partial}{\partial\beta}\{\mathrm{sgn}(\beta)\pi(\theta)\} = 0\end{aligned}$$

où $\theta = (x_0, \alpha, \beta, \sigma^2)$ et

$$s = \Sigma(x_i - \bar{x})^2, \;\; e(x_0) = [(n+k)s + nk(x_0 - \bar{x})^2]/nk\,.$$

Les solutions de cette équation différentielle sont alors de la forme

$$\pi(x_0, \alpha, \beta, \sigma^2) \propto e(x_0)^{(d-1)/2}|\beta|^d g(\sigma^2)\,, \tag{3.17}$$

où g est arbitraire. Par exemple, si $g(\sigma^2) = (\sigma^2)^{-a/2}$, la loi a posteriori correspondante est propre si $(n + k + a - 2d - 5) > 0$. Dans ce cas, les lois a priori de référence sont aussi coïncidantes, c'est-à-dire satisfont (3.17), comme l'illustre le Tableau 3.5 pour quatre ordres différents sur les paramètres. ‖

En général, des lois a priori de référence (inverses) sont coïncidantes lorsque le paramètre d'intérêt, λ, et le paramètre de nuisance, ω, sont orthogonaux au sens de l'information de Fisher

Tab. 3.5. Lois a priori de référence coïncidantes associées à différents ordres pour le modèle de calibration linéaire (3.16).

Partition	a priori
$(x_0, \alpha, \beta, \sigma^2)$	$\lvert\beta\rvert(\sigma^2)^{-5/2}$
$x_0, \alpha, \beta, \sigma^2$	$e(x_0)^{-1/2}(\sigma^2)^{-1}$
$x_0, \alpha, (\sigma^2, \beta)$	$e(x_0)^{-1/2}(\sigma^2)^{-3/2}$
$x_0, (\alpha, \beta), \sigma^2$	$e(x_0)^{-1/2}(\sigma^2)^{-1}$
$x_0, (\alpha, \beta, \sigma^2)$	$e(x_0)^{-1/2}(\sigma^2)^{-2}$

$$I(\lambda, \eta) = \begin{pmatrix} I_{11} & 0 \\ 0 & I_{22} \end{pmatrix},$$

comme détaillé dans Tibshirani (1989), et aussi lorsqu'on utilise l'ordre inverse (ω, λ) pour construire l'a priori de référence, comme cela est expliqué dans Berger *et al.* (1998).

Au-delà de la difficulté technique de cette approche, il est conceptuellement peu élégant d'imposer à une loi a priori des propriétés fréquentistes, alors même que cette loi permet de conditionner en x plutôt que de recourir à des propriétés sur le long terme. Tenter de réconcilier les deux approches (bayésienne et fréquentiste) ne doit pas être rejeté systématiquement, comme cela est expliqué dans le Chapitre 5, mais ce changement de paradigme est plutôt gênant, comme l'illustre Rousseau (1997) qui doit recourir à la randomisation, en violation du principe de vraisemblance. Nous ne le recommandons donc pas.

3.5.6 D'autres approches

Des alternatives à une analyse bayésienne non informative sont décrites dans Berger (1985b, Chapitre 3) et Kass et Wasserman (1996). Nous mentionnons par exemple Rissanen (1983, 1990), qui recourt à la théorie de la transmission d'information de Shannon (1948). Considérant la transmission d'un message binaire par un appareil physique, la loi a priori non informative pour un modèle $f(x|\theta)$ est la longueur minimale d'un message décrivant ce modèle. Dans le cas le plus simple, ces lois sont similaires à celle de Jeffreys. Cette similarité devrait se vérifier en général, de part les connexions qui existent entre information statistique et théorie de l'information. Une revue récente de cette théorie de la *complexité stochastique* est donnée par Dawid (1992) ; voir aussi Hansen et Yu (2000).

Notons aussi que la mise en œuvre des tests requiert des lois a priori particulières, comme le signalent Hansen et Yu (2000) et Kass et Wasserman (1996). Nous traiterons ce point particulier dans le Chapitre 5.

3.6 Validation a posteriori et robustesse

Même dans les situations où l'information a priori est disponible, il est rare de pouvoir proposer une détermination exacte de la loi a priori $\pi(\theta)$ à partir de cette information, ne serait-ce que parce que le pouvoir de discrimination des individus est fini et la détermination des queues de distribution est impossible en pratique. Dans la plupart des cas, une certaine imprécision sur la loi a priori employée dans une inférence bayésienne demeure donc.

Si l'information a priori est riche, la loi a priori sera bien entendu mieux définie que dans un cadre non informatif. Cependant, il est important dans tous les cas de s'assurer que l'impact de cette indétermination de la loi a priori sur les quantités a posteriori soit bien évalué et que la partie arbitraire de l'a priori ne soit pas prédominante. L'étude de ces aspects est dite *analyse de sensibilité* (ou *de robustesse*). La notion de robustesse et la construction d'outils appropriés pour traiter ce problème particulier apparaissent dans les travaux de Good (1983) et Berger (1982a, 1984, 1985b, 1990). D'autres références sont Berger et Berliner (1986), Berger et Sellke (1987), Berger et Delampady (1987), O'Hagan (1988), Sivaganesan et Berger (1989),Walley (1991), Wasserman (1992) et Abraham et Daurés (2000).

Suivant la classification de Berger (1990), nous considérons que l'incertitude portant sur la loi a priori π peut se représenter par une classe Γ de lois a priori, à laquelle π est supposée appartenir. Ces classes peuvent être déterminées selon des critères pratiques ou subjectifs. Les types de classes de robustesse les plus couramment rencontrés dans la littérature sont :

(i) *Classes de lois conjuguées.* Ces classes sont typiquement choisies pour des raisons pratiques, parce qu'elles fournissent en général des bornes explicites pour les quantités d'intérêt. Par exemple, Das Gupta et Studden (1988) considèrent le cas où $x \sim \mathcal{N}_p(\theta, I_p)$ et $\theta \sim \mathcal{N}_p(0, \Sigma)$, avec $\Sigma_1 \preceq \Sigma \preceq \Sigma_2$, la relation d'ordre $\preceq$ étant vérifiée lorsque la différence des deux matrices est semi-définie positive. Les critiques déjà évoquées sur les lois conjuguées s'appliquent bien entendu dans ce cadre et ce d'autant plus que la classe résultante ne contient que des lois "de convenance", dont assez peu sont compatibles avec l'information a priori.

(ii) *Classes à moments déterminés.* L'hypothèse que l'information a priori (limitée) ne peut se traduire que par des bornes sur certains moments de π correspond à la classe

$$\Gamma_M = \{\pi;\ a_i \leq \mathbb{E}^\pi[\theta^i] \leq b_i, i = 1, \ldots, k\}.$$

Cependant, Γ_M n'est pas tellement plus satisfaisante que la classe précédente, car elle impose des conditions fortes sur les queues de la loi a priori. De plus, elle contient des lois peu raisonnables, notamment des lois à support fini[26].

[26] Plus précisément, les bornes portant sur les quantités a posteriori sont atteintes dans la plupart des cas par des lois à support fini, pour des raisons de convexité.

(iii) *Classes de voisinages.* Introduites par Huber (1964b) pour la détection de points aberrants, les *classes d'ϵ-contamination* d'une loi π_0,

$$\Gamma_{\epsilon,\mathcal{Q}} = \{\pi = (1-\epsilon)\pi_0 + \epsilon q;\ q \in \mathcal{Q}\},$$

sont souvent utilisées dans les études de robustesse. Dans l'expression ci-dessus $\mathcal{Q}$ est une classe de distributions choisie en fonction de la précision de l'information a priori. Berger et Berliner (1986) et Berger (1990) donnent des exemples où de telles classes peuvent être utilisées. Le problème majeur lié à l'utilisation de $\Gamma_{\epsilon,\mathcal{Q}}$ est la détermination difficile de ϵ et de $\mathcal{Q}$, notamment à partir du degré d'incertitude sur π_0. Mais des techniques d'estimation de *mélanges* peuvent être utiles dans un tel cadre, lorsque l'information a priori est construite à partir d'un échantillon d'observations passées (éventuellement fictives) (voir la Section 6.4). Une autre relation est de considérer un véritable voisinage associé à une distance comme celles de Hellinger ou de Kullback-Leibler (voir la Section 2.5.4 et Zucchini, 1999). La difficulté est alors de choisir l'*échelle* de tels voisinages.

(iv) *Classes sous-spécifiées.* De telles classes résultent d'une construction de la loi a priori sur une sous-σ-algèbre, c'est-à-dire pour un ensemble plus fruste d'événements que celui d'intérêt. Cette approche est directement reliée aux développements axiomatiques de la Note 3.8.1, puisque l'ordre sur les vraisemblances relatives n'engendre pas forcément une loi a priori sur l'ensemble des boréliens. Par exemple, il se peut que certains des quantiles de la loi a priori soient déterminés,

$$\Gamma_Q = \{\pi;\ \ell_i \leq \int_{I_i} \pi(\theta)\, d\theta \leq u_i, i = 1, \ldots, m\}$$

où $I_1, \ldots, I_m$ est une partition de Θ. Ces classes sont préférables à (ii), mais il peut malgré tout être nécessaire de retirer de Γ_Q certaines lois a priori peu raisonnables, comme dans O'Hagan (1988). Cependant, cette approche semble être la plus réaliste, car, par exemple, il est généralement plus facile de déterminer des fractiles que des moments. Cette approche semble aussi la plus facile à mettre en œuvre parmi celles présentées ici.

(v) *Classes de rapport de densités.* Partant d'une construction subjective de la loi a priori comme dans le cas précédent, une autre solution est de considérer une représentation sous forme d'histogramme. Dès lors, l'incertitude sur l'information a priori peut se représenter par des bornes supérieure et inférieure pour la densité π, ce qui donne la classe

$$\Gamma_R = \{\pi;\ L(\theta) \leq \pi(\theta) \leq U(\theta)\},$$

où L et U sont données. Le choix de ces fonctions est délicat et a des conséquences importantes, car, si elles sont similaires, toutes les lois dans Γ_R auront le même type de queues ; voir DeRobertis et Hartigan (1981) et Abraham et Daurés (2000) pour des classes similaires.

Berger (1990) et Wasserman (1992) développent des outils numériques pour le calcul de bornes sur les quantités a posteriori, pour les classes ci-dessus. De fait, l'approche par robustesse remplace l'estimateur standard $\varrho(\pi)$ par l'ensemble des valeurs possibles pour cet estimateur lorsque la loi a priori π varie dans la classe Γ,

$$\varrho^L = \inf_{\pi\in\Gamma} \varrho(\pi), \qquad \varrho^U = \sup_{\pi\in\Gamma} \varrho(\pi).$$

Goutis (1990, 1994) (voir l'Exemple 3.6) donne une illustration de cette approche pour la classe (ii). Le Chapitre 5 en donne une autre pour l'obtention de bornes conservatrices sur la probabilité a posteriori d'une hypothèse nulle.

Une approche plus conservatrice de la notion de robustesse est de construire des *lois a priori robustes*, qui sont des lois paramétrées aussi peu dépendantes que possible de petites variations de l'information a priori. Par exemple, on peut montrer que les lois de Student sont préférables aux lois normales pour un modèle normal, même si ces dernières sont conjuguées pour ce modèle et qu'elles sont en fait d'entropie maximale dans certains cas (voir Zellner, 1971, Angers, 1987, et Angers et MacGibbon, 1990).

De même, les lois *poly-t* obtenues comme un produit de densités de Student sont utilisées dans l'analyse économétrique des *équations simultanées* pour la même raison (voir Drèze, 1976a, Richard et Tompa, 1980, et Bauwens, 1984). Le plus souvent, ces lois a priori robustes auront des queues épaisses, au contraire des lois conjuguées.

Une autre façon d'accroître la robustesse des lois conjuguées est d'introduire une modélisation hiérarchique. L'approche bayésienne hiérarchique est présentée dans le Chapitre 10, mais il semble d'ores et déjà tout à fait intuitif que l'ajout d'un niveau supplémentaire dans la modélisation a priori puisse améliorer la robustesse de la loi a priori. Considérons une loi conjuguée $\pi_1(\theta|\lambda)$ pour $f(x|\theta)$. Comme il est expliqué ci-dessus, des classes comme (i) ne sont pas très robustes et, de plus, nécessitent la spécification de bornes pour les hyperparamètres λ. Puisque ces hyperparamètres sont (partiellement ou totalement) inconnus, une extension naturelle (dans un cadre bayésien) est d'introduire une loi a priori non informative π_2 sur λ (ou une loi hyper a priori compatible avec l'information disponible). Cette modélisation donne la structure hiérarchique suivante :

$$\begin{aligned} \lambda &\sim \pi_2(\lambda), \\ \theta|\lambda &\sim \pi_1(\theta|\lambda), \\ x|\theta &\sim f(x|\theta). \end{aligned}$$

La loi a priori sur θ est alors la marginale de $\pi_1(\theta|\lambda)\pi_2(\lambda)$, après intégration par rapport à λ,

$$\pi(\theta) = \int \pi_1(\theta|\lambda)\pi_2(\lambda)d\lambda. \tag{3.18}$$

Cette loi a priori n'est généralement pas conjuguée, mais le but principal de cette extension hiérarchique est bien d'éviter le cadre trop restrictif des lois conjuguées. En intégrant sur les hyperparamètres λ, on obtient une distribution (3.18) qui se caractérise généralement par des queues plus épaisses que les lois conjuguées. Par exemple, la loi de Student peut s'écrire comme (3.18), où π_2 est une loi gamma inverse (voir l'Exemple 3.17). Les formulations hiérarchiques sont aussi intéressantes d'un point de vue numérique, comme expliqué dans le Chapitre 6.

D'autres approches prennent en compte la fonction de coût dans l'analyse de robustesse, afin d'obtenir un estimateur qui soit conservateur à l'égard de toutes les lois a priori possibles $\pi \in \Gamma$. Par exemple, δ^* peut être la solution de

$$\inf_{\delta} \sup_{\pi\in\Gamma} r(\pi,\delta) \qquad \text{ou} \qquad \inf_{\delta} \sup_{\pi\in\Gamma} [r(\pi,\delta) - r(\pi,\delta^{\pi})],$$

la première quantité étant le *risque Γ-minimax* et la seconde le *regret Γ-minimax* , comme l'ont développé Robbins (1951) et Good (1952) ; voir Berger et Berliner (1986), Berger (1985b), et Kempthorne (1988) pour de plus amples références.

La littérature sur la robustesse bayésienne s'est considérablement accrue ces dernières années et nous renvoyons les lecteurs aux articles cités ci-dessus pour de plus amples références. Pour conclure ce chapitre, remarquons que le choix de la loi a priori détermine l'inférence bayésienne qui en résulte, que ce choix est parfois trivial et parfois très délicat, mais qu'il doit se justifier dans tous les cas à partir de l'information a priori et, de plus, qu'une analyse de robustesse doit être mise en œuvre, afin d'établir l'impact sur l'a posteriori qu'un changement dans la loi a priori implique. Bien entendu, cette analyse dépendra de la façon dont on évalue l'impact sur les quantités d'intérêt, comme par exemple sur les *coûts* utilisés dans le processus d'estimation. Ceci permet d'utiliser la connaissance de la fonction de coût pour déterminer une loi a priori non informative, mais cette approche a été peu explorée, même si de nombreux bayésiens ont remarqué que fonction de coût et loi a priori ne peuvent être distinguées (voir notamment Lindley, 1985, et l'Exercice 3.58.) Un dernier avertissement aux lecteurs pour noter que l'influence de l'a priori est souvent sous-estimée par les utilisateurs, alors qu'elle peut avoir des conséquences inattendues sur l'inférence résultante. Dès lors, il est nécessaire de recourir dès que possible à d'autres valeurs pour les hyperparamètres, mais aussi à d'autres types de lois, afin d'établir l'impact réel du choix de la loi a priori sur l'inférence qui en résulte[27].

[27]Insistons de nouveau sur l'erreur commune qui consiste à croire que prendre des lois propres de grandes variances est un substitut acceptable aux lois non informatives.

3.7 Exercices

Section 3.1

3.1 (Dupuis, 1995b) Rappelons que la distribution bêta $\mathscr{B}e(\alpha,\beta)$ a pour densité

$$\pi(\theta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta}, \qquad 0\leq\theta\leq 1\,.$$

a. Donner l'espérance de la distribution $\mathscr{B}e(\alpha,\beta)$.

b. Montrer qu'il existe une bijection entre (α,β) et le triplet (μ,θ_0,θ_1), où $\pi(\theta\in[\theta_0,\theta_1])=p$ et μ est l'espérance de la distribution.

c. Quelles sont les conditions sur (μ,θ_0,θ_1) pour l'existence de (α,β) ?

Section 3.2.3

3.2 (Seidenfeld, 1987) Soit θ la variable aléatoire correspondant au résultat d'un lancer de dé à six faces.

a. Si π est la distribution de θ, donner l'a priori d'entropie maximale associé à l'information $\mathbb{E}[\theta]=3.5$.

b. Montrer que, si A est l'événement "θ est impair", la distribution actualisée $\pi(\cdot|A)$ est $(1/3,0,1/3,0,1/3,0)$.

c. Montrer que la loi a priori d'entropie maximale associée aux contraintes $\mathbb{E}[\theta]=3.5$ et $\mathbb{E}[\mathbb{I}_A]=1$ est $(.22,0,.32,0,.47)$.

[*Note :* Seidenfeld (1987) et Kass et Wasserman (1996) utilisent cet exemple pour montrer que l'approche de l'entropie maximale n'est pas toujours compatible avec le principe bayésien d'actualisation donné par (1.13).]

3.3 Montrer que, si les contraintes (3.1) sont toutes associées à des fonctions g_k de la forme $g_k(\theta)=\mathbb{I}_{(-\infty,a_k]}(\theta)$, il n'existe pas d'a priori d'entropie maximale lorsque $\Theta=\mathbb{R}$ et π_0 est la mesure de Lebesgue sur $\mathbb{R}$.

3.4 Soit $\theta\in\mathbb{R}$ et une loi a priori π telle que $\text{var}^{\pi}(\theta)=1$, $\pi(\theta<-1)=0.1$, et $\pi(\theta>1)=0.1$. Calculer l'a priori d'entropie maximale associé à la mesure de Lebesgue sur $\mathbb{R}$, si ce calcul est possible.

3.5 Soit π_0 une mesure de référence pour la méthode de l'entropie maximale et π_0' une mesure absolument continue par rapport à π_0.

a. Donner des exemples où les lois a priori d'entropie maximale associées à π_0 et π_0' coïncident.

b. Appliquer ce résultat au cas où π_0 est la mesure de Lebesgue sur $\mathbb{R}$, π_0' est la distribution $\mathscr{N}(0,1)$, et les contraintes (3.1) sont $\mathbb{E}^{\pi}[\theta]=0$, $\text{var}^{\pi}(\theta)=\sigma^2$, en fonction de la valeur de σ.

3.6 Soit $\theta\in\mathbb{R}_+$. Déterminer s'il existe une loi a priori d'entropie maximale sous la contrainte $\mathbb{E}^{\pi}[\theta]=\mu$ pour $\pi_0(\theta)=1$ et $\pi_0(\theta)=1/\theta$.

3.7 Soit $x\sim\mathscr{P}(\theta)$.

a. Déterminer la loi a priori d'entropie maximale associée à $\pi_0(\theta)=1/\sqrt{\theta}$ et $\mathbb{E}^{\pi}[\theta]=2$.

b. Déterminer les hyperparamètres de la loi a priori π lorsque π est de la forme
(i) $\mathscr{E}xp(\mu)$;
(ii) $\mathscr{G}(2,\varrho)$.

c. Calculer les trois lois a posteriori correspondantes lorsque $x = 3$ et comparer les estimateurs de Bayes de θ sous le coût $\mathrm{L}(\theta, \delta) = \theta(\theta - \delta)^2$.

Section 3.2.4

3.8 Déterminer les lois a priori dans l'Exemple 3.5, lorsque les premier et troisième quartiles sont 2 et -2, et la médiane est 0.

3.9 Soient $x \sim \mathscr{B}(n, \theta)$ et $\theta \sim \mathscr{B}e(\alpha, \beta)$. Déterminer s'il existe des valeurs de α, β telles que $\pi(\theta|x)$ soit la loi a priori uniforme sur $[0, 1]$, même pour une unique valeur de x.

3.10 Soient $x \sim \mathscr{P}a(\alpha, \theta)$, distribué selon une loi de Pareto, et $\theta \sim \mathscr{B}e(\mu, \nu)$. Montrer que, si $\alpha < 1$ et $x > 1$, un certain choix de μ et ν fait de $\pi(\theta|x)$ la loi a priori uniforme sur $[0, 1]$.

Section 3.3.1

3.11 Donner l'expression de $\pi(\theta|x)$ lorsque π est un mélange fini de distributions continues. En particulier, calculer les poids a posteriori. En déduire les résultats de l'Exemple 3.22.

3.12 Déterminer les *distributions symétriques*, c'est-à-dire telles que distributions d'échantillonnage et distributions conjuguées appartiennent à la même famille paramétrée.

3.13 Cet exercice montre que la notion de famille minimale conjuguée est en général sans intérêt.

a. En utilisant les notations de la Proposition 3.19, montrer que l'ensemble des λ dans l'expression $\pi(\theta|\mu, \lambda)$ peut se restreindre à ceux variant dans $\lambda_0 + \mathbb{N}$, pour n'importe quel $\lambda_0 > 0$.

b. En déduire que, si $\lambda_0 - \lambda_0' \notin \mathbb{Z}$, les familles conjuguées associées à $\lambda_0 + \mathbb{N}$ et $\lambda_0' + \mathbb{N}$ sont disjointes.

c. En conclure que l'intersection de toutes les familles conjuguées est vide.

3.14 Soit une population divisée en k catégories (ou *cellules*), se caractérisant par une probabilité p_i d'appartenir à la cellule i pour chaque individu ($1 \leq i \leq n$). Une suite (π_k) de lois a priori sur $p^k = (p_1, \ldots, p_k)$, $k \in \mathbb{N}$, est dite *cohérente* si tout regroupement de cellules en m catégories donne la loi a priori π_m pour les probabilités transformées.

a. Déterminer les conditions de cohérence sur la suite (π_k).

b. Dans le cas particulier où π_k est une loi de Dirichlet $\mathscr{D}_k(\alpha_1, \ldots, \alpha_k)$, exprimer ces conditions en fonction des α_k.

c. Est-ce que l'a priori de Jeffreys engendre une suite cohérente ?

d. Même question pour $\pi_k(p^k) \propto \prod_i p_i^{-1/k}$, comme proposé par Perk (1947).

Section 3.3.3

3.15 Montrer que toute distribution tirée d'une famille exponentielle peut se généraliser en une pseudo-famille exponentielle, par l'ajout de contraintes paramétriques sur le support de x. Commenter la modification de la statistique exhaustive.

3.16 Montrer que, si le support de $f(x|\theta)$ ne dépend pas de θ et s'il existe une famille a priori conjuguée paramétrée $\mathscr{F} = \{\pi(\theta|\lambda),\ \lambda \in \Lambda\}$ avec $\dim(\Lambda) < +\infty$, $f(x|\lambda)$ appartient nécessairement à une famille exponentielle. (*Indication :* C'est une conséquence du lemme de Pitman-Koopman.)

3.17 Donner une statistique exhaustive associée à l'échantillon $x_1, \ldots, x_n$ d'une loi de Pareto $\mathscr{P}a(\alpha, \theta)$.

3.18 Donner une statistique exhaustive associée à l'échantillon $x_1, \ldots, x_n$ d'une loi normale tronquée

$$f(x|\theta) \propto e^{-(x-\theta)^2/2} \mathbb{I}_{[\theta-c,\theta+c]}(x),$$

où c est connu.

3.19 *(Brown, 1986b) Montrer que toute famille exponentielle peut se reparamétriser en une famille exponentielle naturelle. Montrer aussi que la dimension de cette *reparamétrisation naturelle* ne dépend pas du choix de la reparamétrisation.

3.20 *(Dynkin, 1951) Montrer que les lois normales et les lois de la forme $c\log(y)$, où $y \sim \mathscr{G}(\alpha, \beta)$, sont les seules lois appartenant à la fois à une famille exponentielle et à une famille de position. En déduire que les lois normales sont les seules lois appartenant à une famille exponentielle et à symétrie sphérique (voir l'Exercice 1.1).

3.21 *(Lauritzen, 1996) Soient $X = (x_{ij})$ et $\Sigma = (\sigma_{ij})$ des matrices $m \times m$ symétriques définies positives. La *loi de Wishart*, $\mathscr{W}_m(\alpha, \Sigma)$, est définie par la densité

$$p_{\alpha,\Sigma}(X) = \frac{|X|^{\frac{\alpha-(m+1)}{2}} \exp(-\mathrm{tr}(\Sigma^{-1}X)/2)}{\Gamma_m(\alpha)|\Sigma|^{\alpha/2}},$$

où $\mathrm{tr}(A)$ est la trace de A et

$$\Gamma_m(\alpha) = 2^{\alpha m/2}\pi^{m(m-1)/4} \prod_{i=1}^{m} \Gamma\left(\frac{\alpha - i + 1}{2}\right).$$

a. Montrer que cette loi appartient à une famille exponentielle. Donner sa représentation naturelle et calculer l'espérance de $\mathscr{W}_m(\alpha, \Sigma)$.

b. Montrer que, si $z_1, \ldots, z_n \sim \mathscr{N}_m(0, \Sigma)$,

$$\sum_{i=1}^{n} z_i z_i' \sim \mathscr{W}_m(n, \Sigma).$$

c. Montrer que les moments de cette loi sont donnés par

$$\mathbb{E}[X|\alpha, \Sigma] = \alpha\Sigma, \qquad \mathrm{Cov}(X) = 2\alpha\Sigma \otimes \Sigma.$$

d. Montrer que l'espérance de l'inverse X^{-1} est

$$\mathbb{E}[X^{-1}|\alpha, \Sigma] = \frac{1}{\alpha - p - 1}\Sigma, \qquad \alpha > p + 1.$$

3.22 *(Pitman, 1936) Démontrer le *lemme de Pitman-Koopman* : *Si, pour $n \geq n_0$, il existe T_n de $\mathbb{R}^n$ dans $\mathbb{R}^k$ tel que $T_n(x_1, \ldots, x_n)$ est exhaustive pour $x_1, \ldots, x_n$ observations i.i.d. de $f(x|\theta)$, la distribution f appartient nécessairement à une famille exponentielle lorsque le support de f ne dépend pas de θ.* Étudier le cas où le support de f dépend de θ.

3.23 Montrer que la loi *gaussienne inverse*, de densité

$$(\pi)^{-1/2} z^{-3/2} \exp\{\theta_1 z + \theta_2(1/z) - (2\theta_1\theta_2)^{1/2} + (1/2)\log(-2\theta_2)\}$$

où $z \in \mathbb{R}_+$ et $\theta_1, \theta_2 \in \mathbb{R}_-$, est exponentielle mais non régulière.

3.24 *(Morris, 1982) Une *famille exponentielle restreinte* sur $\mathbb{R}$ est définie par

$$P_\theta(x \in A) = \int_A \exp\{\theta x - \psi(\theta)\}\, dF(x), \qquad \theta \in \Theta. \tag{3.19}$$

a. Montrer que, si $0 \in \Theta$, F est nécessairement une fonction de répartition. Si cette condition n'est pas vérifiée, montrer que la transformation de F en

$$dF_0(x) = \exp\{\theta_0 x - \psi(\theta)\}\, dF(x),$$

pour une valeur arbitraire $\theta_0 \in \Theta$ et le remplacement de θ par $\theta - \theta_0$ redonne le même résultat.

b. Montrer que, au sens restreint, $\mathscr{B}e(m\mu, m(1-\mu))$ et la loi log-normale $\mathscr{LN}(\alpha, \sigma^2)$ n'appartiennent pas à une famille exponentielle.

c. Si $\mu = \psi'(\theta)$ est l'espérance de la distribution (3.19), la *fonction de variance* de cette distribution est définie par $V(\mu) = \psi''(\theta) = \mathrm{var}_\theta(x)$. Montrer que V est effectivement une fonction de μ et que, de plus, si l'espace de variation de μ, Ω, est connu, le couple (V, Ω) caractérise complètement la famille (3.19) par

$$\psi\left(\int_{\mu_0}^{\mu} \frac{dm}{V(m)}\right) = \int_{\mu_0}^{\mu} \frac{m\, dm}{V(m)}.$$

(Noter que $\theta = \int_{\mu_0}^{\mu} dm/V(m)$.) Montrer que $V(\mu) = \mu^2$ définit *deux* familles, selon que $\Omega = \mathbb{R}^-$ ou $\Omega = \mathbb{R}^+$.

d. Montrer que $V(\mu) = \mu(1-\mu)/(m+1)$ correspond à la fois à la loi binomiale $\mathscr{B}(m, \mu)$ et à $\mathscr{B}e(m\mu, m(1-\mu))$. En déduire que la caractérisation par V n'est valide que pour les familles exponentielles naturelles.

e. Montrer que les familles exponentielles de *fonction de variance quadratique*, données par

$$V(\mu) = v_0 + v_1\mu + v_2\mu^2, \tag{3.20}$$

incluent les distributions suivantes : normale, $\mathscr{N}(\mu, \sigma^2)$, Poisson, $\mathscr{P}(\mu)$, gamma, $\mathscr{G}(r, \mu/r)$, binomiale, $\mathscr{B}(m, m\mu)$ et négative binomiale, $\mathscr{N}eg(r, p)$, qu'on peut définir comme le nombre de succès avant le r-ième échec, avec $\mu = rp/(1-p)$.

f. Montrer que les lois normales (respectivement, de Poisson) sont les seules distributions exponentielles naturelles de fonction de variance constante (respectivement, de degré un).

g. Supposons $v_2 \neq 0$ dans (3.20) et définissons $d = v_1^2 - 4v_0v_2$, le discriminant de (3.20), et $a = 1$ si $d = 0$, $a = \sqrt{dv_2}$ sinon. Montrer que $x^* = aV'(x)$ est une transformation linéaire de x, de fonction de variance

$$V^*(\mu^*) = s + v_2(\mu^*)^2, \tag{3.21}$$

où $\mu^* = aV'(\mu)$ et $s = -\mathrm{sign}(dv_2)$. Montrer qu'il est suffisant de considérer V^* pour caractériser les familles exponentielles naturelles de fonction de variance quadratique, au sens où les autres familles s'obtiennent par inversion de la transformation linéaire.

h. Montrer que (3.21) correspond à six cas possibles selon le signe de v_2 et la valeur de s $(-1, 0, 1)$. Éliminer les deux cas impossibles et identifier les familles données à la question e. ci-dessus. Montrer que le cas restant est

$v_2 > 0$, $s = 1$. Pour $v_2 = 1$, montrer que ce cas correspond à la distribution de $x = \log\{y/(1-y)\}/\pi$, où

$$y \sim \mathscr{B}e\left(\frac{1}{2} + \frac{\theta}{\pi}, \frac{1}{2} - \frac{\theta}{\pi}\right), \qquad |\theta| < \frac{\pi}{2},$$

et

$$f(x|\theta) = \frac{\exp[\theta x + \log(\cos(\theta))]}{2\cosh(\pi x/2)}. \tag{3.22}$$

[*Note :* La formule de réflexion $B(0.5 + t, 0.5 - t) = \pi/\cos(\pi t)$ peut être utile.] Les distributions générées par les transformations linéaires de (3.22) sont notées $\mathrm{GHS}(r, \lambda)$ (pour *generalized hyperbolic secant*), avec $\lambda = \tan(\theta)$, $r = 1/v_2$, et $\mu = r\lambda$. Montrer que la densité de $\mathrm{GHS}(r, \lambda)$ peut s'écrire

$$f_{r,\lambda}(x) = \left(1 + \lambda^2\right)^{-r/2} \exp\{x \arctan(\lambda)\} f_{r,0}(x)$$

(ne pas chercher à obtenir une expression explicite de $f_{r,0}$).

[*Note :* L'Exercice 10.33 exhibe d'autres propriétés des familles exponentielles à variance quadratique en termes de familles conjuguées et d'estimateurs de Bayes. L'Exercice 6.11 montre comment des polynômes orthogonaux peuvent être associés à chaque distribution d'une famille exponentielle à variance quadratique.]

3.25 Comparer les familles exponentielles usuelles avec les distributions (2.9) obtenues dans le Chapitre 2 et vérifier si elles génèrent des *estimateurs universels*.

3.26 Montrer que, pour toute famille exponentielle, l'espace naturel N est convexe.

3.27 Prouver la décomposition de l'Exemple 3.17
(i) directement ; et
(ii) via la représentation usuelle d'une distribution de Student.

3.28 Une alternative à la régression logistique introduite dans l'Exemple 3.21 est le modèle *probit*, tel que

$$P_\alpha(y_i = 1) = 1 - P_\alpha(y_i = 0) = \Phi(\alpha^t x_i), \qquad i = 1, \ldots, n,$$

où Φ est la fonction de répartition d'une loi normale centrée réduite.
a. Montrer que ce second modèle n'appartient pas à une famille exponentielle, même conditionnellement aux x_i.
b. Les observations y_i peuvent être considérées comme les fonctions indicatrices $\mathbb{I}_{z_i \le \alpha^t x_i}$ où z_i est une variable aléatoire non observée $\mathscr{N}(0, 1)$. Montrer que, si les z_i sont connus, la mesure de Lebesgue donne une loi a posteriori explicite. [*Note :* Le caractère intéressant de cette remarque apparaîtra plus clairement au Chapitre 6, car les *données manquantes* $z_1, \ldots, z_n$ peuvent être simulées.]

Section 3.3.4

3.29 Pour une distribution quelconque d'une famille exponentielle, déterminer des contraintes pour que la loi a priori d'entropie maximale soit aussi une loi conjuguée.

3.30 Un modèle de régression linéaire classique peut s'écrire $y \sim \mathscr{N}_p(X\beta, \sigma^2 I_p)$ où X est une matrice $p \times q$ et $\beta \in \mathbb{R}^q$. Lorsque X est connu, donner la paramétrisation naturelle de cette famille exponentielle et obtenir les lois a priori conjuguées sur (β, σ^2). Généraliser au cas $\mathscr{N}_p(X\beta, \Sigma)$, avec Σ connu.

3.31 Soit $x \sim \mathcal{N}(\theta, \theta)$ avec $\theta > 0$.
a. Déterminer l'a priori de Jeffreys $\pi^J(\theta)$.
b. Établir si la loi de x appartient à une famille exponentielle et construire les lois a priori conjuguées sur θ.
c. Utiliser la Proposition 3.20 pour relier les hyperparamètres des lois conjuguées à l'espérance de θ.

3.32 Montrer que, si $x \sim \mathcal{B}e(\theta_1, \theta_2)$, il existe des lois conjuguées pour $\theta = (\theta_1, \theta_2)$, mais que celles-ci ne permettent pas un calcul analytique des quantités a posteriori, à l'exception de $\mathbb{E}^\pi[\theta_1/(\theta_1 + \theta_2)|x]$, suivant la Proposition 3.20.

3.33 *(Robert, 1991) La distribution normale inverse généralisée $\mathcal{IN}(\alpha, \mu, \tau)$ a pour densité

$$K(\alpha, \mu, \tau)|\theta|^{-\alpha} \exp\left\{-(\frac{1}{\theta} - \mu)^2/2\tau^2\right\},$$

avec $\alpha > 0$, $\mu \in \mathbb{R}$, et $\tau > 0$.
a. Montrer que cette densité est bien définie et que la constante de normalisation est

$$K(\alpha, \mu, \tau)^{-1} = \tau^{\alpha-1} e^{-\mu^2/2\tau^2} 2^{(\alpha-1)/2}\ \Gamma(\frac{\alpha-1}{2})\ {}_1F_1\left(\frac{\alpha-1}{2}; 1/2; \frac{\mu^2}{2\tau^2}\right),$$

où ${}_1F_1$ est la *fonction confluente hypergéométrique* (voir Abramowitz et Stegun, 1964).
b. Montrer que cette distribution généralise celle de $y = 1/x$ pour $x \sim \mathcal{N}(\mu, \tau^2)$. Vérifier que la constante de normalisation ci-dessus est correcte dans ce cas particulier.
c. En déduire que l'espérance de $\mathcal{IN}(\alpha, \mu, \tau)$ existe pour $\alpha > 2$ et vaut

$$\mathbb{E}_{\alpha,\mu,\tau}[\theta] = \frac{\mu}{\tau^2}\ \frac{{}_1F_1(\frac{\alpha-1}{2}; 3/2; \mu^2/2\tau^2)}{{}_1F_1(\frac{\alpha-1}{2}; 1/2; \mu^2/2\tau^2)}.$$

d. Montrer que ces distributions $\mathcal{I}N(\alpha, \mu, \tau)$ constituent une famille conjuguée pour le modèle multiplicatif $\mathcal{N}(\theta, \theta^2)$.

3.34 Montrer que la distribution de Student $\mathcal{T}_p(\nu, \theta, \tau^2)$ n'admet pas de famille conjuguée autre que la famille triviale $\mathcal{F}_0$.

3.35 La Proposition 3.19 établit l'existence d'une famille conjuguée pour toute famille exponentielle, de la forme (3.8),

$$\pi(\theta|\lambda, \mu) = \exp\{\theta \cdot \mu - \lambda\psi(\theta)\}K(\mu, \lambda).$$

a. Montrer que la distribution (3.8) est en fait bien définie pour $\lambda > 0$ et $(\mu/\lambda) \in \mathring{N}$, intérieur de N.
b. Calculer cette constante K pour des distributions normale, gamma et négative binomiale.
c. En déduire (en recourant à une certaine reparamétrisation) que la fonction de vraisemblance $\ell(\theta|x)$ est une distribution a priori particulière pour les familles exponentielles et donner l'a priori correspondant pour les familles ci-dessus.
d. Cette propriété caractérise-t-elle les familles exponentielles ? Donner un contre-exemple.

3.36 * Démontrer la Proposition 3.20 et sa réciproque dans le cas continu. Appliquer aux distributions du Tableau 3.4.

3.37 Montrer que les distributions du Tableau 3.4 sont en fait conjuguées
(i) directement ; et
(ii) en utilisant la Proposition 3.20.

3.38 Soit $x \sim \mathcal{G}(\theta, \beta)$, c'est-à-dire $f_\beta(x|\theta) = \frac{\beta^\theta}{\Gamma(\theta)} x^{\theta-1} e^{-\beta x}$.

a. Peut-on construire une famille conjuguée pour cette distribution ?

b. Traiter le cas $\theta \in \mathbb{N}$.

c. Même question pour $x \sim \mathcal{B}e(1, \theta)$.

3.39 Montrer que, pour des familles exponentielles, un accroissement du nombre de niveaux hiérarchiques ne modifie pas la nature conjuguée de l'a priori résultant si des lois conjuguées avec des paramètres d'échelle constants sont utilisées à tous les niveaux de la hiérarchie. (Considérer par exemple le cas normal.)

3.40 *(Robert, 1993b) Soit $f(x|\theta)$ prise dans une famille exponentielle,

$$f(x|\theta) = e^{\theta \cdot x - \psi(\theta)} h(x), \qquad x \in \mathbb{R}^k,$$

et $\pi_0(\theta|x_0, \lambda)$ une loi a priori conjuguée,

$$\pi_0(\theta|x_0, \lambda) = e^{\theta \cdot x_0 - \lambda\psi(\theta)}.$$

Nous cherchons à obtenir une estimation dite objective de $\nabla\psi(\theta)$, à partir d'une loi a priori arbitraire $\pi_0(\theta|x_0, \lambda)$. Dans ce but, nous remplaçons π_0 par la distribution $\pi_1(\theta|x_1, \lambda)$ définie par la relation

$$\mathbb{E}^{\pi_1}[\nabla\psi(\theta)] = \mathbb{E}^{\pi_0}[\nabla\psi(\theta)|x], \tag{3.23}$$

afin de réduire l'influence de x_0.

a. En déduire la relation entre x_1 et x_0.

b. Nous itérons le processus d'actualisation (3.23) afin d'éliminer, autant que possible, l'influence de x_0 et nous construisons de cette façon une suite $\pi_n(\theta|x_n, \lambda)$ de lois a priori conjuguées. Donner la relation entre x_n et x_{n-1} et en déduire la limite de la suite (x_n).

c. Donner la limite correspondante des estimateurs de Bayes de $\nabla\psi(\theta)$. Comment caractérisez-vous l'estimateur résultant ? S'agit-il toujours d'un estimateur de Bayes ?

d. Dans le cas particulier où $x \sim \mathcal{N}(\theta, 1)$, le paramètre d'intérêt est $h(\theta) = e^{-\theta}$. Donner l'estimateur $h(\theta)$ obtenu de cette façon, en utilisant la formule d'actualisation itérative

$$\mathbb{E}^{\pi_n}[h(\theta)] = \mathbb{E}^{\pi_{n-1}}[h(\theta)|x].$$

e. Considérer le cas $x \sim \mathcal{G}(\alpha, \theta)$ et $h(\theta) = \theta^k$ afin de montrer que cette méthode itérative, appelée *rétroaction d'a priori*, ne converge pas toujours vers l'estimateur du maximum de vraisemblance.

f. Montrer que la limite de cet estimateur obtenu par rétroaction d'a priori lorsque λ tend vers $+\infty$ est l'estimateur du maximum de vraisemblance de $h(\theta)$, pour une fonction arbitraire h et toute famille exponentielle.

Section 3.4

3.41 Dans le cadre de l'Exemple 3.22, construire une loi a priori en observant quelques pièces et en imposant un mélange de lois bêta, comme dans Diaconis et Ylvisaker (1985). Choisir l'une de ces pièces et calculer la distribution a posteriori de θ, la probabilité d'obtenir pile, après dix lancers et cinquante lancers.

3.42 Déduire les lois a posteriori de l'Exemple 3.22 de la relation de récurrence $\Gamma(a+1) = a\Gamma(a)$ sur la fonction gamma.

3.43 Soient $x \sim \mathcal{N}(0,1)$ et $\theta \sim \mathcal{T}_1(5,0,1)$.

a. Établir une méthode d'approximation de la loi a priori par un mélange de :
(i) deux lois normales ; et
(ii) cinq lois normales.

b. Dans chaque cas, donner l'approximation de l'espérance a posteriori de θ correspondante pour $x = 1$, et comparer avec la valeur exacte.

Section 3.5.1

3.44 Soit $x_1, \ldots, x_n \sim \mathcal{N}(\mu+\nu, \sigma^2)$, avec $\pi(\mu,\nu,\sigma) \propto 1/\sigma$.

a. Montrer que la distribution a posteriori n'est pas définie pour tout n.

b. Étendre ce résultat aux modèles surparamétrisés avec des lois a priori impropres.

Les exercices suivants (3.45-3.51) *traitent du paradoxe de marginalisation à travers plusieurs exemples et démontrent que celui-ci ne peut avoir lieu qu'avec des lois a priori impropres. Dawid* et al. *(1973), Stone (1976) et Jaynes (1980) proposent des solutions partielles à ce paradoxe. Notons qu'une explication fondamentale est que la loi a priori impropre* $\pi(d\eta, d\theta) = \pi(\eta)\, d\eta\, d\theta$ *ne correspond pas à la loi pseudo-marginale* $\pi(d\eta) = \pi(\eta)\, d\eta$.

3.45 *(Dawid *et al.*, 1973) Soient n variables aléatoires $x_1, \ldots, x_n$, telles que les ξ premières d'entre elles suivent la loi $\mathcal{E}xp(\eta)$ et les $n - \xi$ restantes suivent $\mathcal{E}xp(c\eta)$, où c est une constante connue et ξ prend ses valeurs dans $\{1, 2, \ldots, n-1\}$.

a. Donner la forme de la distribution a posteriori de ξ lorsque $\pi(\xi,\eta) = \pi(\xi)$ et montrer qu'elle ne dépend que de $z = (z_2, \ldots, z_n)$, avec $z_i = x_i/x_1$.

b. Montrer que la distribution de z, $f(z|\xi)$, ne dépend que de ξ.

c. Montrer que la loi a posteriori $\pi(\xi|x)$ ne peut pas s'écrire comme une loi a posteriori pour $z \sim f(z|\xi)$, quelle que soit $\pi(\xi)$, bien qu'elle ne dépende que de z. Comment expliquez-vous ceci ?

d. Montrer que ce paradoxe n'a pas lieu lorsque $\pi(\xi,\eta) = \pi(\xi)\eta^{-1}$.

3.46 *(Dawid *et al.*, 1973) Soient u_1, u_2, s^2 tels que

$$u_1 \sim \mathcal{N}(\mu_1, \sigma^2), \qquad u_2 \sim \mathcal{N}(\mu_2, \sigma^2), \qquad s^2 \sim \sigma^2 \chi^2_\nu/\nu,$$

et $\zeta = (\mu_1 - \mu_2)/(\sigma\sqrt{2})$ est le paramètre d'intérêt. La loi a priori est

$$\pi(\mu_1, \mu_2, \sigma) = \frac{1}{\sigma}.$$

a. Montrer que la loi a posteriori $\pi(\zeta|x)$ ne dépend que de

$$z = \frac{u_1 - u_2}{s\sqrt{2}}.$$

b. Montrer que la distribution de z ne dépend que de ζ, mais que pourtant un paradoxe apparaît ; il est impossible de calculer $\pi(\zeta|x)$ à partir de $f(z|\zeta)$, même si $\pi(\zeta|x)$ ne dépend que de z.

c. Montrer que ce paradoxe disparaît lorsque

$$\pi(\mu_1, \mu_2, \sigma) = \frac{1}{\sigma^2}.$$

3.47 *(Dawid *et al.*, 1973) Soient

$$x_{11}, \ldots, x_{1n} \sim \mathcal{N}(\mu_1, \sigma^2),$$
$$x_{21}, \ldots, x_{2n} \sim \mathcal{N}(\mu_2, \sigma^2),$$

$2n$ variables aléatoires indépendantes.

a. Le paramètre d'intérêt est $\xi = (\xi_1, \xi_2) = (\mu_1/\sigma, \mu_2/\sigma)$ et la loi a priori est

$$\pi(\mu_1, \mu_2, \sigma) = \sigma^{-p}.$$

Montrer que $\pi(\xi|x)$ ne dépend que de $z = (z_1, z_2) = (\bar{x}_1/s, \bar{x}_2/s)$ et que la loi de z ne dépend que de ξ. Calculer la valeur de p qui évite ce paradoxe.

b. Le paramètre d'intérêt est désormais $\zeta = \xi_1$. Montrer que $\pi(\zeta|x)$ ne dépend que de z_1 et que $f(z_1|\xi)$ ne dépend que de ζ. Donner la valeur de p qui évite ce paradoxe.

c. Mêmes questions pour $\sigma \sim \mathcal{P}a(\alpha, \sigma_0)$.

3.48 *(Dawid *et al.*, 1973) Soient (x_1, x_2) distribués selon :

$$f(x_1, x_2|\theta) \propto \int_0^{+\infty} t^{2n-1} \exp\left[-\frac{1}{2}\left\{t^2 + n(x_1 t - \zeta)^2 + n(x_2 t - \xi)^2\right\}\right] dt,$$

avec $\theta = (\zeta, \xi)$. Justifier cette distribution en recourant au cadre de l'Exercice 3.47. La loi a priori sur θ est $\pi(\theta) = 1$.

a. Montrer que $\pi(\zeta|x)$ ne dépend que de x_1 et que $f(x_1|\theta)$ ne dépend que de ζ, mais que $\pi(\zeta|x)$ ne peut pas être déduite de $x_1 \sim f(x_1|\zeta)$.

b. Montrer que, pour toute loi $\pi(\theta)$ telle que $\pi(\zeta|x)$ ne dépend que de x_1, $\pi(\zeta|x)$ n'est pas proportionnelle à $\pi(\zeta) f(x_1|\zeta)$.

3.49 *(Jaynes, 1980) Dans le cadre de l'Exercice 3.45, prendre $\pi(\xi, \eta) = \pi(\xi)\pi(\eta)$.

a. Montrer que

$$\pi(\xi|x) \propto \pi(\xi) c^{-\xi} \int_0^{+\infty} \eta^{-n} \exp(-\eta x_1 Q) \pi(\eta)\, d\eta,$$

où

$$Q = \sum_{i=1}^{\xi} z_i + c \sum_{\xi+1}^{n} z_i.$$

b. Déterminer si le paradoxe a lieu pour $\pi(\eta) = \eta^{-k}$ $(k > -n - 1)$.

c. Même question pour $\eta \sim \mathcal{P}a(\alpha, \eta_0)$.

3.50 *(Jaynes, 1980) Soit

$$f(y, z|\eta, \zeta) \propto \frac{\zeta^z \eta^y (1-\eta)^{z-y}}{y!(z-y)!} \qquad (0 \leq y \leq z),$$

avec $0 < \eta < 1$.

a. Montrer que $f(z|\eta,\zeta)$ ne dépend que de ζ et calculer la distribution $f(y, z|\eta,\zeta)$ à partir de $f(y|z,\eta,\zeta)$.
b. Montrer que le paradoxe n'a lieu pour aucun $\pi(\eta)$.

3.51 *(Dawid *et al.*, 1973) Soient $x=(y,z)$ de loi $f(x|\theta)$ et $\theta=(\eta,\xi)$. Supposons que $\pi(\xi|x)$ ne dépende que de z et $f(z|\theta)$ que de ξ.
a. Montrer que le paradoxe est évité lorsque $\pi(\theta)$ est une loi propre.
b. Généraliser au cas où $\int \pi(\eta,\xi)\,d\eta = \pi(\xi)$ et déterminer si le paradoxe est ainsi évité.

Section 3.5.3

3.52 Reprenant l'Exemple 3.32 et pour $x \sim \mathscr{B}(n,p)$, trouver une loi a priori sur n telle que $\pi(n|x)$ soit $\mathscr{N}eg(x,p)$.

3.53 * Reprenant l'Exemple 3.34,

a. Montrer que l'estimateur de Bayes de $\eta = ||\theta||^2$ sous un coût quadratique pour $\pi(\eta) = 1/\sqrt{\eta}$ et $x \sim \mathscr{N}(\theta, I_p)$ peut s'écrire

$$\delta^\pi(x) = \frac{{}_1F_1(3/2; p/2; ||x||^2/2)}{{}_1F_1(1/2; p/2; ||x||^2/2)},$$

où ${}_1F_1$ est la fonction confluente hypergéométrique.

b. Déduire du développement limité de ${}_1F_1$ le développement asymptotique de δ^π (pour $||x||^2 \to +\infty$).

c. Comparer δ^π avec δ_0 tel que $\delta_0(x) = ||x||^2 - p$.

d. Étudier le comportement de ces estimateurs sous un coût quadratique pondéré

$$\mathrm{L}(\delta,\theta) = \frac{(||\theta||^2 - \delta)^2}{2||\theta||^2 + p}$$

et conclure.

3.54 Trouver une transformation de θ, $\eta = g(\theta)$, telle que l'information de Fisher $I(\eta)$ soit constante pour :
(i) une loi de Poisson, $\mathscr{P}(\theta)$;
(ii) une loi gamma, $\mathscr{G}(\alpha,\theta)$, avec $\alpha = 1, 2, 3$; et
(iii) une loi binomiale, $\mathscr{B}(n,\theta)$.

3.55 En supposant que $\pi(\theta) = 1$ soit une loi a priori acceptable pour des paramètres réels, montrer que cette loi générale correspond à $\pi(\sigma) = 1/\sigma$ si $\sigma \in \mathbb{R}^+$ et à $\pi(\varrho) = 1/\varrho(1-\varrho)$ si $\varrho \in [0,1]$, pour les transformations naturelles $\theta = \log(\sigma)$ et $\theta = \log(\varrho/(1-\varrho))$.

3.56 *(Saxena et Alam, 1982) Dans un cadre identique à celui de l'Exercice 3.53 :

a. Donner l'estimateur du maximum de vraisemblance de $||\theta||^2$ lorsque $x \sim \mathscr{N}(\theta, I_p)$.

b. Montrer que l'estimateur du maximum de vraisemblance obtenu à partir de $z = ||x||^2$ vérifie l'équation implicite

$$1 = \frac{z}{\sqrt{\lambda z}} \frac{I_{p/2}(\sqrt{\lambda z})}{I_{(p-1)/2}(\sqrt{\lambda z})} \qquad (z > p),$$

où I_ν est la *fonction modifiée de Bessel* (voir Abramowitz et Stegun, 1964 , ou l'Exercice 4.36).

c. Utiliser un développement limité de I_ν pour montrer que l'estimateur du maximum de vraisemblance $\hat{\lambda}$ vérifie

$$\hat{\lambda}(z) = z - p + 0.5 + O(1/z).$$

d. Montrer que $\hat{\lambda}$ est dominé par $(z-p)^+$ sous un coût quadratique.

3.57 L'information de Fisher n'est pas définie lorsque le support de $f(x|\theta)$ dépend de θ. Considérer les cas suivants :

$$\text{(i) } x \sim \mathscr{U}_{[-\theta,\theta]}; \quad \text{(ii) } x \sim \mathscr{P}a(\alpha,\theta); \quad \text{(iii) } f(x|\theta) \propto e^{-(x-\theta)^2/2}\mathbb{I}_{[0,\theta]}(x).$$

3.58 Montrer qu'une approximation du second ordre des coûts d'entropie et de Hellinger introduits dans la Section 2.5.4 est $(\theta-\delta)^2 I(\theta)$. Ce résultat est-il une justification supplémentaire pour utiliser la loi a priori de Jeffreys ?

3.59 Soit $x \sim \mathscr{P}(\theta)$.

a. Déterminer l'a priori de Jeffreys π^{J} et évaluer si l'a priori invariant par transformation d'échelle $\pi_0(\theta) = 1/\theta$ est préférable.

b. Donner la loi a priori d'entropie maximale pour la mesure de référence π_0^{J} et les contraintes $\mathbb{E}^\pi[\theta] = 1$, $\mathrm{var}^\pi(\theta) = 1$. Que se passe-t-il si on remplace π par π_0 ?

c. En fait, x est le nombre de voitures traversant une voie ferrée pendant une durée T. Montrer que x est distribué selon une loi de Poisson $\mathscr{P}(\theta)$ si la durée entre deux arrivées est distribuée selon $\mathscr{E}xp(\lambda)$; noter que $\theta = \lambda T$.

d. Justifier l'utilisation de π_0 à l'aide de la construction de la loi Poisson établie ci-dessus.

Section 3.5.4

3.60 Pour $x \sim \mathscr{N}(\theta, \sigma^2)$, donner la loi a priori de référence pour les ordres $\{\theta, \sigma\}$ et $\{\sigma, \theta\}$.

3.61 Soient $\theta \in [a, b]$ et $\pi(\theta) \propto 1/\theta$.

a. Déterminer la constante de normalisation de π.

b. Calculer $p_i = \pi(i \leq \theta < i+1)$ pour $a \leq i \leq b-1$.

c. En déduire la limite de p_i lorsque a tend vers 0 ou b tend vers ∞. [*Note :* Cet exercice est relié au *problème des entrées de tableau*, c'est-à-dire au fait que dans beaucoup de tableaux numériques la fréquence du premier chiffre significatif est $\log_{10}(1 + i^{-1})$ $(1 \leq i \leq 9)$. Voir Berger 1985b, p. 86, pour une présentation détaillée.]

3.62 *(Kass et Wasserman, 1996) Montrer que l'a priori de référence obtenu à partir de l'a priori de Jeffreys pour θ_1 fixé, $\pi(\theta_2|\theta_1)$, et de l'a priori de Jeffreys pour la loi marginale (3.14) peut aussi s'écrire

$$\pi(\theta_1, \theta_2) \propto \pi(\theta_2|\theta_1) \exp\left\{ \int \pi(\theta_2|\theta_1) \log \sqrt{|\mathbf{I}|/|\mathbf{I}_{22}|} d\theta_2 \right\},$$

où $\mathbf{I}$ est l'information de Fisher et $\mathbf{I}_{22}$ est la composante de $\mathbf{I}$ associée à θ_2.

Section 3.6

3.63 *(Berger, 1990) Soit $\Gamma_{\epsilon,\mathscr{Q}}$ la classe de lois définie en Section 3.6 (iii), avec

$$\mathscr{Q} = \{ \text{ distributions unimodales symétriques en } \theta_0 \}.$$

Lorsque π varie dans $\mathscr{Q}$, la loi marginale

$$m(\pi) = \int f(x|\theta)\pi(\theta)\,d\theta$$

varie entre des bornes supérieure et inférieure m^U et m^L.

a. Montrer que toute distribution unimodale symétrique en θ_0 peut s'écrire comme un mélange de distributions uniformes symétrique en θ_0, $\mathscr{U}_{[\theta_0-a,\theta_0+a]}$.

b. En déduire que

$$m^U = \sup_{\pi\in\Gamma_\epsilon,\mathscr{Q}} m(\pi) = (1-\epsilon)m(\pi_0) + \epsilon \sup_{z>0} \int_{\theta_0-z}^{\theta_0+z} \frac{f(x|\theta)}{2z}\,d\theta.$$

c. Si la quantité d'intérêt est le *facteur de Bayes*,

$$B(\pi) = \frac{f(x|\theta_0)}{\int_{\theta\neq\theta_0} f(x|\theta)\pi_1(\theta)\,d\theta},$$

où π_1 est la loi π conditionnée par $\theta \neq \theta_0$ et π_0 est la masse de Dirac en θ_0, montrer que

$$B^L = \inf_{\pi\in\Gamma_\epsilon,\mathscr{Q}} B(\pi) = \frac{f(x|\theta)}{\epsilon \sup_z \int_{\theta_0-z}^{\theta_0+z} (f(x|\theta)/2z)\,d\theta}.$$

3.64 Soit la classe des lois a priori

$$\Gamma = \{\mathscr{N}(\mu,\tau^2),\ 0 \le \mu \le 2,\ 2 \le \tau^2 \le 4\}$$

avec $x \sim \mathscr{N}(\theta, 1)$.

a. Étudier les variations de $\mathbb{E}^\pi[\theta|x]$ et $\mathrm{var}^\pi(\theta|x)$ pour $\pi \in \Gamma$.

b. Étudier $\varrho(\pi, \delta^{\pi'})$ pour $\pi, \pi' \in \Gamma$ et $\delta^\pi(x) = \mathbb{E}^\pi[\theta|x]$, $\mathrm{L}(\theta,\delta) = (\theta-\delta)^2$ afin de déterminer l'estimateur minimax pour la classe Γ.

3.65 *(Walley, 1991) Supposons que, au lieu de définir une loi a priori π sur la σ-algèbre de Θ, on définisse des bornes supérieure et inférieure pour π, notées $\overline{\pi}$ et $\underline{\pi}$. Pour tout événement A, $\underline{\pi}(A)$ représente la somme maximale qu'on est prêt à parier pour obtenir une unité monétaire si A a lieu. De même, $1-\overline{\pi}(A)$ est la somme minimale qu'on est prêt à parier que A n'ait pas lieu.

a. Montrer que, si la loi a priori π est connue, $\underline{\pi} = \pi = \overline{\pi}$.

b. Montrer qu'on doit imposer $\underline{\pi}(A) + \underline{\pi}(A^c) \le 1 \le \overline{\pi}(A) + \overline{\pi}(A^c)$ pour tout A pour éviter une *perte certaine*.

c. Si $\underline{\pi}(A\cup B)$ est la somme maximale qu'on est prêt à parier sur $A\cup B$, montrer que $\underline{\pi}(A\cup B) \ge \underline{\pi}(A) + \underline{\pi}(B)$ et, de même, que $\overline{\pi}(A\cup B) \le \overline{\pi}(A) + \overline{\pi}(B)$.

3.66 ***(Suite de l'Exercice 3.65)** Si on considère plutôt des *paris*, c'est-à-dire des fonctions X à valeurs réelles définies sur un espace mesurable Ω correspondant à des récompenses variables, dépendant de l'état d'incertitude $\omega \in \Omega$, il est alors aussi possible de définir *des prévisions supérieure et inférieure*, $\overline{P}$ et $\underline{P}$, où $\underline{P}(X)$ est le prix maximal acceptable pour la récompense X et $\overline{P}(X)$ le prix de vente minimal.

a. Un pari est *désirable* s'il est possible que quelqu'un le contracte. Justifier les axiomes suivants :

(A) Si $\sup_\omega X(\omega) < 0$, alors X n'est pas désirable ;

(B) Si $\inf_\omega X(\omega) > 0$, alors X est désirable ;

(C) Si X est désirable et $\lambda > 0$, alors λX est désirable ; et

(D) Si X et Y sont tous les deux désirables, alors $X + Y$ est désirable.

b. Justifier les axiomes de *cohérence* suivants sur $\underline{P}$ et montrer qu'ils correspondent aux axiomes (B), (C) et (D) ci-dessus :
(P_1) $\underline{P}(X) \geq \inf_\omega X(\omega)$;
(P_2) $\underline{P}(\lambda X) = \lambda \underline{P}(X)$; et
(P_3) $\underline{P}(X+Y) \geq \underline{P}(X) + \underline{P}(Y)$.

c. Pour une prévision inférieure $\underline{P}$ donnée, la *prévision supérieure conjuguée* est définie par $\overline{P}(X) = -\underline{P}(-X)$. Montrer que, si $\underline{P}$ est cohérente et $\overline{P}$ est la conjuguée de $\underline{P}$, celles-ci satisfont

$$\inf_\omega X(\omega) \leq \underline{P}(X) \leq \overline{P}(X) \leq \sup_\omega X(\omega)\,,$$

et en déduire que $\overline{P}$ est une fonction convexe.

d. Montrer que, lorsque $\underline{P}$ est *autoconjuguée*, alors $\underline{P}(X) = \overline{P}(X)$ et vérifie les contraintes de linéarité suivantes :

$$\underline{P}(X+Y) = \underline{P}(X) + \underline{P})Y) \qquad \text{et} \qquad \underline{P}(\lambda X) = \lambda \underline{P}(X), \quad \lambda \in \mathbb{R}.$$

3.67 ***(Suite de l'Exercice 3.66)** On dit qu'une prévision inférieure $\underline{P}$ *évite une perte certaine* si, pour tout $n \geq 1$ et tout ensemble de paris $X_1, \ldots, X_n$,

$$\sup_\omega \sum_{i=1}^n X_i - \underline{P}(X_i) \geq 0.$$

a. Montrer que $\underline{P}$ évite une perte certaine si et seulement si

$$\sup_\omega \sum_{i=1}^n \lambda_i (X_i - \underline{P}(X_i)) \geq 0$$

pour tout $n \geq 1$, tout ensemble de paris $X_1, \ldots, X_n$ et tout $\lambda_i \geq 0$.

b. Sous l'hypothèse que $\underline{P}$ évite une perte certaine, montrer que, pour tout $\lambda \geq 0$,

$$\begin{aligned} &\underline{P}(\lambda X) \leq \lambda \overline{P}(X), \qquad \overline{P}(\lambda X) \geq \lambda \underline{P}(X), \\ &\underline{P}(\lambda X + (1-\lambda)Y) \leq \lambda \overline{P}(X) + (1-\lambda)\overline{P}(Y). \end{aligned}$$

où $\overline{P}$ est la prévision supérieure conjuguée.

c. Une prévision inférieure est *cohérente* si

$$\sup_\omega \left[\sum_{i=1}^n (X_i - \underline{P}(X_i)) - m(X_0 - \underline{P}(X_0)) \right] \geq 0$$

pour tout m, n et tout ensemble de paris $X_0, \ldots, X_n$. Montrer que $\underline{P}$ est cohérente si et seulement si elle satisfait les axiomes (P_1), (P_2), et (P_3).

d. Montrer que la linéarité est équivalente à la cohérence plus l'autoconjugaison, si on définit la *linéarité* comme

$$\sup_\omega \left\{ \sum_{i=1}^n X_i(\omega) - \sum_{j=1}^m Y_i(\omega) \right\} \geq \sum_{i=1}^n \underline{P}(X_i) - \sum_{j=1}^m \underline{P}(Y_i)$$

pour tout m, n et tout ensemble de paris $X_1, \ldots, X_n, Y_1, \ldots, Y_m$.

e. Montrer que $\underline{P}$ est une prévision linéaire si et seulement si $\underline{P}(X+Y) = \underline{P}(X) + \underline{P}(Y)$ et $\underline{P}(X) \geq \inf_\omega X(\omega)$. En déduire que $\underline{P}$ est une prévision linéaire si et seulement si elle satisfait la condition de linéarité, (P_2), et
(P_4) si $X \geq 0$, alors $\underline{P}(X) \geq 0$; et
(P_5) $\underline{P}(1) = 1$.

Note 3.8.3

3.68 Appliquer la décomposition de Dalal et Hall (1983) aux cas suivants :
(i) $x \sim \mathcal{N}(\theta, I_p)$, $\theta \sim \mathcal{T}_p(m, 0, \tau^2)$; et
(ii) $x \sim \mathcal{N}eg(N, p)$, $p/(1-p) \sim \mathcal{G}(1/2, 1/2)$.

3.69 *Trouver les mesures naturelles ν_m de Dalal et Hall (1983) pour les lois du Tableau 3.6.

3.8 Notes

3.8.1 Construction axiomatique de lois a priori

Pour démontrer l'existence d'une loi a priori, nous avons besoin, à l'instar de la fonction d'utilité (voir la Section 2.2), de nous fonder sur un ordre des événements (plutôt que des récompenses). Supposons donc que le décideur, le client ou le statisticien soient à même de déterminer une relation d'ordre sur une σ-algèbre $\mathscr{B}(\Theta)$. Cette relation, notée $\preceq$, est telle que $B \prec A$ signifie que A est plus vraisemblable que B, $B \preceq A$, que A est au moins aussi vraisemblable que B, et $B \sim A$, que A et B sont aussi vraisemblables l'un que l'autre. Bien entendu, s'il existe une distribution P de probabilité sur $(\Theta, \mathscr{B}(\Theta))$, P induit directement une relation d'ordre sur $\mathscr{B}(\Theta)$. Nous considérons ci-dessous des hypothèses sous lesquelles la réciproque peut être établie. Une première hypothèse est que la relation d'ordre est *totale* :

(A_1) Pour tout ensemble mesurable A et B, une et seulement une des relations suivantes est satisfaite :

$$A \prec B, \quad B \prec A \quad \text{ou} \quad A \sim B.$$

Une autre hypothèse est :

(A_2) Si A_1, A_2, B_1, B_2 sont quatre ensembles mesurables vérifiant $A_1 \cap A_2 = B_1 \cap B_2 = \emptyset$ et $A_i \preceq B_i$ $(i = 1, 2)$, alors $A_1 \cup A_2 \preceq B_1 \cup B_2$. De plus, si $A_1 \prec B_1$, $A_1 \cup A_2 \prec B_1 \cup B_2$.

Cette hypothèse naturelle entraîne *la transitivité* de la relation d'ordre. L'hypothèse suivante empêche l'existence d'ensembles mesurables de vraisemblance négative (donc moins vraisemblables que l'ensemble vide) :

(A_3) Pour tout événement A, $\emptyset \preceq A$ et $\emptyset \prec \Theta$.

La condition supplémentaire $\emptyset \prec \Theta$ évite le cas trivial où tous les événements sont équivalents. Il est aussi nécessaire de permettre la comparaison d'une suite infinie d'événements.

(A_4) Si $A_1 \supset A_2 \supset \cdots$ est une suite décroissante d'ensembles mesurables et B est un événement donné tel que $B \preceq A_i$ pour tout i, alors

$$B \preceq \bigcap_{i=1}^{+\infty} A_i.$$

Cette hypothèse assure en quelque sorte la continuité de l'ordre des préférences et est reliée à la propriété de σ-additivité des mesures de probabilité. Cependant, les axiomes (A_1)–(A_4) sont insuffisants pour obtenir l'existence d'une distribution de probabilité à partir de l'ordre des vraisemblances. En fait, passer d'une échelle de comparaison qualitative à une comparaison quantitative requiert une dernière hypothèse.

(A_5) Il existe une variable aléatoire X sur $(\Theta, \mathscr{B}(\Theta))$ de *distribution uniforme* sur $[0,1]$, c'est-à-dire telle que, pour tout I_1, I_2, intervalles de $[0,1]$,

$$\{X \in I_1\} \preceq \{X \in I_2\}$$

si et seulement si

$$\lambda(I_1) \leq \lambda(I_2),$$

où λ est la mesure de Lebesgue.

Cette hypothèse supplémentaire permet alors d'établir le résultat d'existence suivant (voir DeGroot, 1970, pour une démonstration).

Théorème 3.40. *Sous les axiomes* (A_1)–(A_5)*, il existe une distribution P telle que $P(A) \leq P(B)$ si et seulement si $A \preceq B$.*

Comparés à l'obtention d'une fonction d'utilité dans le Chapitre 2, les développements précédents sur les fondations axiomatiques de la loi a priori sont plus limités. Une première raison est que les hypothèses ci-dessus et le cadre formel correspondant sont plus difficiles à justifier. En fait, le fait qu'un statisticien soit à même d'exprimer la *vraisemblance* d'un événement signifie qu'il a, consciemment ou pas, construit un modèle probabiliste sous-jacent et, donc, que la construction précédente est en quelque sorte tautologique. L'hypothèse (A_5) est particulièrement forte et peut rarement être vérifiée en pratique. Notez cependant que, jusqu'à un certain point, la même critique peut être faite à l'égard de la construction de la fonction d'utilité.

Une seconde raison de cette limitation est plus terre à terre. Selon le Théorème 3.40, le décideur peut construire une loi a priori à partir de son ordre des vraisemblances. Cependant, il est très vraisemblable, surtout si Θ n'est pas fini, que cet ordre sera *grossier*, c'est-à-dire que la σ-algèbre $\mathscr{B}(\Theta)$ correspondante ne sera pas la σ-algèbre borélienne usuelle sur Θ, empêchant par là même l'utilisation des distributions classiques sur θ. Cependant, il est rassurant de pouvoir justifier l'utilisation d'une loi a priori par d'autres raisonnements que ceux de l'approche fréquentiste, supposant la répétabilité des expériences, même si cela est d'un intérêt limité en pratique.

3.8.2 Échangeabilité et lois a priori conjuguées

Bernardo et Smith (1994, Section 4.3) justifient partiellement l'existence de lois a priori par la notion d'*échangeabilité* :

Définition 3.41. *Une suite $(x_1, \ldots, x_n)$ de variables aléatoires est* finiment échangeable *si la distribution jointe $p(x_1, \ldots, x_n)$ est invariante par toute permutation d'indices des variables aléatoires, c'est-à-dire*

$$p(x_1, \ldots, x_n) = p(x_{(1)}, \ldots, x_{(n)}),$$

Une suite infinie $(x_n)_n$ *est* infiniment échangeable *si toute suite extraite finie est finiment échangeable.*

Bien que l'hypothèse d'échangeabilité ne soit pas toujours raisonnable (voir Bernardo et Smith, 1994, Section 4.2.2, pour des exemples), il existe beaucoup de situations pour lesquelles l'ordre dans lequel les données ont été obtenues n'a effectivement pas d'importance. Les conséquences de cette hypothèse d'infinie échangeabilité sur l'existence de lois a priori sont de plus tout à fait intéressantes. Par exemple, si $(x_n)_n$ est une suite infinie de variables aléatoires prenant valeurs dans $\{0,1\}$, de Finetti (1972) a démontré qu'il existe une mesure de probabilité $\pi(\theta)$ telle que, pour tout n, la loi jointe de $(x_1,\ldots,x_n)$ puisse s'écrire

$$p(x_1,\ldots,x_n) = \int_0^1 \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} d\pi(\theta),$$

c'est-à-dire que, conditionnellement à θ, les x_i sont des variables aléatoires i.i.d. de Bernoulli $\mathscr{B}(\theta)$. Comme l'ont montré Bernardo et Smith (1994, Section 4.3.2), cette propriété s'étend aux variables aléatoires prenant leurs valeurs dans un ensemble fini, disons $\{1,2,\ldots,k\}$, celles-ci étant alors multinomiales, conditionnellement au vecteur $\theta = (\theta_1,\ldots,\theta_k)$.
Dans le cas général où les x_i sont à valeurs réelles et infiniment échangeables, une représentation intéressante est aussi disponible, de la forme

$$p(x_1,\ldots,x_n) = \int \prod_{i=1}^n F(x_i) d\pi(F)\,,$$

où F est une fonction de répartition et π est une mesure de probabilité sur l'espace des fonctions de distribution (voir Chow et Teicher, 1988, pour une formulation plus précise de ce résultat, dont les aspects les plus subtils touchant à la théorie de la mesure dépassent le cadre de ce livre). Cette représentation est intrinsèquement non paramétrique (voir la Note 1.8.2), mais Bernardo et Smith (1994, Section 4.6) traitent d'autres notions d'échangeabilité qui permettent de revenir à un cadre paramétrique.

3.8.3 Approximation par des mélanges continus de lois a priori conjuguées

Soit une densité prise dans une famille exponentielle et écrite sous la forme

$$f(x|\theta) = \exp\{x\cdot\tau(\theta) - \gamma(\theta)\},$$

avec $\mathbb{E}[x] = \theta$. (Cette paramétrisation est dite *paramétrisation en moyenne*; voir Brown (1986b, Chapitre 3.) Une suite de lois conjuguées naturelles est donnée par $(m \in \mathbb{N})$

$$h_m(\theta|s) = \exp\{s\cdot\tau(\theta) - m\gamma(\theta)\}c_m(s), \tag{3.24}$$

où $c_m(s)$ est la constante de normalisation. Rappelons que la loi a priori (3.24) correspond à l'actualisation d'une loi a priori plate sur θ pour m observations *fictives* (ou *virtuelles*) $\tilde{x}_1,\ldots,\tilde{x}_m$ de $f(x|\theta)$, telles que $s = \sum_{i=1}^m \tilde{x}_i$.
En fait, la fonction (3.24) peut aussi être considérée comme la densité de s pour une mesure $d\nu_m$ appelée *mesure naturelle*. Si $\mathscr{S}_m$ est l'espace dans lequel s varie, et dQ_m est une mesure de probabilité sur $\mathscr{S}_m$,

$$v_m(\theta) = \int_{\mathscr{S}_m} h_m(\theta|s)\, dQ_m(s) \tag{3.25}$$

est un mélange (continu) de lois a priori conjuguées. Pour une loi a priori π sur Θ, on définit

$$dQ_m(s) = \frac{\pi(s/m)\, d\nu_m(s)}{\int_{\mathscr{S}_m} \pi(t/m)\, d\nu_m(t)},$$

ce qui donne une approximation de π, comme le montre le lemme suivant.

Théorème 3.42. *Si ν_m est absolument continue par rapport à la mesure de Lebesgue ou par rapport à la mesure de comptage sur $\mathscr{S}_m$, et admet pour densité $f_m(s)$, et si $f_m(s)$ converge uniformément sur $\mathscr{S}_m$ vers 1 lorsque m tend vers $+\infty$, alors*

$$v_m(\theta) \longrightarrow \pi(\theta)$$

point par point, et globalement pour la norme L_1.

Tab. 3.6. Approximation d'une loi a priori par mélange de lois conjuguées. (*Source :* Dalal et Hall, 1983.)

Distrib. $f(x,\theta)$	$\tau(\theta)$, $\gamma(\theta)$	$c_m(s)$	$h_m(\theta\|s)$
Normal $\mathscr{N}(\theta,1)$	θ, $\theta^2/2$	$\sqrt{m}\varphi(s/\sqrt{m})$	$\theta \sim \mathscr{N}\left(\frac{s}{m}, \frac{1}{m}\right)$
Gamma $\mathscr{G}\left(\frac{\beta}{\theta}, \theta\right)$	$-\frac{\beta}{\theta}$, $-\beta\log\left(\frac{\beta}{\theta}\right)$	$\frac{s^{m\beta-1}}{\beta\,\Gamma(m\beta-1)}$	$\frac{1}{\theta} \sim \mathscr{G}(s\beta, m\beta-1)$
Poisson $\mathscr{P}(\theta)$	$\log\theta$, θ	$m^{s+1}/\Gamma(s+1)$	$\theta \sim \mathscr{G}(m, s+1)$
Bernoulli $\mathscr{B}(1,\theta)$	$\log\frac{\theta}{1-\theta}$, $\log\frac{1}{1-\theta}$	$\frac{(m+1)!}{s!(m-s)!}$	$\theta \sim \mathscr{B}e(s+1, m-s+1)$
Neg. bin. $\mathscr{N}eg\,(r,\ r/r+\theta)$	$\log\frac{\theta}{r+\theta}$, $r\log(r+\theta)$	$\frac{r^{mr}(mr+s-1)!}{rs!(mr-2)!}$	$\frac{r}{r+\theta} \sim \mathscr{B}e(mr-1, s+1)$

De plus, cette approximation reste valide a posteriori, au sens de la distance de la variation totale, définie comme

$$||\pi - \tilde{\pi}||_{TV} = \sup_A |\pi(A) - \tilde{\pi}(A)|\,,$$

et donc toujours inférieure à 1. Il s'agit donc en quelque sorte d'un résultat d'approximation plus faible, relativement à la norme L_1 du Théorème 3.42.

Théorème 3.43. *Si p_m, la loi marginale de x sous h_m, est finie et si $\pi(\theta)$ et $\pi(\theta|x)$ sont régulières, $v_m(\theta|x)$ converge vers $\pi(\theta|x)$, point par point et au sens de la variation totale.*

La loi a posteriori approchée est, pour n observations et $t = \sum_{i=1}^{n} x_i$,

$$v_m(\theta|n,t) = \frac{\int_{\mathscr{S}_m} h_{m+n}(\theta|s+t)\dfrac{c_m(s)}{c_{m+n}(s+t)}\pi(s/m)\,d\nu_m(s)}{\int_{\mathscr{S}_m}\dfrac{c_m(s')}{c_{m+n}(s'+t)}\pi(s'/m)\,d\nu_m(s')}$$

et le Tableau 3.6 donne les valeurs de τ, γ et c_m pour quelques lois usuelles.

Comparés aux résultats de Diaconis et Ylvisaker (1985), les Théorèmes 3.42 et 3.43 sont effectivement plus généraux et assurent, de plus, la convergence des lois a posteriori. L'inconvénient cependant est que cette approche ne conserve pas l'avantage principal des lois a priori conjuguées, à savoir leur simplicité. Les méthodes de simulation présentées dans le Chapitre 6 sont donc nécessaires pour le calcul de ces estimateurs de Bayes.

3.8.4 Correction de Bartlett

Dans la théorie asymptotique standard, la statistique du rapport de vraisemblance

$$\varpi_n = 2\left\{\sum_{i=1}^{n} f(x_i|\hat{\theta}) - \sum_{i=1}^{n} f(x_i|\hat{\theta}_0)\right\},$$

est distribuée approximativement selon une loi du χ^2_k, où $\hat{\theta}_0$ et $\hat{\theta}$ sont les estimateurs du maximum de vraisemblance contraint et non contraint, et où k est le nombre de contraintes (Gouriéroux et Monfort, 1996). Bartlett (1937) remarque qu'un meilleur ajustement à une loi du χ^2_k est obtenu lorsque ϖ_n est remplacé par $k\varpi_n/\mathbb{E}_\theta[\varpi_n]$, au sens où (Lawley, 1956)

$$P_\theta\left(\frac{k\varpi_n}{\hat{E}} \le t\right) = \chi^2_k(t) + O\left(n^{-2}\right),$$

où $\hat{E}$ est un estimateur approprié de $\mathbb{E}_\theta[\varpi_n]$ et $\chi^2_k(t)$ est la fonction de répartition d'un χ^2_k. La *correction de Bartlett* permet ainsi de réduire l'erreur d'approximation (par un χ^2_k) de $O\left(n^{-1}\right)$ à $O\left(n^{-2}\right)$.

Comme l'ont noté DiCiccio et Stern (1994), si $\theta = (\psi, \varphi)$ et si la contrainte sur θ est que ψ soit fixé, le rapport de vraisemblance dépend de ψ, $\varpi_n = \varpi_n(\psi)$. Bickel et Ghosh (1990) ont établi que la correction de Bartlett s'étend à la loi a posteriori de $\varpi_n(\psi)$, c'est-à-dire qu'il existe une correction de $\varpi_n(\psi)$ telle que

$$P\left(\varpi_n(\psi)\times\left(1-\frac{A_B}{k}\right) \le t \,\middle|\, x,\ldots,x_n\right) = \chi^2_k(t) + O\left(n^{-2}\right),$$

où A_B se déduit d'un développement de l'espérance a posteriori

$$\mathbb{E}[\varpi_n(\psi)|x_1,\ldots,x_n] = k + A_B + O\left(n^{-3/2}\right)$$

et est d'ordre $O\left(n^{-1}\right)$. DiCiccio et Stern (1994) ont aussi montré que cette approximation du second ordre par un χ^2_k reste valide pour une statistique du rapport de vraisemblance ajustée, $\varpi_n(\psi) + \omega_n(\psi)$, où $\omega_n(\psi)$ est $O(1)$. Par exemple, Kass et Steffey (1989) utilisent

$$\omega_n(\psi) = \frac{-1}{2} \log \left(\frac{\det \ell_{\varphi\varphi}(\hat{\theta}(\psi))}{\det \ell_{\varphi\varphi}(\hat{\theta})} \right) + \log \left(\frac{\pi(\hat{\theta}(\psi))}{\pi(\hat{\theta})} \right) ,$$

où $\hat{\theta}(\psi)$ et $\hat{\theta}$ sont les estimateurs du maximum de vraisemblance contraint et non contraint, et $\ell_{\varphi\varphi}$ est la matrice des dérivées secondes de la log-vraisemblance pour le paramètre de nuisance φ. DiCiccio et Stern (1994) établissent la correction correspondante A_B, tandis que DiCiccio et Stern (1993) donnent le facteur de correction de la statistique du rapport a posteriori

$$\kappa^{\pi} = 2 \left\{ \log \pi(\hat{\psi}|x) - \log \pi(\psi|x) \right\} .$$

où $\hat{\psi}$ est l'estimateur MAP marginal de ψ.

Exemple 3.44. (DiCiccio et Stern, 1993) Soit le modèle de régression normale

$$y_i \sim \mathcal{N} \left(\sum_{j=1}^{k} u_{ij}\beta_j, \sigma^2 \right) , \qquad i = 1, \ldots, n,$$

associé à une loi a priori impropre plate sur (β, η), pour $\eta = \log \sigma$. Si le paramètre d'intérêt est η (ou σ^2), alors

$$\kappa^{\pi} = (n - k + 2) \left[\frac{n\varrho}{n - k + 2} - \log \frac{n\varrho}{n - k + 2} - 1 \right] ,$$

où $\varrho = \hat{\sigma}^2/\sigma^2$ et le terme de correction $\aleph^{\pi}$, tel que $(1 + \aleph^{\pi})^{-1}\kappa^{\pi}$ soit χ^2_p à un terme $O(n^{-2})$ près, est $\aleph^{\pi}(\eta) = n^{-1}/3$. Lorsque $\xi = (\beta_1, \ldots, \beta_p)$ est le paramètre d'intérêt, $\aleph^{\pi}(\xi) = (1 + p/2)n^{-1}$. ‖

4

Estimation bayésienne ponctuelle

> "There is always something new from you,' Perrin growled. 'Can't you tell us what to expect once in a while, instead of explaining after it happens ?"
>
> **Robert Jordan**, *The Dragon Reborn.*

4.1 Inférence bayésienne

4.1.1 Introduction

Quand la loi a priori $\pi(\theta)$ est disponible, la loi a posteriori $\pi(\theta|x)$ peut être construite formellement à partir de l'observation x, de distribution $f(x|\theta)$. Cette loi de mise à jour est alors un résumé complet de l'information disponible sur le paramètre θ, résumé qui intègre simultanément l'information a priori et l'information apportée par l'observation x. (Évidemment, ceci reste vrai pour un échantillon $x_1, \ldots, x_n$, mais on peut revenir généralement à la situation précédente grâce à une statistique exhaustive.) La version bayésienne du principe de vraisemblance implique par conséquent que l'inférence sur θ dépend entièrement de la loi a posteriori $\pi(\theta|x)$. Même si θ n'est pas nécessairement conçue comme *variable aléatoire*, la loi $\pi(\theta|x)$ peut être utilisée comme une distribution de probabilité habituelle pour décrire les propriétés de θ. Les indicateurs résumant $\pi(\theta|x)$ tels que moyenne, mode, variance, médiane a posteriori, sont par exemple des estimateurs potentiels. Notamment, lorsque la quantité d'intérêt est $h(\theta)$, un estimateur possible de $h(\theta)$ est la moyenne a posteriori $\mathbb{E}^{\pi}[h(\theta)|x]$. (Comme il a été dit dans la Section 3.5, quand la loi

π est une loi non informative, quelques difficultés de *marginalisation* peuvent se produire et il est parfois nécessaire de construire une nouvelle loi a priori de référence pour le paramètre d'intérêt $h(\theta)$.)

4.1.2 Estimateur MAP

S'il faut faire un choix entre les quantités a posteriori données ci-dessus, ce choix est impossible sans critère de coût, de sorte à définir correctement la notion de "meilleur estimateur". Néanmoins, un estimateur de référence de θ fondé sur $\pi(\theta|x)$ est l'*estimateur du maximum a posteriori (MAP)*, défini comme le mode a posteriori. Notons que l'estimateur MAP maximise aussi $\ell(\theta|x)\pi(\theta)$ et, par conséquent, ne requiert pas le calcul de la loi marginale.

Cet estimateur est associé au coût $0-1$, comme on l'a vu dans la Section 2.5.3, dans le cas particulier $\theta \in \{0,1\}$. Dans le cas continu, puisque, pour tout $\delta \in \Theta$,

$$\int_\Theta \mathbb{I}_{\delta \neq \theta} \pi(\theta|x) d\theta = 1 \,,$$

la fonction de coût $0-1$ peut être remplacée par une suite de coûts, $\mathrm{L}_\varepsilon(d,\theta) = \mathbb{I}_{||\theta-d||>\varepsilon}$, et l'estimateur MAP est alors la limite des estimateurs de Bayes associés à L_ε, quand ε tend vers 0. Il peut aussi être associé à la suite de fonctions de coût L^p, $\mathrm{L}^p(d,\theta) = ||\theta - d||^p$, quand p tend vers l'infini.

Cet estimateur naturel peut s'exprimer comme un *estimateur du maximum de vraisemblance pénalisée* au sens classique (Akaike, 1978, 1983). Notons que les propriétés d'optimalité asymptotique pour un estimateur de maximum de vraisemblance habituel (cohérence, efficacité) sont maintenues pour ces extensions bayésiennes, sous certaines conditions de régularité sur f et π (voir la Note 1.8.4, et Ibragimov et Has'minskii, 1981). Cette extension des propriétés asymptotiques de l'estimateur du maximum de vraisemblance est raisonnable intuitivement, car, lorsque la taille de l'échantillon tend vers l'infini, l'information contenue dans cet échantillon devient prédominante par rapport à l'information *fixe* apportée par la loi a priori π. Cependant, les estimateurs MAP sont asymptotiquement équivalents aux estimateurs du maximum de vraisemblance classiques[28], et, de plus, ont l'avantage d'être disponibles pour des tailles finies d'échantillons.

Exemple 4.1. Soient $x \sim \mathscr{B}(n,p)$. Nous avons vu dans le chapitre précédent que la loi de Jeffreys est dans ce cas la loi bêta $\mathscr{B}e(1/2,1/2)$, soit

$$\pi^*(p) = \frac{1}{B(1/2,1/2)} p^{-1/2}(1-p)^{-1/2} \,,$$

[28] Cette équivalence avec le maximum de vraisemblance n'est, bien sûr, plus valide lorsque le nombre de paramètres croît avec le nombre d'observations, où des incohérences peuvent apparaître (Diaconis et Freedman, 1986).

en omettant la fonction indicatrice $\mathbb{I}_{[0,1]}(p)$ pour simplifier les notations. Deux autres lois non informatives ont été proposées, respectivement par Laplace (1786) et Haldane (1931) (voir aussi l'Exercice 4.4),

$$\pi_1(p) = 1 \qquad \text{et} \qquad \pi_2(p) = p^{-1}(1-p)^{-1}\,.$$

Les estimateurs MAP correspondant sont alors, pour $n > 2$,

$$\begin{aligned}\delta^*(x) &= \max\left(\frac{x-1/2}{n-1}, 0\right),\\ \delta_1(x) &= \frac{x}{n},\\ \delta_2(x) &= \max\left(\frac{x-1}{n-2}, 0\right).\end{aligned}$$

Quand $n = 1$, δ^* et δ_2 sont égaux à δ_1. Pour $n = 2$ et $x = 1$, l'estimateur δ_2 est aussi égal à δ_1, qui est un estimateur du maximum de vraisemblance habituel. On voit bien que, quand n est grand, les trois estimateurs sont effectivement équivalents. ‖

Exemple 4.2. Soit $x \sim \mathscr{C}(\theta, 1)$, c'est-à-dire

$$f(x|\theta) = \frac{1}{\pi}\left[1 + (x-\theta)^2\right]^{-1},$$

et $\pi(\theta) = \frac{1}{2}e^{-|\theta|}$. L'estimateur MAP de θ est alors $\delta^*(x) = 0$, puisque le maximum de $\exp(-|\theta|)[1 + (x-\theta)^2]^{-1}$ est atteint en $\theta = 0$, quelle que soit la valeur de x ! Ce comportement surprenant d'un estimateur qui ne dépend pas de x peut s'expliquer par le caractère plat de la fonction de vraisemblance, qui n'est pas suffisamment informative relativement à une loi a priori très précise. Bien entendu, d'un point de vue pratique, cet estimateur est sans intérêt, mais ce paradoxe disparaît lorsque le nombre d'observations augmente (Exercices 4.6 et 4.7). ‖

4.1.3 Principe de vraisemblance

L'inférence bayésienne apparaît comme une façon efficace de mettre en œuvre le principe de vraisemblance, puisqu'elle fournit un estimateur, en sélectionnant, comme dans l'Exemple 4.3 ci-dessous, l'un des maxima de la fonction de vraisemblance. Comme l'ont souligné Savage (1954) et Berger et Wolpert (1988), de nombreuses considérations philosophiques et pratiques relient le principe de vraisemblance à une approche bayésienne robuste. En particulier, ceci permet l'élimination de quelques paradoxes classiques, comme ceux de Stein (1962b), Stone (1976), Fraser *et al.* (1984) et Le Cam (1990). L'exemple suivant illustre la résolution du paradoxe de Fraser *et al.* (1984). (Voir aussi Joshi, 1967b, pour une analyse plus générale de ce phénomène.)

Exemple 4.3. (Berger et Wolpert, 1988) Soit $\mathscr{X} = \Theta = \mathbb{N}^*$ et

$$f(x|\theta) = \frac{1}{3} \text{ pour } x = \begin{cases} \theta/2, 2\theta, 2\theta+1 & \text{si } \theta \text{ est pair,} \\ (\theta-1)/2, 2\theta, 2\theta+1 & \text{si } \theta \neq 1 \text{ est impair,} \\ 1, 2, 3 & \text{si } \theta = 1. \end{cases}$$

La fonction de vraisemblance est alors

$$\ell(\theta|x) = \frac{1}{3} \text{ pour } \theta = \begin{cases} x/2, 2x, 2x+1 & \text{si } x \text{ est pair,} \\ (x-1)/2, 2x, 2x+1 & \text{si } x \neq 1 \text{ est impair,} \\ 1, 2, 3 & \text{si } x = 1, \end{cases} \tag{4.1}$$

et les trois valeurs de θ pour lesquelles $\ell(\theta|x) \neq 0$ sont pondérées de la même manière par la fonction de vraisemblance. Considérons les trois estimateurs suivants :

$$\delta_1(x) = \begin{cases} x/2 & \text{si x est pair,} \\ (x-1)/2 & \text{si } x \neq 1 \text{ est impair,} \\ 1 & \text{si } x = 1, \end{cases}$$

et

$$\delta_2(x) = 2x, \qquad \delta_3(x) = 2x+1.$$

Ils sont équivalents du point de vue du principe de vraisemblance, car la fonction de vraisemblance est constante sur son support, mais δ_2 et δ_3 sont des estimateurs relativement sous-optimaux puisque

$$P(\delta_2(x) = \theta) = P(x = \theta/2) = \begin{cases} 1/3 & \text{si } \theta \text{ est pair,} \\ 0 & \text{sinon,} \end{cases}$$

$$P(\delta_3(x) = \theta) = P(x = (\theta-1)/2) = \begin{cases} 1/3 & \text{si } \theta \neq 1 \text{ est impair,} \\ 0 & \text{sinon,} \end{cases}$$

tandis que

$$P(\delta_1(x) = \theta) = \begin{cases} 1 & \text{si } \theta = 1, \\ 2/3 & \text{sinon.} \end{cases}$$

L'estimateur δ_1 est donc préférable pour des coûts comme le coût $0-1$. Quand l'information disponible sur le modèle se réduit à la fonction de vraisemblance (4.1), une loi non informative possible sur θ est $\pi(\theta) = 1/\theta$, car θ peut être considéré approximativement comme un paramètre d'échelle. Dans ce cas,

$$\pi(\theta|x) \propto \frac{1}{3\theta} \left[\mathbb{I}_{\delta_1(x)}(\theta) + \mathbb{I}_{\delta_2(x)}(\theta) + \mathbb{I}_{\delta_3(x)}(\theta) \right]$$

et cette loi a posteriori donne $\delta_1(x)$ comme étant quatre fois plus probable que $\delta_2(x)$ ou $\delta_3(x)$. On peut aussi montrer que $P^\pi(\theta = \delta_1(x)|x) \simeq 2/3$ pour

x grand. Cela permet de justifier le choix de δ_1. Une modélisation a priori plus informative conduirait à une conclusion similaire (car une distribution convenable $\pi(\theta)$ doit décroître pour θ suffisamment grand). ‖

Berger et Wolpert (1988) fournissent des résolutions similaires aux paradoxes exhibés par Stein (1962b) et Stone (1976). Un avantage immédiat de l'approche bayésienne, comparativement à d'autres mises en œuvre du principe de vraisemblance est qu'elle traite les paramètres de nuisance intervenant dans la fonction de vraisemblance en les marginalisant. En fait, si $\ell(\theta, \tau|x)$ dépend aussi du paramètre de nuisance τ, une construction naturelle d'une estimation $\hat{\theta}$ de θ est de considérer le maximum de vraisemblance intégré

$$\int \ell(\theta, \tau|x)\pi(\theta, \tau)\, d\tau$$

au lieu d'une vraisemblance "profilée" plus classique,

$$\max_{\tau} \ell(\theta, \tau|x)\pi(\theta, \tau).$$

Voir aussi Basu (1988) pour une analyse étendue du traitement des paramètres de nuisance.

4.1.4 Espace des paramètres restreint

Berger (1985b) remarque l'intérêt d'une approche bayésienne non informative pour des *espaces des paramètres restreints*, la loi a priori étant simplement la troncation d'une loi non informative sans contrainte.

D'un point de vue classique, le calcul d'estimateurs du maximum de vraisemblance restreints est souvent compliqué, notamment quand les contraintes sont non linéaires (voir Robertson *et al.*, 1988). En revanche, la mise en œuvre d'une approche bayésienne via des méthodes de simulation de Monte Carlo (voir le Chapitre 6) permet un calcul aisé des estimateurs de Bayes. (Cet avantage peut même être utilisé pour calculer des estimateurs du maximum de vraisemblance restreints à travers des techniques bayésiennes. Voir Geyer et Thompson, 1992, Robert et Hwang, 1996 et Robert et Casella, 2004, Chapitre 5.)

Exemple 4.4. Soit l'estimation du *modèle de régression linéaire*

$$y = b_1 X_1 + b_2 X_2 + \epsilon, \tag{4.2}$$

qui relie les *revenus directs* (X_1), les *revenus de l'épargne* (X_2) et l'*épargne* (y). Une estimation précise des *taux d'épargne* b_1 et b_2 peut aider le gouvernement à déterminer les taux d'intérêt ou la politique fiscale. Les taux d'intérêt sont évidemment contraints par $0 \leq b_1,\ b_2 \leq 1$. Soit un échantillon

$(y_1, X_{11}, X_{21}), \ldots, (y_n, X_{1n}, X_{2n})$ de (4.2) et supposons que les erreurs ϵ_i soient indépendantes et distribuées selon $\mathcal{N}(0,1)$, c'est-à-dire que $y_i \sim \mathcal{N}(b_1 X_{1i} + b_2 X_{2i}, 1)$. La loi non informative correspondante est alors la loi propre

$$\pi(b_1, b_2) = \mathbb{I}_{[0,1]}(b_1)\mathbb{I}_{[0,1]}(b_2)$$

et la moyenne a posteriori est donnée par $(i = 1, 2)$

$$\mathbb{E}^\pi[b_i|y_1, \ldots, y_n] = \frac{\int_0^1 \int_0^1 b_i \prod_{j=1}^n \varphi(y_j - b_1 X_{1j} - b_2 X_{2j})\, db_1\, db_2}{\int_0^1 \int_0^1 \prod_{j=1}^n \varphi(y_j - b_1 X_{1j} - b_2 X_{2j})\, db_1\, db_2},$$

où φ est la densité de la loi normale centrée. Si on note par $(\hat{b}_1, \hat{b}_2)$ l'estimateur des moindres carrés non contraints de (b_1, b_2), qui est aussi l'estimateur du maximum de vraisemblance régulier de (b_1, b_2), la loi a posteriori non contrainte sur (b_1, b_2) est

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} \hat{b}_1 \\ \hat{b}_2 \end{pmatrix}, \left(X^t X\right)^{-1} \right), \tag{4.3}$$

avec

$$X = \begin{pmatrix} X_{11} & X_{21} \\ \vdots & \vdots \\ X_{1n} & X_{2n} \end{pmatrix}.$$

Par conséquent, l'estimateur de Bayes restreint est donné par $(i = 1, 2)$

$$\delta_i^\pi(y_1, \ldots, y_n) = \frac{\mathbb{E}^\pi \left[b_i \mathbb{I}_{[0,1]^2}(b_1, b_2) | y_1, \ldots, y_n \right]}{P^\pi \left((b_1, b_2) \in [0,1]^2 | y_1, \ldots, y_n \right)},$$

où le terme de droite est calculé sous la loi (4.3). Si on indique

$$\Sigma = (X^t X)^{-1} = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{12} & \sigma_{22}^2 \end{pmatrix},$$

la loi conditionnelle de b_1 est

$$b_1 | b_2 \sim \mathcal{N}\left(\hat{b}_1 + \sigma_{12}(b_2 - \hat{b}_2)/\sigma_{22}^2, \sigma_{11}^2 - \sigma_{12}^2 \sigma_{22}^{-2} \right).$$

Alors

$$\begin{aligned} &P^\pi \left((b_1, b_2) \in [0,1]^2 | y_1, \ldots, y_n \right) \\ &\quad = \int_0^1 \left\{ \Phi\left(\frac{1 - \hat{b}_1 - \sigma_{12}(b_2 - \hat{b}_2)/\sigma_{22}^2}{\sqrt{\sigma_{11}^2 - \sigma_{12}^2 \sigma_{22}^{-2}}} \right) \right. \\ &\qquad \left. - \Phi\left(\frac{-\hat{b}_1 - \sigma_{12}(b_2 - \hat{b}_2)/\sigma_{22}^2}{\sqrt{\sigma_{11}^2 - \sigma_{12}^2 \sigma_{22}^{-2}}} \right) \right\} \sigma_{22}^{-1} \varphi\left(\frac{b_2 - \hat{b}_2}{\sigma_{22}} \right) db_2 \end{aligned}$$

et

$$\mathbb{E}^\pi[b_i \mathbb{I}_{[0,1]^2}(b_1,b_2)|y_1,\ldots,y_n] = \int_0^1 \left[\hat{b}_1 + \frac{\sigma_{12}}{\sigma_{22}^2}(b_2-\hat{b}_2) + (\sigma_{11}^2 - \sigma_{12}^2\sigma_{22}^{-2})^{1/2} \left\{ \varphi\left(\frac{1-\hat{b}_1-\sigma_{12}(b_2-\hat{b}_2)/\sigma_{22}^2}{\sqrt{\sigma_{11}^2-\sigma_{12}^2\sigma_{22}^{-2}}}\right) - \varphi\left(\frac{-\hat{b}_1-\sigma_{12}(b_2-\hat{b}_2)/\sigma_{22}^2}{\sqrt{\sigma_{11}^2-\sigma_{12}^2\sigma_{22}^{-2}}}\right)\right\}\right] \sigma_{22}^{-1}\varphi\left(\frac{b_2-\hat{b}_2}{\sigma_{22}}\right) db_2.$$

Notons qu'on peut obtenir la seconde intégrale sous forme explicite en utilisant la fonction de répartition Φ d'une Gaussienne centrée réduite, mais le dénominateur ne peut pas être calculé de façon analytique. Il est donc plus efficace de calculer les deux intégrales par une (seule) simulation de Monte Carlo (voir le Chapitre 6).

Si b_1 et b_2 sont indépendants a posteriori, c'est-à-dire si $\sigma_{12} = 0$, l'estimateur de Bayes est explicite et donné par $(i = 1, 2)$

$$\mathbb{E}^\pi[b_i|y_1,\ldots,y_n] = \hat{b}_i - \sigma_{ii}\frac{\exp\{-(1-\hat{b}_i)^2/2\sigma_{ii}^2\} - \exp\{-\hat{b}_i^2/2\sigma_{ii}^2\}}{\sqrt{2\pi}\{\Phi((1-\hat{b}_i)/\sigma_{ii}) - \Phi(-\hat{b}_1/\sigma_{ii})\}}.$$

‖

Notons que la modélisation bayésienne est encore plus appropriée lorsqu'il s'agit d'incorporer une information *vague*, c'est-à-dire dans des cas où une restriction sur l'espace des paramètres est probable mais pas certaine. Le Chapitre 10 démontre qu'une manière typique de traiter ces cas est d'utiliser une modélisation empirique ou hiérarchique.

4.1.5 Précision des estimateurs de Bayes

Puisque la loi a posteriori $\pi(\theta|x)$ est complètement disponible, il est possible d'associer à un estimateur $\delta^\pi(x)$ de $h(\theta)$ une évaluation de la *précision* de l'estimation via, par exemple, l'*erreur quadratique a posteriori*,

$$\mathbb{E}^\pi[(\delta^\pi(x) - h(\theta))^2|x],$$

égale à $\text{var}^\pi(h(\theta)|x)$ lorsque $\delta^\pi(x) = \mathbb{E}^\pi[h(\theta)|x]$. De la même façon, dans un cadre multidimensionnel, la matrice de covariance caractérise la performance des estimateurs. Ces indications additionnelles fournies par la loi a posteriori illustrent l'avantage opérationnel de l'approche bayésienne, car l'approche classique a souvent des difficultés à motiver le choix de ces évaluations.

De plus, les mesures d'évaluation bayésiennes sont toujours conditionnelles[29], tandis que l'approche fréquentiste doit recourir à des bornes supérieures au moyen du principe minimax, car le paramètre θ est inconnu (voir Berger et Robert, 1990, pour une comparaison des deux approches).

Exemple 4.5. (Suite de l'Exemple 4.1) Soit l'estimateur du maximum de vraisemblance de p, $\delta_1(x) = x/n$. Alors

$$\begin{aligned}\mathbb{E}^\pi[(\delta_1(x)-p)^2|x] &= \mathbb{E}^\pi[(p-x/n)^2|x] \\ &= \left(\frac{x+1/2}{n+1}-\frac{x}{n}\right)^2 + \frac{(x+1/2)(n-x+1/2)}{(n+1)^2(n+2)} \\ &= \frac{(x-n/2)^2}{(n+1)^2n^2} + \frac{(x+1/2)(n-x+1/2)}{(n+1)^2(n+2)}, \end{aligned} \tag{4.4}$$

car $\pi(p|x)$ est la loi bêta $\mathscr{B}e(x+1/2, n-x+1/2)$. D'un point de vue fréquentiste, le risque de l'estimateur du maximum de vraisemblance est

$$\mathbb{E}_p[(\delta_1(x)-p)^2] = \text{var}(x/n) = \frac{p(1-p)}{n}$$

et

$$\sup_p p(1-p)/n = 1/4n.$$

En développant (4.4), il est facile de vérifier que le maximum de (4.4) est

$$1/[4(n+2)],$$

quantité toujours plus petite que $1/4n$. Le principal avantage de (4.4) est de fournir malgré tout une réponse *modulable* pour l'évaluation de δ_1, car (4.4) varie entre $1/[4(n+2)]$ et $3/[4(n+1)(n+2)]$. Bien évidemment, une approximation fréquentiste de $p(1-p)/n$ peut aussi être proposée, à savoir $(x/n)(1-x/n)/n$. Cette évaluation souffre alors de l'inconvénient opposé, car il varie trop largement, comme le montre la Figure 4.1. Il peut même prendre la valeur 0 quand x vaut 0 ou n. Un comportement similaire est discuté par Berger (1990) dans un cadre général. ‖

4.1.6 Prévision

L'inférence bayésienne peut être aussi mise en œuvre dans des problèmes de *prévision*. Si $x \sim f(x|\theta)$ et $z \sim g(z|x,\theta)$, où z ne dépend pas nécessairement de x, la *distribution prédictive* de z après observation de x est donnée par

[29] En fait, il existe des contreparties bayésiennes aux inégalités de Cramér-Rao utilisées dans l'évaluation des estimateurs non biaisés. Il s'agit des bornes de *Van Trees* (Gill et Levit, 1995), utilisées en traitement de signal et dans d'autres domaines, comme l'ont illustré Bergman *et al.* (2001)

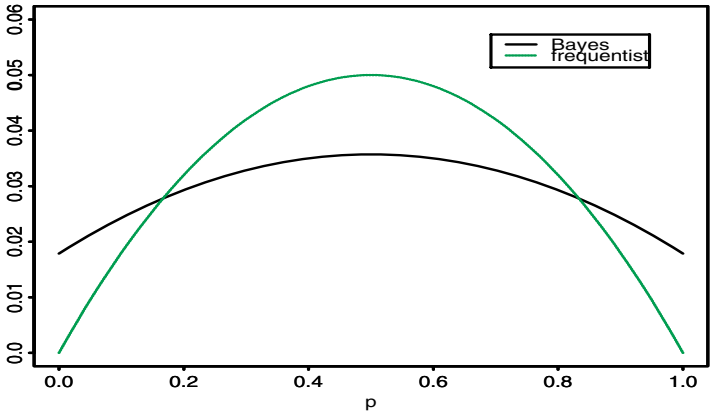

Fig. 4.1. Comparaison des évaluations bayésienne et fréquentiste de l'erreur d'estimation dans le cas binomial ($n = 3$).

$$g^{\pi}(z|x) = \int_{\Theta} g(z|x,\theta)\pi(\theta|x)\, d\theta. \tag{4.5}$$

La distribution de z est alors assez logiquement moyennée sur les valeurs de θ relativement à la loi a posteriori, qui est aussi la distribution actualisée de θ. Il est possible d'utiliser (4.5) pour calculer la moyenne et la variance prédictive de la variable aléatoire z. Dans la Section 4.3.1, nous considérons un exemple particulier de détermination d'une distribution prédictive discrète (voir aussi l'Exercice 4.41).

Exemple 4.6. Le modèle AR (1), où AR signifie *autorégressif*, est un modèle dynamique qui définit la distribution d'une variable au temps t ($1 \leq t \leq T$), x_t, conditionnellement à l'observation précédente x_{t-1}, comme

$$x_t = \varrho x_{t-1} + \epsilon_t\,,$$

où les ϵ_t sont i.i.d. $\mathcal{N}(0,\sigma^2)$. (Ce modèle sera considéré en détail dans la Section 4.5.) Pour une suite d'observations donnée jusqu'au temps $T-1$, $x_{1:(T-1)} = (x_1,\ldots,x_{(T-1)})$, la distribution prédictive de x_T est alors

$$x_T|x_{1:(T-1)} \sim \int \frac{1}{\sqrt{2\pi}}\sigma^{-1}\exp\{-(x_T-\varrho x_{T-1})^2/2\sigma^2\}\pi(\varrho,\sigma|x_{1:(T-1)})d\varrho d\sigma\,,$$

où $\pi(\varrho,\sigma|x_{1:(T-1)})$ peut être formulée explicitement (Exercice 4.14). ‖

Notons que l'approche de la Théorie de la Décision développée dans les sections suivantes s'applique aussi à la prédiction, même si nous ne mentionnerons plus ce point par la suite. En fait, si un coût de prédiction $\mathrm{L}(z,\delta)$

est disponible, un prédicteur $\delta(x)$ peut être choisi qui minimise l'erreur de prédiction moyenne (l'espérance étant calculée par rapport à la distribution prédictive (4.5)) ; voir l'Exercice 4.46.

4.1.7 Retour à la décision

Étant donné l'étendue des utilisations possibles de la loi a posteriori, certains considèrent qu'on devrait fournir aux clients la loi a posteriori afin qu'ils puissent l'utiliser à leur guise. Bien que la communication de $\pi(\theta|x)$ soit en effet envisageable pour de petites dimensions, sa complexité rend en général difficile l'extraction de l'information qu'elle contient. La loi a posteriori est évidemment essentielle dans le processus de décision, mais il revient au statisticien d'assister plus avant le décideur, afin d'extraire les caractéristiques d'intérêt de $\pi(\theta|x)$. Par conséquent, nous sommes de nouveau confrontés au problème important de sélection d'un estimateur et nous avons vu dans le Chapitre 2 que cette sélection n'est efficace et cohérente que lorsqu'elle est fondée sur un critère de *coût*. Les sections qui suivent mettent en avant la théorie bayésienne de la décision, avec une attention particulière aux cas normaux et d'échantillonnage. Bien que formellement rattachés à la Théorie de la Décision, tests et régions de confiance sont traités séparément dans le chapitre suivant (Chapitre 5).

4.2 Théorie bayésienne de la décision

4.2.1 Estimateurs de Bayes

Rappelons que, pour une fonction de coût $\mathrm{L}(\theta, \delta)$ et une loi a priori (ou une mesure) π, la *règle de Bayes* $\delta^\pi(x)$ est solution de

$$\min_\delta \mathbb{E}^\pi[\mathrm{L}(\theta, \delta)|x].$$

Selon la complexité du coût L et de la loi a posteriori $\pi(\theta|x)$, l'estimateur δ^π sera déterminé analytiquement ou numériquement.

Comme nous l'avons montré dans le Chapitre 2, les solutions associées à des coûts classiques sont formellement connues et correspondent aux caractéristiques usuelles d'une distribution (moyenne, médiane, fractiles, etc.). Par exemple, l'estimateur de Bayes associé au coût quadratique est la moyenne a posteriori (Proposition 2.41 et Corollaire 2.42). Bien sûr, cette construction formelle des estimateurs de Bayes classiques n'évite pas toujours le recours à une approximation numérique, particulièrement dans des cas multidimensionnels.

Exemple 4.7. Soit $x \sim \mathcal{N}_p(\theta, I_p)$. Comme nous l'avons indiqué dans la Section 3.6, la loi de Student fournit une alternative robuste à la loi normale conjuguée pour l'estimation de θ. Soit donc $\theta \sim \mathcal{T}_p(\alpha, 0, \tau^2 I_p)$, c'est-à-dire

$$\pi(\theta|\alpha,\tau) = \frac{\Gamma((\alpha+p)/2)}{(\alpha\tau\pi)^{p/2}\Gamma(\alpha/2)} \left(1 + \frac{||\theta||^2}{\alpha\tau^2}\right)^{-(\alpha+p)/2}.$$

Par conséquent,

$$\pi(\theta|x) \propto \left(1 + \frac{||\theta||^2}{\alpha\tau^2}\right)^{-(\alpha+p)/2} e^{-||x-\theta||^2/2},$$

qui ne conduit pas à une expression explicite de la loi a posteriori. Cependant, il est malgré tout possible de réduire le problème de calcul à celui d'une *intégrale simple*, pour toute valeur de p, comme l'a montré Dickey (1968). En effet, si $\theta \sim \mathcal{T}_p(\alpha, 0, \tau^2 I_p)$, la loi a posteriori de θ peut s'écrire comme un mélange caché (voir l'Exemple 3.17),

$$\begin{aligned} \theta|z &\sim \mathcal{N}_p(0, \tau^2 z I_p), \\ z^{-1} &\sim \mathcal{G}(\alpha/2, \alpha/2), \end{aligned}$$

où z est une variable aléatoire auxiliaire. Conditionnellement à z, la loi a posteriori de θ est

$$\theta|x, z \sim \mathcal{N}_p\left(\frac{x}{1+\tau^2 z}, \frac{\tau^2 z}{1+\tau^2 z} I_p\right)$$

et, comme

$$\pi(z|x) \propto (1+\tau^2 z)^{-p/2} e^{-||x||^2/2(1+\tau^2 z)} \pi(z),$$

on calcule l'estimateur de Bayes comme étant

$$\begin{aligned} \delta^\pi(x) &= \int_0^{+\infty} \mathbb{E}^\pi[\theta|x,z]\pi(z|x)\,dz \\ &= x\, \frac{\int_0^{+\infty} (1+\tau^2 z)^{-(p+2)/2} e^{-||x||^2/2(1+\tau^2 z)} z^{-(\alpha+2)/2} e^{-\alpha/2z}\,dz}{\int_0^{+\infty} (1+\tau^2 z)^{-p/2} e^{-||x||^2/2(1+\tau^2 z)} z^{-(\alpha+2)/2} e^{-\alpha/2z}\,dz}. \end{aligned}$$

Cet estimateur peut donc s'exprimer comme une intégrale simple pour toute valeur de p. ‖

Cependant, des décompositions subtiles comme celle de l'exemple ci-dessus ne sont pas toujours possibles et le calcul d'un estimateur de Bayes nécessite alors une méthode d'approximation générale comme celles décrites dans le Chapitre 6.

En revanche, un résultat intéressant est que, quand la loi marginale $m(x)$ est disponible, l'espérance a posteriori du paramètre naturel d'une famille exponentielle se calcule aisément.

Lemme 4.8. *Soit $f(x|\theta) = h(x)e^{\theta\cdot x-\psi(\theta)}$, une distribution d'une famille exponentielle. Pour toute loi a priori π, la moyenne a posteriori de θ est donnée par*

$$\delta^\pi(x) = \nabla \log m_\pi(x) - \nabla \log h(x), \tag{4.6}$$

où ∇ est l'opérateur gradient et m_π est la loi marginale associée à π.

Preuve. L'espérance a posteriori est donnée par

$$\begin{aligned}\mathbb{E}^\pi[\theta_i|x] &= \frac{\int_\Theta \theta_i h(x)e^{\theta\cdot x-\psi(\theta)}\pi(\theta)\,d\theta}{m_\pi(x)} \\ &= \left(\frac{\partial}{\partial x_i}\int_\Theta h(x)e^{\theta\cdot x-\psi(\theta)}\pi(\theta)\,d\theta\right)\frac{1}{m_\pi(x)} - \left(\frac{\partial}{\partial x_i}h(x)\right)\frac{1}{h(x)} \\ &= \frac{\partial}{\partial x_i}\left[\log m_\pi(x) - \log h(x)\right].\end{aligned}$$

□

Notons que ce lemme est satisfait pour tout π ; il apparaît comme le résultat dual du calcul des moments de $f(x|\theta)$ à partir de la dérivée de ψ dans une famille exponentielle (voir le Lemme 3.13). Son intérêt pratique est hélas plutôt limité, car le calcul de la loi marginale est généralement assez délicat et connaître $m_\pi(x)$ explicitement équivaut à connaître $\pi(\theta|x)$ explicitement[30].

Exemple 4.9. Nous avons introduit dans la Note 2.5.4 l'estimateur de James-Stein tronqué,

$$\delta^{\mathrm{JS}}(x) = \left(1 - \frac{p-2}{||x||^2}\right)^+ x$$

quand $x \sim \mathcal{N}_p(\theta, I_p)$. Dans le cas normal, (4.6) s'écrit

$$\delta^\pi(x) = x + \nabla \log m_\pi(x).$$

Bien qu'il existe une fonction m telle que δ^{JS} peut s'écrire comme ci-dessus (voir Bock, 1988), m n'est pas une loi marginale et cet estimateur ne peut pas être de Bayes : il vaut 0 sur l'ouvert $\{||x||^2 < p-2\}$ et devrait être nul partout du fait de la contrainte d'analycité. ‖

L'expression (4.6) des estimateurs de Bayes est aussi utile pour l'établissement de résultats liés à l'*effet de Stein*, soit pour établir les conditions de domination comme dans Stein (1981), George (1986a), Berger et Robert (1990)

[30] Une conséquence théorique de ce lemme est que les estimateurs de Bayes sont des fonctions *analytiques* (ou holomorphes) si la famille exponentielle considérée est telle que la fonction h qui l'engendre est holomorphe, puisque m_π/h est alors la transformée de Laplace de $e^{-\psi(\theta)}\pi(\theta)$. Le Chapitre 8 établit un critère d'inadmissibilité à partir de cette propriété.

et Brandwein et Strawderman (1990), soit pour caractériser l'admissibilité de certains estimateurs comme dans Bock (1988) et Brown (1988) ; voir l'Exercice 4.44.

Tab. 4.1. Estimateurs de Bayes du paramètre θ sous coût quadratique pour les lois a priori conjuguées des familles exponentielles usuelles.

Loi de x	Loi conjuguée	Moyenne a posteriori
Normale $\mathcal{N}(\theta, \sigma^2)$	Normale $\mathcal{N}(\mu, \tau^2)$	$\dfrac{\mu\sigma^2 + \tau^2 x}{\sigma^2 + \tau^2}$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	$\dfrac{\alpha + x}{\beta + 1}$
Gamma $\mathcal{G}(\nu, \theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	$\dfrac{\alpha + \nu}{\beta + x}$
Binomiale $\mathcal{B}(n, \theta)$	Bêta $\mathcal{B}e(\alpha, \beta)$	$\dfrac{\alpha + x}{\alpha + \beta + n}$
Binomiale négative $\mathcal{N}eg(n, \theta)$	Bêta $\mathcal{B}e(\alpha, \beta)$	$\dfrac{\alpha + n}{\alpha + \beta + x + n}$
Multinomiale $\mathcal{M}_k(n; \theta_1, \ldots, \theta_k)$	Dirichlet $\mathcal{D}(\alpha_1, \ldots, \alpha_k)$	$\dfrac{\alpha_i + x_i}{\left(\sum_j \alpha_j\right) + n}$
Normale $\mathcal{N}(\mu, 1/\theta)$	Gamma $\mathcal{G}(\alpha/2, \beta/2)$	$\dfrac{\alpha + 1}{\beta + (\mu - x)^2}$

4.2.2 Les lois a priori conjuguées

Dans le cas particulier des lois a priori conjuguées, les espérances a posteriori des paramètres naturels admettent évidemment des expressions explicites ; c'est d'ailleurs pratiquement le seul cas où des expressions analytiques sont disponibles dans une telle généralité. Le Tableau 4.1 présente les estimateurs de Bayes associés aux distributions usuelles et à leurs lois a priori conjuguées. Notons que, quand plusieurs observations de $f(x|\theta)$ sont disponibles, on retrouve les mêmes lois a priori conjuguées et que seuls les paramètres dans l'estimateur sont modifiés, ceci en raison des propriétés d'exhaustivité des familles exponentielles (Section 3.3.3).

Exemple 4.10. Si $x_1, ..., x_n$ sont des observations indépendantes de $\mathcal{N}eg(m, \theta)$ et si $\theta \sim \mathcal{B}e(\alpha, \beta)$, la loi a posteriori de θ est la distribution bêta

$$\mathscr{B}e\left(\alpha+mn,\sum_{i=1}^{n}x_i+\beta\right) \quad \text{et} \quad \delta^\pi(x_1,...,x_n)=\frac{\alpha+mn}{\alpha+\beta+mn+\sum_{i=1}^{n}x_i}.$$

Ce résultat est une conséquence directe du fait que $\sum_{i=1}^{n} x_i \sim \mathscr{N}eg(mn,\theta)$. ‖

Exemple 4.11. Soient n observations x_1, ..., x_n de $\mathscr{U}([0,\theta])$ et prenons $\theta \sim \mathscr{P}a(\theta_0,\alpha)$. Alors

$$\theta|x_1,...,x_n \sim \mathscr{P}a(\max(\theta_0,x_1,...,x_n),\alpha+n)$$

et

$$\delta^\pi(x_1,...,x_n)=\frac{\alpha+n}{\alpha+n-1}\max(\theta_0,x_1,...,x_n).$$

Ainsi, par comparaison avec l'estimateur du maximum de vraisemblance,

$$\delta_0(x_1,...,x_n)=\max(x_1,...,x_n),$$

l'estimateur de Bayes donne une estimation plus "optimiste" de θ, car

$$\frac{\alpha+n}{\alpha+n-1}>1\,.$$

Dans le cas limite où $\alpha=0$ et $\theta_0=0$, on retrouve le meilleur estimateur équivariant de θ sous coût quadratique $\frac{n}{n-1}\delta_0(x_1,\ldots,x_n)$ (voir le Chapitre 9), qui est plus grand que δ^π quand $\theta_0=0$. Ce comportement de rétrécissement de δ^π pour $\alpha\neq 1$ s'explique par le choix de π, qui décroît avec θ, et favorise donc les valeurs de θ proches de θ_0. ‖

De même, rappelons que l'estimation d'une fonction de θ, $g(\theta)$, sous coût quadratique, donne comme estimateur de Bayes $\delta^\pi(x)=\mathbb{E}^\pi[g(\theta)|x]$.

Exemple 4.12. Soit $x \sim \mathscr{G}(\nu,\theta)$, où le paramètre de forme ν est connu, et $\theta \sim \mathscr{G}(\alpha,\beta)$. Le paramètre d'intérêt est $1/\theta$, l'espérance de x. Sous le coût quadratique

$$\mathrm{L}(\theta,\delta)=\left(\delta-\frac{1}{\theta}\right)^2,$$

l'estimateur de Bayes est alors

$$\begin{aligned}\delta_1^\pi(X)&=\frac{(\beta+x)^{\alpha+\nu}}{\Gamma(\alpha+\nu)}\int_0^{+\infty}\frac{1}{\theta}\,\theta^{\alpha+\nu-1}e^{-(\beta+x)\theta}d\theta\\&=\frac{\beta+x}{\alpha+\nu-1}.\end{aligned}$$

‖

Sous un coût quadratique renormalisé (ou pondéré),

$$\mathrm{L}(\theta, \delta) = w(\theta) \parallel \delta - \theta \parallel_Q^2,$$

où Q est une matrice $p \times p$ symétrique semi-définie positive, l'estimateur de Bayes associé est

$$\delta^\pi(x) = \frac{\mathbb{E}^\pi[\theta w(\theta)|x]}{\mathbb{E}^\pi[w(\theta)|x]}.$$

Exemple 4.13. (Suite de l'Exemple 4.12) Un coût invariant par changement d'échelle ne dépend pas de l'unité de mesure et peut être plus pertinent pour une estimation de $1/\theta$. Par exemple, le coût

$$\mathrm{L}(\theta, \delta) = \theta^2 \left(\delta - \frac{1}{\theta}\right)^2$$

donne l'estimateur de Bayes

$$\begin{aligned}\delta_2^\pi(x) &= \frac{\mathbb{E}^\pi\left[\theta^2/\theta \mid x\right]}{\mathbb{E}^\pi\left[\theta^2 \mid x\right]} \\ &= \frac{\int_0^{+\infty} \theta\theta^{\alpha+\nu-1} e^{-(\beta+x)\theta} d\theta}{\int_0^{+\infty} \theta^{\alpha+\nu+1} e^{-(\beta+x)\theta} d\theta} \\ &= \frac{\beta + x}{\alpha + \nu + 1} = \frac{\alpha + \nu - 1}{\alpha + \nu + 1}\, \delta_1^\pi(x).\end{aligned}$$

‖

Insistons de nouveau sur le fait que, même pour les lois a priori conjuguées, le fait que l'estimateur de Bayes de toute fonction de θ s'exprime comme une espérance a posteriori n'évite pas nécessairement le calcul numérique, car une intégration analytique peut être impossible, en particulier dans les problèmes multidimensionnels.

Exemple 4.14. Soient $x \sim \mathcal{N}_p(\theta, I_p)$ et $h(\theta) = ||\theta||^2$. Le coût considéré dans Saxena et Alam (1982) est

$$\mathrm{L}(\theta, \delta) = \frac{(\delta - ||\theta||^2)^2}{2||\theta||^2 + p}$$

car, si $\delta_0(x) = ||x||^2 - p$,

$$R(\delta_0, \theta) = \frac{1}{2||\theta||^2 + p}\, \mathbb{E}(||x||^2 - ||\theta||^2 - p)^2 = 2$$

et δ_0 a un risque constant. Sans cette renormalisation, tous les estimateurs ont un risque maximal égal à $+\infty$, tandis que sous L, l'estimateur δ_0 est minimax. Alors, même pour une loi a priori conjuguée, $\mathcal{N}_p(0, \tau^2 I_p)$, le calcul de

$$\delta^\pi(x) = \frac{\mathbb{E}^\pi[||\theta||^2/(2||\theta||^2+p)|x]}{\mathbb{E}^\pi[1/(2||\theta||^2+p)|x]}$$

ne peut pas être effectué analytiquement. ‖

Dans les exemples précédents, nous avons eu largement recours au coût quadratique, car il constitue un coût standard et permet, autant que possible, des calculs explicites. Nous renvoyons les lecteurs au Chapitre 2 pour des critiques sur le caractère arbitraire des coûts standard et l'opposition entre coûts concaves bornés et coûts convexes non bornés, les premiers conduisant à un paradoxe d'amateurs du risque et les seconds à une plus grande instabilité des procédures en résultant (voir Kadane et Chuang, 1978, Smith, 1988, et les Exercices 4.1 et 4.15). Malgré tout, il faut remarquer que, lorsque la fonction de coût est vraiment déterminée par le décideur, celle-ci est généralement complexe et nécessite le plus souvent une minimisation numérique pour aboutir à l'estimateur de Bayes.

4.2.3 Estimation du coût

Pour un coût donné, $\mathrm{L}(\theta,\delta)$, on peut aussi chercher à évaluer les performances de l'estimateur de Bayes $\delta^\pi(x)$. Cette évaluation peut s'interpréter d'un point de vue décisionnel comme l'estimation du coût $\mathrm{L}(\theta,\delta^\pi(x))$ par $\gamma(x)$, sous une seconde fonction de coût, comme

$$\tilde{\mathrm{L}}(\theta,\delta^\pi,\gamma) = \left[\gamma(x) - \mathrm{L}(\theta,\delta^\pi(x))\right]^2. \tag{4.7}$$

De nouveau, le coût quadratique (4.7) n'est pas plus justifié comme choix automatique dans ce contexte que dans d'autres cas d'estimation. Mais, en dehors de son côté pratique, le choix du coût quadratique peut se défendre par l'absence de justification en termes d'utilité et, par conséquent, une perception plus proche de l'erreur comme une variance. Sous (4.7), l'évaluation bayésienne des performances de δ^π est donnée par le résultat suivant.

Proposition 4.15. *L'estimateur de Bayes du coût* $\mathrm{L}(\theta,\delta^\pi(x))$ *sous* (4.7) *pour la loi a priori* π *est*

$$\gamma^\pi(x) = \mathbb{E}^\pi[\mathrm{L}(\theta,\delta^\pi(x))|x].$$

Ce résultat découle directement de la Proposition 2.41, puisque, conditionnellement à x, le but est d'estimer une fonction particulière de θ sous un coût quadratique. Notons que la dépendance de cette fonction à x n'a pas d'importance d'un point de vue bayésien, car, une fois x *observé*, x est *fixé*. De même, pour un coût d'erreur absolue, l'estimateur de Bayes du coût est la médiane de la distribution a posteriori de $\mathrm{L}(\theta,\delta^\pi(x))$, moins facile à obtenir. Quand L est le coût quadratique, la variance a posteriori, $\mathrm{var}^\pi(x)$, est par conséquent l'estimateur de Bayes du coût associé avec δ^π.

L'estimation du coût dans une perspective fréquentiste a été étudiée par Johnstone (1998) et Rukhin (1988a,b), le premier montrant que, pour un estimateur minimax avec un risque constant p, l'évaluation $\gamma(x) = p$ n'est pas nécessairement admissible sous (4.7). Berger (1984, 1985a) (voir aussi Lu et Berger, 1989a,b) développe un concept additionnel pour l'estimation du coût appelé *validité fréquentiste* : un estimateur γ du coût $\mathrm{L}(\theta, \delta(x))$ est *valide en fréquence* si

$$\mathbb{E}_\theta[\gamma(x)] \geq R(\theta, \delta(x)), \qquad \theta \in \Theta,$$

c'est-à-dire si cet estimateur ne sous-estime jamais sur le long terme l'erreur résultant de l'utilisation de δ. Une telle restriction peut sembler intuitivement satisfaisante, mais elle est fondée sur la justification à la base de la notion d'*estimation sans biais*, et cette restriction contredit le principe de vraisemblance.

Robert et Casella (1994) proposent une approche purement décisionnelle de l'estimation du coût pour des régions de confiance (voir le Chapitre 5). Si $C(x)$ est une région de confiance pour θ, le coût usuel pour son estimation est le coût $0-1$,

$$\mathrm{L}(C(x), \theta) = 1 - \mathbb{I}_{C(x)}(\theta)\,.$$

Un estimateur du coût $\gamma(x)$ évalue donc le taux de couverture de $C(x)$ et approche en quelque sorte la probabilité de couverture de la région de confiance. Hwang et Brown (1991) ont ainsi montré que, pour les régions de confiance usuelles C_0, dans un cadre normal, l'estimateur constant

$$\alpha = P(\theta \notin C_0(x))$$

est admissible parmi les estimateurs valides en fréquence, mais est inadmissible pour $p > 5$ en l'absence de cette restriction (voir la Section 5.5).

Exemple 4.16. Soient $x \sim \mathcal{N}_p(\theta, \sigma^2 I_p)$ et $\theta \sim \mathcal{N}_p(0, \tau^2 I_p)$. Sous un coût quadratique,

$$\delta^\pi(x) = \frac{\sigma^2}{\sigma^2+\tau^2}\, x \quad \text{et} \quad V^\pi(x) = \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\, p.$$

En revanche, l'approche fréquentiste donne $+\infty$ comme risque maximal pour δ^π et est donc mal adaptée à ce problème. ‖

4.3 Modèles d'échantillonnage

Dans cette section, nous considérons trois problèmes d'échantillonnage pour lesquels une approche bayésienne est facile à mettre en œuvre. Notons tout d'abord que, en général, les *modèles discrets* nécessitent moins d'information a priori pour construire une loi a priori. Le premier problème que nous

considérons est lié à la *règle de succession de Laplace*, introduite en 1774 par Laplace. Le deuxième problème a été étudié sous le nom de *"problème du tramway"* par Neyman dans les années 1930. La dernière section étudie les modèles de *capture-recapture*, qui sont très intéressants pour la biologie animale et pour d'autres modèles d'estimation de population. Ces trois problèmes ont comme point commun le fait qu'ils proposent une inférence sur une population finie ou sur une sous-population. Il s'agit de cas où une certaine partie de l'information a priori est habituellement disponible, ou bien de cas où on peut faire le choix d'une loi a priori non informative sans (grande) ambiguïté.

4.3.1 Règle de succession de Laplace

Considérons le modèle hypergéométrique $\mathscr{H}(N, N_1, x)$ standard : Soit une population de taille N divisée en deux sous-populations de tailles inconnues respectives N_1 et $N_2 = N - N_1$. Lors d'un tirage sans remise de x individus dans cette population, x_1 individus appartiennent à la première sous-population et $x_2 = x - x_1$ à la seconde. Lorsque aucune information n'est disponible sur N_1, la loi non informative est

$$\pi(N_1) = \frac{1}{N+1}\mathbb{I}_{\{0,1,\ldots,N\}}(N_1)$$

et la loi a posteriori correspondante de N_1 est $(x_1 \leq N_1 \leq N - (x - x_1))$

$$\pi(N_1|x_1) = \frac{\binom{N_1}{x_1}\binom{N-N_1}{x-x_1}}{\sum_{i=0}^{N}\binom{i}{x_1}\binom{N-i}{x-x_1}} = \frac{\binom{N_1}{x_1}\binom{N-N_1}{x-x_1}}{\binom{N+1}{x+1}}\,.$$

Soit E l'événement que le tirage suivant donnera un individu de la première sous-population, p étant la probabilité de E. Alors

$$P(E|N_1, x_1) = \frac{N_1 - x_1}{N - x}.$$

Donc

$$\begin{aligned} P(E, N_1|x_1) &= \frac{N_1 - x_1}{N - x}\frac{\binom{N_1}{x_1}\binom{N-N_1}{x-x_1}}{\binom{N+1}{x+1}} \\ &= \frac{x_1 + 1}{N - x}\frac{\binom{N_1}{x_1+1}\binom{N-N_1}{x-x_1}}{\binom{N+1}{x+1}}, \end{aligned}$$

et

$$p = P(E|x_1) = \frac{x_1 + 1}{N - x}\frac{\binom{N+1}{x+2}}{\binom{N+1}{x+1}} = \frac{x_1 + 1}{x + 2},$$

qui est indépendant de N. Par conséquent, la loi prévisionnelle de l'appartenance du $(x+1)$-ième tirage est une loi de Bernoulli, $\mathscr{B}(1, (x_1 + 1)/(x + 2))$.

Laplace, considérant le cas particulier $x = x_1$, déduit sa règle de succession : *Si n premiers tirages donnent tous un élément de la même sous-population, la probabilité que le tirage suivant donne à nouveau un élément de cette population est* $\frac{n+1}{n+2}$. Une conséquence de la règle de succession de Laplace est que la probabilité que toute la population soit du même type que les n premières observations est $\frac{n+1}{N+1}$. Certains critiquent[31] cette règle de succession comme étant biaisée en faveur de la sous-population la plus importante, car les populations rares ne seront pas détectées (voir aussi Popper, 1983). Au contraire, Jeffreys (1961, Section 3.2.2) soutient que, au moins dans le domaine de la physique, cette règle conduit assez souvent à rejeter les lois considérées.

4.3.2 Le problème du tramway

Jeffreys (1961) pose le problème suivant, qu'il attribue à Neyman :

> *"Une personne voyageant dans un pays étranger doit changer de train à un embranchement et aller dans une ville qui lui est totalement inconnue. Elle n'en connaît pas la taille. La première chose qu'elle y voit est un tramway numéroté* 100. *Que peut-elle en déduire sur le nombre de tramways dans la ville ? On peut supposer que les tramways sont numérotés en ordre croissant à partir de* 1*."*

Clairement, ce problème a des applications moins anecdotiques. Par exemple, on peut lui rattacher une partie des problèmes de *coïncidence* décrits dans Diaconis et Mosteller (1989).

Exemple 4.17. Soit un phénomène *cyclique* de période inconnue T et à K états possibles (crises boursières, occurrences de comètes, mutations génétiques, feux de signalisation, etc.) ; on observe qu'aux temps t_1 et t_2, le phénomène est dans le même état. Le problème inférentiel est de déduire T de l'observation de la différence $t_2 - t_1$. ‖

Dans le cas du problème du tramway, le nombre N de lignes peut prendre les valeurs $1, 2, \ldots$ Il est préférable de considérer une loi non informative de la forme

$$\pi(N) = \frac{1}{N},$$

plutôt qu'une loi uniforme sur $\mathbb{N}^*$, car N peut s'interpréter comme un paramètre d'échelle. (De plus, la loi a priori uniforme ne donne pas une loi a posteriori proprement définie.) Si T est le numéro relevé, il est supposé distribué selon la loi uniforme

$$f(t|N) = P(T = t|N) = \frac{1}{N} \qquad (t = 1, 2, \ldots, N).$$

[31] Il peut sembler curieux de critiquer un résultat mathématique ! La critique porte en fait sur le choix de la loi a priori, voire de l'axiomatique bayésienne.

Ainsi,

$$\pi(N|T) \propto \frac{1}{N^2}\mathbb{I}_{(N\geq T)}$$

et

$$P^\pi(N \geq n_0|T) = \frac{\sum_{n=n_0}^{+\infty} 1/n^2}{\sum_{n=T}^{+\infty} 1/n^2} \approx \frac{\int_{n_0}^{+\infty}(1/x^2)dx}{\int_{T}^{+\infty}(1/x^2)dx} = \frac{T}{n_0}.$$

Dans ce cas, la médiane a posteriori est approximativement $N^\pi(T) \approx 2T$, estimateur communément retenu pour le problème du tramway. En fait, notons que la moyenne de T conditionnellement à N est $\frac{N-1}{2} \approx \frac{N}{2}$.

4.3.3 Modèles de capture-recapture

Lorsqu'on travaille avec une *loi hypergéométrique* $\mathscr{H}(N, n, p)$, le paramètre d'intérêt est le plus souvent p comme dans le cas de la règle de succession de Laplace, mais il peut aussi arriver que la taille de la population, N, soit inconnue et qu'on cherche à l'estimer. Plus généralement, dans les cas où le recensement d'une population est impossible (ou trop coûteux), il faut trouver une méthode d'estimation de la taille de cette population.

Exemple 4.18. Sur une île de Terre-Neuve vit une harde de cerfs isolée de tout prédateur. Pour éviter que les cerfs ne rompent l'équilibre écologique de l'île, il est nécessaire de réguler cette population en maintenant un nombre de cerfs inférieur à quarante. Un recensement annuel de tous les cerfs prendrait cependant trop de temps. ‖

On pourrait mentionner plusieurs exemples en biologie, sociologie, psychologie, météorologie, écologie, etc., où une évaluation statistique de la taille de la population est nécessaire. Par exemple, les méthodes de capture-recapture exposées ici sont utilisées dans des recensements en France comme aux États-Unis pour dénombrer certaines populations *sous-comptabilisées*, car mal estimées par les techniques habituelles de recensement, comme les populations nomades, les sans-abri ou les immigrants illégaux[32]. L'approche habituelle est appelée *capture-recapture*, car elle consiste à observer au moins deux échantillons successifs de la population d'intérêt et a été d'abord utilisée en biologie animale, où les individus sont effectivement capturés ; voir Seber (1983, 1986) et Pollock (1991) pour une présentation générale.

Dans cette section, nous utilisons le cadre général de Wolter (1986), qui montre que la plupart des modèles de capture-recapture peuvent être décrits par une distribution multinomiale pour chaque individu i ($1 \leq i \leq N$) dans la population. Le Tableau 4.2 donne les probabilités de capture, avec $p^i_{11} + p^i_{12} + p^i_{21} + p^i_{22} = 1$. Par exemple, p^i_{12} représente la probabilité d'être capturé dans

[32] Un exemple frappant de l'efficacité de cette méthode est donné dans McKeganey *et al.* (1992) pour l'estimation du nombre de prostituées dans la ville de Glasgow.

Tab. 4.2. Paramètres de probabilité pour une expérience de capture-recapture.

		Échantillon 2	
		capturé	manqué
Échantillon 1	capturé	p_{11}^i	p_{12}^i
	manqué	p_{21}^i	p_{22}^i

Tab. 4.3. Partition de la population selon le modèle du Tableau 4.2.

		Échantillon 2	
		capturé	manqué
Échantillon 1	capturé	n_{11}	n_{12}
	manqué	n_{21}	n_{22}

le premier échantillon seulement. Après les deux expériences de capture, la population est divisée en quatre sous-populations comme le montre le Tableau 4.3, avec $n_{11}+n_{12}+n_{21}+n_{22} = N$ (la quatrième taille d'échantillon n_{22} étant inconnue). Pour le modèle le plus simple, dit *uniforme*, chaque individu a la même probabilité p d'être capturé dans les deux expériences. Par conséquence, $p_{11} = p^2$, $p_{12} = p_{21} = p(1-p)$ et $p_{22} = (1-p)^2$. La vraisemblance peut s'écrire

$$L(N, p|n_{11}, n_{12}, n_{21}) = \binom{N}{n_{11}\ n_{21}\ n_{21}} p^{n_\cdot}(1-p)^{2N-n_\cdot},$$

où $n_\cdot = 2n_{11} + n_{12} + n_{21}$ est le nombre total d'individus capturés et

$$\binom{N}{n_{11}\ n_{12}\ n_{21}} = \frac{N!}{n_{11}!\, n_{21}!\, n_{12}!\, n_{22}!}$$

est le *coefficient multinomial*. Pour $\pi(N, p) = \pi(N)\pi(p)$ avec $\pi(p)$ une distribution $\mathscr{B}e(\alpha, \beta)$, la loi a posteriori conditionnelle sur p est

$$\pi(p|N, n_{11}, n_{12}, n_{21}) \propto p^{\alpha+n_\cdot-1}(1-p)^{\beta+2N-n_\cdot-1},$$

c'est-à-dire

$$p|N, n_\cdot \sim \mathscr{B}e(\alpha + n_\cdot, \beta + 2N - n_\cdot).$$

Malheureusement, la loi a posteriori marginale de N est assez compliquée. Par exemple, si $\pi(N) = 1$, elle satisfait

$$\pi(N|n_\cdot) \propto \binom{N}{n_+} \frac{B(\alpha + n_\cdot, \beta + 2N - n_\cdot)}{B(\alpha, \beta)}, \tag{4.8}$$

où $n_+ = n_{11}+n_{12}+n_{21}$ est le nombre d'individus capturés qui sont différents. Cette distribution est appelée parfois loi *bêta-Pascal* (voir Raiffa et Schlaifer, 1961), mais elle n'admet pas d'expression explicite. La même difficulté a lieu lorsque $\pi(N) = 1/N$ comme dans Castledine (1981) ou si $\pi(N)$ est une loi de Poisson $\mathscr{P}(\lambda)$ comme dans Raftery (1988), George et Robert (1992) et

Dupuis (1995a,b). Bien entendu, N prenant des valeurs entières, il est toujours possible de calculer le facteur de normalisation dans (4.8) en sommant sur N. Mais, outre le temps requis pour le calcul, les erreurs d'approximation peuvent devenir importantes quand N et n_+ prennent des valeurs élevées. Notons que, pour la loi a priori de Poisson sur N, on a

$$N - n_+ | n_+, p \sim \mathscr{P}((1-p)^2\lambda),$$

donc les distributions a posteriori conditionnelles sont "accessibles" (le Chapitre 6 utilise cette propriété). Les extensions du modèle uniforme sont décrites dans Wolter (1986), George et Robert (1992) et Dupuis (1995a,b).

Un modèle plus simple utilisé dans un cadre de capture-recapture est le modèle hypergéométrique, dit aussi *modèle de Darroch* (Darroch, 1958), dans lequel les tailles des deux échantillons $n_1 = n_{11} + n_{12}$ et $n_2 = n_{11} + n_{21}$ sont fixées. Dans ce cas, la description ci-dessus ne s'applique plus et la seule variable aléatoire qui reste est n_{11}, de loi $\mathscr{H}(N, n_2, \frac{n_1}{N})$. En effet, les valeurs n_1 et n_2 ne sont pas déterminées à l'avance, mais sont plutôt déterminées par un critère d'arrêt généralement inconnu. Cependant, si la loi a priori sur N est non informative et de support discret, le calcul des estimateurs de Bayes est du même ordre de complexité. Néanmoins, le modèle de Darroch peut s'écrire comme un cas particulier du modèle de Wolter (voir l'Exercice 4.35), ce qui permet d'utiliser les mêmes techniques d'approximation développées pour le modèle de Wolter dans ce cadre (voir le Chapitre 6).

Pour le modèle de Darroch, l'estimateur classique de N est l'estimateur du maximum de vraisemblance

$$\hat{N} = \frac{n_1}{(n_{11}/n_2)},$$

qui égalise la proportion dans la population (n_1/N) et la proportion dans l'échantillon (n_{11}/n_2). Cet estimateur présente un inconvénient majeur : il ne peut pas être utilisé lorsque $n_{11} = 0$. Il faut alors de nouveau tirer n_3 individus et observer n_{22} individus déjà présents dans le premier ou le second échantillon. Puisque le nombre d'individus marqués augmente avec le nombre d'échantillons, la probabilité de n'observer que des nouveaux individus à chaque tirage diminue. Il est cependant peu raisonnable de réclamer un échantillon supplémentaire alors que l'objectif initial du modèle statistique était de réduire les coûts d'échantillonnage.

Une analyse bayésienne ne souffre pas de ce défaut, car elle arrive à une conclusion même lorsque $n_{11} = 0$. À partir d'une distribution a priori[33] π sur N, il est facile de calculer la loi a posteriori $\pi(N = n|n_{11})$ et de mener une inférence sur N.

[33]Cette loi a priori aura une influence importante sur l'inférence résultante si n_{11} est petit. Voir l'Exemple 3.1 pour une illustration de détermination de loi a priori dans un contexte réaliste.

Exemple 4.19. (Suite de l'Exemple 4.18) Les règles de natalité et de mortalité des cerfs impliquent que le nombre de cerfs varie entre trente-six et cinquante. Une étude biologique plus approfondie sur l'espérance de vie des cerfs peut certainement aider à construire un modèle de loi a priori sur N, mais nous utiliserons ici une distribution uniforme sur $\{36, \ldots, 50\}$. Si on observe $n_1 = n_2 = 5$, la formule de Bayes,

$$\pi(N = n|n_{11}) = \frac{\binom{n_1}{n_{11}}\binom{n_2}{n_2 - n_{11}} / \binom{n}{n_2} \pi(N = n)}{\sum_{k=36}^{50} \binom{n_1}{n_{11}}\binom{n_2}{n_2 - n_{11}} / \binom{k}{n_2} \pi(N = k)},$$

permet d'obtenir le Tableau 4.4, qui fournit la loi a posteriori de N.

Puisque la loi a posteriori complète de N est disponible, nous pouvons calculer la moyenne, la médiane et le mode a posteriori de N (ou tout autre estimateur de Bayes). Le Tableau 4.5 donne les espérances a posteriori pour les différentes valeurs de n_{11} (on les comparera avec l'estimateur classique $25/n_{11}$ pour $n_{11} \neq 0$, qui varie beaucoup plus avec n_{11}).

Tab. 4.4. Loi a posteriori de la taille de la population de cerfs, $\pi(N|n_{11})$.

	n_{11}					
N	0	1	2	3	4	5
36	0.058	0.072	0.089	0.106	0.125	0.144
37	0.059	0.072	0.085	0.098	0.111	0.124
38	0.061	0.071	0.081	0.090	0.100	0.108
39	0.062	0.070	0.077	0.084	0.089	0.094
40	0.063	0.069	0.074	0.078	0.081	0.082
41	0.065	0.068	0.071	0.072	0.073	0.072
42	0.066	0.068	0.067	0.067	0.066	0.064
43	0.067	0.067	0.065	0.063	0.060	0.056
44	0.068	0.066	0.062	0.059	0.054	0.050
45	0.069	0.065	0.060	0.055	0.050	0.044
46	0.070	0.064	0.058	0.051	0.045	0.040
47	0.071	0.063	0.056	0.048	0.041	0.035
48	0.072	0.063	0.054	0.045	0.038	0.032
49	0.073	0.062	0.052	0.043	0.035	0.028
50	0.074	0.061	0.050	0.040	0.032	0.026

Tab. 4.5. Espérance a posteriori de la taille de la population de cerfs, $\mathbb{E}[N|n_{11}]$.

n_{11}	0	1	2	3	4	5
$\mathbb{E}(N|n_{11})$	43.32	42.77	42.23	41.71	41.23	40.78

Si, au lieu d'une erreur quadratique, nous utilisons le coût

$$\mathrm{L}(N,\delta) = \begin{cases} 10(\delta - N) & \text{si } \delta > N, \\ N - \delta & \text{sinon,} \end{cases} \tag{4.9}$$

afin d'éviter une surestimation du nombre de cerfs (ce qui aurait des conséquences plus dramatiques pour l'avenir de la harde qu'une sous-estimation), l'estimateur de Bayes est le fractile (1/11) de $\pi(N|n_{11})$, donné dans le Tableau 4.6 pour différentes valeurs de n_{11}. Notons que, dans ce cas, les estimateurs prennent nécessairement des valeurs entières. ‖

Tab. 4.6. Estimateur de la taille de la population de cerfs sous une perte asymétrique (4.9).

n_{11}	0	1	2	3	4	5
$\delta^\pi(n_{11})$	37	37	37	36	36	36

Une application bayésienne très intéressante de l'inférence du modèle de capture-recapture est donnée par Mosteller et Wallace (1984). Elle concerne *l'authentification d'œuvres par la linguistique statistique* lorsque l'origine de certaines de ces œuvres est incertaine. Par exemple, Mosteller et Wallace (1984) étudient les *Federalist Papers*, une collection d'articles écrits en 1787 afin de soutenir la nouvelle Constitution des États-Unis. Douze de ces articles sont attribués soit à Hamilton, soit à Madison. À partir d'écrits authentifiés de ces deux auteurs, Mosteller et Wallace (1984) calculent la fréquence des trente mots les plus courants et, en utilisant l'approche du modèle de capture-recapture, déduisent que les douze articles auraient été écrits par Madison. Efron et Thisted (1976) ont aussi utilisé cette méthode dans l'étude du vocabulaire de Shakespeare pour authentifier plus tard dans Thisted et Efron (1987) un poème récemment découvert comme ayant été effectivement écrit par Shakespeare.

4.4 Le cas particulier du modèle normal

4.4.1 Introduction

Lorsque Gauss introduisit la distribution normale aux alentours de 1810, Laplace estima qu'il s'agissait en fait de la loi d'erreur *idéale* (voir l'Exemple 1.12). Par la suite, s'appuyant sur le Théorème Central Limit, les statisticiens de la première moitié du XIXième siècle se référaient presque toujours à la distribution normale (Stigler, 1986). Il y a, bien sûr, de nombreux phénomènes pour lesquels un modèle normal n'est pas applicable, mais ce dernier reste

considérablement utilisé, en particulier en économétrie et dans des domaines où on peut justifier l'approximation du Théorème Central Limit (physique particulaire, etc.). En réalité, l'approximation normale est souvent justifiée par des raisons asymptotiques (voir aussi Cox et Reid, 1987). Il est donc intéressant d'étudier en détail cette distribution particulière d'un point de vue bayésien.

Pour l'observation d'une distribution normale multivariée, $\mathcal{N}_p(\theta, \Sigma)$, de matrice de covariance connue Σ, la loi conjuguée est aussi normale, $\mathcal{N}_p(\mu, A)$, et la loi a posteriori $\pi(\theta|x)$ est

$$\mathcal{N}_p\left(x - \Sigma(\Sigma + A)^{-1}(x - \mu), (A^{-1} + \Sigma^{-1})^{-1}\right).$$

Sous un coût quadratique, l'estimateur de Bayes est alors la moyenne a posteriori

$$\begin{aligned}\delta^\pi(x) &= x - \Sigma(\Sigma + A)^{-1}(x - \mu)\\ &= \left(\Sigma^{-1} + A^{-1}\right)^{-1}\left(\Sigma^{-1}x + A^{-1}\mu\right);\end{aligned}$$

notons que $\delta^\pi(x)$ peut s'écrire comme une combinaison convexe de l'observation, x, et de la moyenne a priori, μ, les poids étant proportionnels à l'inverse de la matrice de covariance.

Plus l'information a priori sur θ est précise, plus proche de μ est l'estimateur de Bayes. Notons aussi que l'information a priori (resp., l'observation de x) apporte une réduction de la variance de Σ (respectivement, de A) à $\left(\Sigma^{-1} + A^{-1}\right)^{-1}$. Pour des observations répétées du modèle normal ci-dessus, $x_1, ..., x_n$, la statistique exhaustive

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \sim \mathcal{N}_p\left(\theta, \frac{1}{n}\Sigma\right)$$

étend directement l'analyse précédente.

Une critique déjà évoquée dans le Chapitre 3 est que les lois a priori conjuguées normales ne sont pas assez robustes et qu'il serait préférable d'utiliser une loi de Student pour $\pi(\theta)$. La loi de Cauchy, cas limite d'une loi de Student, peut alors être utilisée en raison de ses queues plus lourdes, mais elle empêche encore un calcul exact (voir l'Exemple 4.7), même si Angers (1992) propose une solution analytique reposant sur des fonctions confluentes hypergéométriques.

4.4.2 Estimation de la variance

Dans la plupart des cas, la variance du modèle est partiellement ou totalement inconnue. Il est alors nécessaire de considérer des lois a priori pour le paramètre (θ, Σ). Si la variance est connue à une constante multiplicative près,

σ^2, il est généralement possible de revenir à un cadre unidimensionnel, c'est-à-dire lorsque $x_1, \ldots, x_n$ sont i.i.d. $\mathcal{N}(\theta, \sigma^2)$, pour des raisons d'exhaustivité. (Le cas particulier où seul σ^2 est inconnu est traité dans les Tableaux 3.4 et 4.4.) Si nous définissons les statistiques $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ et $s^2 = \sum_{i=1}^n (x_i - \bar{x})^2$, la vraisemblance peut s'écrire

$$\ell(\theta, \sigma \mid \bar{x}, s^2) \propto \sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}\left\{s^2 + n\,(\bar{x} - \theta)^2\right\}\right]$$

et l'estimateur de Bayes ne dépend que de $\bar{x}$ et s^2. Nous indiquons dans l'Exemple 3.30 que la *loi de Jeffreys* pour ce modèle est $\pi^*(\theta, \sigma) = 1/\sigma^2$ et mentionnons qu'il est préférable de considérer la loi alternative $\tilde{\pi}(\theta, \sigma) = 1/\sigma$ pour des raisons d'invariance. Dans ce cas,

$$\ell(\theta, \sigma \mid \bar{x}, s^2)\tilde{\pi}(\theta, \sigma) \propto \sigma^{-n-1} \exp\left[-\frac{1}{2\sigma^2}\left\{s^2 + n\,(\bar{x} - \theta)^2\right\}\right] . \tag{4.10}$$

Donc,

Proposition 4.20. *Si $x_1, \ldots, x_n$ sont i.i.d. $\mathcal{N}(\theta, \sigma^2)$, la loi a posteriori de (θ, σ) associée à $\tilde{\pi}$ est*

$$\begin{aligned} \theta|\sigma, \bar{x}, s^2 &\sim \mathcal{N}\left(\bar{x}, \frac{\sigma^2}{n}\right), \\ \sigma^2|\bar{x}, s^2 &\sim \mathcal{I}G\left(\frac{n-1}{2}, \frac{s^2}{2}\right). \end{aligned} \tag{4.11}$$

L'équation (4.11) définit vraiment la loi a posteriori de (θ, σ^2), car elle fournit la loi marginale de σ^2 et la loi de θ conditionnellement à σ^2. La démonstration de cette proposition est une conséquence directe de (4.10), puisque

$$\tilde{\pi}(\theta, \sigma^2|\bar{x}, s^2) \propto \sigma^{-1} e^{-n(\bar{x}-\theta)^2/2\sigma^2} \sigma^{-n} e^{-s^2/2\sigma^2} \sigma^{-1},$$

et la loi *gamma inverse* $\mathcal{I}G(\alpha, \beta)$ a pour densité

$$\pi(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha) x^{\alpha+1}} e^{-\beta/x} \mathbb{I}_{(0,+\infty)}(x). \tag{4.12}$$

Par conséquent, la loi a posteriori marginale de σ^2 est du même type que lorsque θ est connu. En revanche, la loi marginale a posteriori de θ diffère, car il vient de (4.11) que

$$\tilde{\pi}(\theta|\bar{x}, s^2) \propto \left\{s^2 + n(\bar{x} - \theta)^2\right\}^{-n/2},$$

c'est-à-dire

$$\theta|\bar{x}, s^2 \sim \mathcal{T}_1\left(n-1, \bar{x}, \frac{s^2}{n(n-1)}\right). \tag{4.13}$$

Pour la loi de Jeffreys, π^*, l'équivalent de (4.13) est une loi de Student à n degrés de liberté, qui est toujours définie, tandis que (4.13) n'est définie que pour $n \geq 2$. (Notons que l'exclusion de $n = 1$ pourrait s'interpréter comme un argument supplémentaire en faveur de $\tilde{\pi}$, car, dans un cadre non informatif, il paraît difficile de proposer une inférence sur le paramètre (θ, σ) tout entier avec une *seule* observation.)

Les lois a posteriori conjuguées ont naturellement la même forme que (4.11). Ces lois présentent cependant une curieuse particularité, à savoir que θ et σ^2 ne sont pas indépendants a priori. Par conséquent, la loi a priori de la moyenne θ dépend de la précision associée à la mesure de la moyenne. Certains cadres d'application peuvent justifier cette dépendance[34], mais ceci n'est pas vrai pour tous les problèmes d'estimation et cette loi peut encore moins être considérée comme une loi a priori représentative standard (voir Berger, 2000). Cependant, ces critiques subjectives ne se doublent pas de propriétés particulièrement négatives des estimateurs résultants.

Soit alors

$$\pi(\theta, \sigma^2) = \pi_1(\theta|\sigma^2)\pi_2(\sigma^2),$$

où π_1 est une distribution normale $\mathcal{N}(\mu, \sigma^2/n_0)$ et π_2 est une loi gamma inverse $\mathcal{I}G(\nu/2, s_0^2/2)$. La loi a posteriori satisfait

$$\begin{aligned}\pi(\theta, \sigma^2|x) &\propto \sigma^{-n-\nu-3} \exp\left\{-\frac{1}{2}\left[s^2 + s_0^2 + n_0(\theta-\mu)^2 + n(\bar{x}-\theta)^2\right]/\sigma^2\right\} \\ &= \sigma^{-n-\nu-3} \exp\left\{-\frac{1}{2}\left[s_1^2 + n_1(\theta-\theta_1)^2\right]/\sigma^2\right\},\end{aligned}$$

où

$$\begin{aligned}n_1 &= n + n_0, \quad \theta_1 = \frac{1}{n_1}\left(n_0\theta_0 + n\bar{x}\right), \\ s_1^2 &= s^2 + s_0^2 + \left(n_0^{-1} + n^{-1}\right)^{-1}(\theta_0 - \bar{x})^2.\end{aligned}$$

Ces lois sont en réalité conjuguées car

$$\begin{aligned}\pi\left(\theta|\bar{x}, s^2, \sigma\right) &\propto \frac{1}{\sigma}\exp\left\{-\frac{n_1(\theta-\theta_1)^2}{2\sigma^2}\right\}, \\ \pi\left(\sigma^2|\bar{x}, s^2\right) &\propto \sigma^{-n-\nu-2}\exp\left\{-s_1^2/2\sigma^2\right\}.\end{aligned}$$

Comme dans le cas non informatif, la loi a posteriori marginale de θ est une loi de Student. Notons que, sauf lorsque π est construit à partir d'observations précédentes (ou virtuelles), n_0 n'est pas une taille d'échantillon ; n_0/n caractérise plutôt la précision de la détermination de la loi a priori, relativement à la précision des observations. En général, n_0 est plus petit que la

[34] Lorsque la loi a priori est construite à partir d'observations passées, il est logique que la variance a priori de θ dépende de σ^2 (conditionnellement).

taille d'échantillon n. Notons aussi que, si n_0/n tend vers 0, nous obtenons le cas limite $\theta|\bar{x}, \sigma^2 \sim \mathcal{N}(\bar{x}, \sigma^2/n)$, correspondant à la loi a posteriori associée à la loi a priori de Jeffreys. Voici donc un exemple supplémentaire du fait que les lois non informatives se présentent souvent comme des limites de lois conjuguées.

L'inférence statistique fondée sur la loi conjuguée ci-dessus nécessite une détermination précise des hyperparamètres $(\theta_0, s_0^2, n_0, \nu)$, afin d'obtenir l'expression des estimateurs de Bayes. Si la détermination de θ_0 et n_0 est plutôt classique, il est généralement plus difficile d'avoir une information a priori sur σ^2. Rappelons que, si $\sigma^2 \sim \mathcal{I}G(\nu/2, s_0^2/2)$, les deux premiers moments sont donnés par $(\nu > 4)$

$$\mathbb{E}^\pi\left[\sigma^2\right] = \frac{s_0^2}{\nu - 2}, \quad \text{var}^\pi(\sigma^2) = \frac{2s_0^4}{(\nu-2)^2(\nu-4)}.$$

Ces formules peuvent alors s'utiliser pour modéliser une information a priori sous une forme conjuguée, c'est-à-dire pour déterminer s_0^2 et ν.

Lorsque le paramètre (θ, Σ) est totalement inconnu, il reste possible de construire des lois a priori conjuguées. Pour n observations $x_1, \ldots, x_n$ de $\mathcal{N}_p(\theta, \Sigma)$, une statistique exhaustive est

$$\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i, \qquad S = \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^t,$$

et

$$\ell(\theta, \Sigma|\bar{x}, S) \propto |\Sigma|^{-n/2} \exp -\frac{1}{2}\left\{n(\bar{x}-\theta)^t\Sigma^{-1}(\bar{x}-\theta) + \text{tr}(\Sigma^{-1}S)\right\}.$$

La forme de la fonction de vraisemblance suggère alors les lois conjuguées suivantes :

$$\begin{aligned} \theta|\Sigma &\sim \mathcal{N}_p\left(\mu, \frac{\Sigma}{n_0}\right), \\ \Sigma^{-1} &\sim \mathcal{W}_p(\alpha, W), \end{aligned} \tag{4.14}$$

où $\mathcal{W}_p$ indique la loi de Wishart, définie dans l'Exercice 3.21. Les lois a posteriori sont alors

$$\begin{aligned} \theta|\Sigma, \bar{x}, S &\sim \mathcal{N}_p\left(\frac{n_0\mu + n\bar{x}}{n_0 + n}, \frac{\Sigma}{n_0 + n}\right), \\ \Sigma^{-1}|\bar{x}, S &\sim \mathcal{W}_p\left(\alpha + n, W_1(\bar{x}, S)\right), \end{aligned}$$

avec

$$W_1(\bar{x}, S)^{-1} = W^{-1} + S + \frac{nn_0}{n + n_0}(\bar{x} - \mu)(\bar{x} - \mu)^t.$$

Notons que ce cas multidimensionnel est la généralisation du cas unidimensionnel considéré au-dessus, car la loi de Wishart $\mathcal{W}_p$ est la généralisation

en dimension p d'une loi du khi deux. Rappelons ici que les deux premiers moments de $\Xi = (\xi_{ij}) \sim \mathscr{W}_p(\alpha, W)$ sont

$$\mathbb{E}[\Xi] = \alpha W, \qquad \text{var}(\xi_{ij}) = 2\alpha w_{ij}^2,$$

et que les hyperparamètres de la loi a priori de Σ peuvent se calculer à partir de

$$\mathbb{E}[\Sigma] = \frac{W^{-1}}{\alpha - p - 1}, \qquad \text{var}(\sigma_{ij}) = \frac{2(w^{ij})^2}{(\alpha - p - 3)(\alpha - p - 1)^2},$$

pour $\Sigma^{-1} \sim \mathscr{W}_p(\alpha, W)$ et $W^{-1} = (w^{ij})$ (Eaton, 1982, Anderson, 1984).

Dans ce cadre, la loi de Jeffreys est aussi un cas limite des lois conjuguées, car Geisser et Cornfield (1963) ont montré qu'elle vaut

$$\pi^{\text{J}}(\theta, \Sigma) = \frac{1}{|\Sigma|^{(p+1)/2}},$$

et donc qu'elle correspond à la limite de lois de Wishart $\mathscr{W}_p(\alpha, W)$ pour Σ^{-1} lorsque W^{-1} tend vers $\mathbf{O}$ et α vers 0. En effet, la densité de Σ lorsque $\Sigma^{-1} \sim \mathscr{W}_p(\alpha, W)$ est

$$f(\Sigma|\alpha, W) \propto |\Sigma|^{-(\alpha+p+1)/2} \exp\left\{-\frac{1}{2}\text{tr}(W^{-1}\Sigma^{-1})\right\}$$

(Anderson, 1984).

4.4.3 Modèles linéaires et G-priors

Le modèle standard de régression,

$$y = X\beta + \epsilon, \tag{4.15}$$

avec $\epsilon \sim \mathscr{N}_k(0, \Sigma)$, $\beta \in \mathbb{R}^p$, peut s'analyser de la même façon que dans la partie précédente lorsque la matrice de covariance Σ est connue, si on travaille conditionnellement à X. En effet, une statistique exhaustive est alors

$$\hat{\beta} = (X^t\Sigma^{-1}X)^{-1}X^t\Sigma^{-1}y\,,$$

estimateur du maximum de vraisemblance et des moindres carrés de β. Celui-ci est distribué selon une loi $\mathscr{N}_p(\beta, (X^t\Sigma^{-1}X)^{-1})$.

Lindley et Smith (1972) ont étudié des lois conjuguées du type

$$\beta \sim \mathscr{N}_p(A\theta, C),$$

où $\theta \in \mathbb{R}^q$ ($q \leq p$). Dans ce modèle, la matrice de régression X est considérée comme constante. En d'autres termes, l'inférence est faite conditionnellement à X. (Habituellement, X est aussi partiellement aléatoire, mais ce conditionnement est justifié par le principe de vraisemblance du moment que la loi de

X ne dépend pas des paramètres du modèle de régression.) Par conséquent, A, C, ou θ peuvent dépendre de X (voir ci-dessous pour l'exemple des *lois a priori simplifiées* de Zellner, 1971). Lorsque la nature stochastique de X *doit* être considérée, l'approche habituelle est d'étudier un modèle à effets aléatoires,

$$y = X_1\beta_1 + X_1X_2\beta_2 + \epsilon,$$

qui peut se décomposer en

$$\begin{aligned} y|\theta_1 &\sim \mathcal{N}_k(X_1\theta_1, \Sigma_1), \\ \theta_1|\theta_2 &\sim \mathcal{N}_p(X_2\theta_2, \Sigma_2), \end{aligned}$$

avec pour loi a priori

$$\theta_2|\theta_3 \sim \mathcal{N}_q(X_3\theta_3, \Sigma_3).$$

Smith (1973) analyse ce modèle et montre que

$$\theta_1|y, \theta_3 \sim \mathcal{N}_p(\theta_1^*, D_1),$$

avec

$$\begin{aligned} \theta_1^* &= D_1\left[\hat{D}_1^{-1}\hat{\theta}_1 + (\Sigma_2 + X_2\Sigma_3X_2^t)^{-1}X_2X_3\theta_3\right], \\ D_1^{-1} &= \hat{D}_1^{-1} + (\Sigma_2 + X_2\Sigma_3X_2^t)^{-1}, \end{aligned}$$

fonction des estimateurs des moindres carrés classiques

$$\hat{D}_1^{-1} = X_2^t\Sigma_1^{-1}X_2, \qquad \hat{\theta}_1 = \hat{D}_1X_2^t\Sigma_1^{-1}y.$$

Par conséquent, l'estimateur de Bayes θ_1^* est une combinaison convexe de l'estimateur des moindres carrés, $\hat{\theta}_1$ et de la moyenne a priori, $X_2X_3\theta_3$.

Nous introduisons ci-dessous un exemple où une structure de variance inconnue permet toujours un calcul analytique des estimateurs de Bayes. Cependant, si la variance Σ est totalement inconnue, il n'est pas possible de construire des lois a priori conjuguées, comme l'avaient remarqué Lindley et Smith (1972). Press (1989) propose une solution dans un cas particulier où des observations *indépendantes* sont disponibles. Dans un cas général, la loi a priori de Jeffreys est de nouveau (Geisser et Cornfield, 1963)

$$\pi^J(\beta, \Sigma) = \frac{1}{|\Sigma|^{(k+1)/2}}.$$

La vraisemblance

$$\ell(\beta, \Sigma|y) \propto |\Sigma|^{-n/2}\exp\left\{-\frac{1}{2}\mathrm{tr}\left[\Sigma^{-1}\sum_{i=1}^{n}(y_i - X_i\beta)(y_i - X_i\beta)^t\right]\right\}$$

suggère alors d'utiliser les lois de Wishart, mais les lois a posteriori marginales sur β ne sont définies que pour des échantillons de taille suffisamment grande

et, de plus, elles ne sont pas explicites, quelle que soit la taille de l'échantillon (voir l'Exercice 4.45).

Dans le cas particulier où la variance du modèle (4.15) est connue à un facteur multiplicatif σ^2 près, il est possible de réécrire le modèle comme $\epsilon \sim \mathcal{N}_k(0, \sigma^2 I_k)$ et l'estimateur des moindres carrés $\hat{\beta}$ a une distribution normale $\mathcal{N}_p(\beta, \sigma^2(X^tX)^{-1})$. Une famille de lois conjuguées pour (β, σ^2) est alors

$$\begin{aligned} \beta|\sigma^2 &\sim \mathcal{N}_p\left(\mu, \frac{\sigma^2}{n_0}(X^tX)^{-1}\right), \\ \sigma^2 &\sim \mathcal{I}G(\nu/2, s_0^2/2), \end{aligned} \tag{4.16}$$

car, si $s^2 = ||y - X\hat{\beta}||^2$, les lois a posteriori sont

$$\begin{aligned} \beta|\hat{\beta}, s^2, \sigma^2 &\sim \mathcal{N}_p\left(\frac{n_0\mu + \hat{\beta}}{n_0+1}, \frac{\sigma^2}{n_0+1}(X^tX)^{-1}\right), \\ \sigma^2|\hat{\beta}, s^2 &\sim \mathcal{I}G\left(\frac{k-p+\nu}{2}, \frac{s^2 + s_0^2 + \frac{n_0}{n_0+1}(\mu-\hat{\beta})^tX^tX(\mu-\hat{\beta})}{2}\right). \end{aligned}$$

En effet,

$$\begin{aligned} &\pi(\beta, \sigma^2|\hat{\beta}, s^2) \propto (\sigma^2)^{-k/2} \exp\left[-\frac{1}{2\sigma^2}\left\{(\beta-\hat{\beta})^tX^tX(\beta-\hat{\beta}) + s^2\right\}\right] \\ &\times \exp\left(-\frac{n_0}{2\sigma^2}(\beta-\mu)^tX^tX(\beta-\mu)\right)\sigma^2)^{-\nu/2-1} \exp\left(-\frac{s_0^2}{2\sigma^2}\right) \\ &\propto (\sigma^2)^{-p/2} \exp\left\{-\frac{n_0+1}{2\sigma^2}\left(\beta - \frac{n_0\mu+\hat{\beta}}{n_0+1}\right)^t X^tX\left(\beta - \frac{n_0\mu+\hat{\beta}}{n_0+1}\right)\right\} \\ &\times\sigma^{-(k-p+\nu+2)} \exp\left[-\frac{1}{2\sigma^2}\left\{s_0^2 + s^2 + \frac{n_0}{n_0+1}(\mu-\hat{\beta})^tX^tX(\mu-\hat{\beta})\right\}\right]. \end{aligned}$$

Bien que (4.16) ne soit qu'un cas particulier de loi conjuguée, plusieurs critiques se sont élevées contre ce choix, développé par Zellner (1971, 1986b) sous le nom de *G-priors* ou *a priori simplifiés*[35]. Ces critiques ne s'adressent pas pour la plupart au problème de l'aspect réducteur d'un modèle conjugué, un argument assez légitime déjà évoqué au Chapitre 3, mais plutôt à la dépendance de la loi a priori à X. On peut soutenir que X est aussi une variable aléatoire et par conséquent qu'un modèle a priori ne devrait pas dépendre de X. En fait, les lois a priori alternatives

$$\beta|\sigma \sim \mathcal{N}_p(\beta_0, \sigma^2 A)$$

constituent aussi une famille conjuguée qui est moins critiquable lorsque A est fixé. Cependant, nous considérons que le débat est plutôt vide de sens car :

[35]Le nom de G-priors provient de l'utilisation dans l'article originel du symbole g comme facteur de $\sigma^2(X^tX)^{-1}$ dans (4.16).

(1) Le modèle de régression est entièrement conditionnel aux variables explicatives. La loi a priori (4.16) peut se voir comme une loi a posteriori par rapport à ces variables (ou, pour élargir l'hypothèse habituelle d'indépendance entre les variables explicatives et les erreurs, comme l'hypothèse de l'indépendance bayésienne avec les paramètres). Cette approche est alors justifiée par les points de vue conditionnel et bayésien, le conditionnement étant alors établi en deux étapes.

(2) Un *G-prior* suggère une distribution constante pour la moyenne de y, $\theta = \mathbb{E}_\theta[y|X]$, plutôt que pour β. La loi a priori est alors déterminée par rapport au sous-espace généré par les colonnes de X et non pas par rapport à une base spéciale de ce sous-espace.

(3) Ce modèle est adéquat pour la prise en compte des problèmes de *multicolinéarité*, car il permet d'assigner une grande variance a priori aux composantes affectées par la multicolinéarité (donc plus difficiles à estimer). (Voir Zellner, 1971, Casella, 1985a, ou Steward, 1987, pour des références sur la multicolinéarité.)

(4) Des points de vue pratique et subjectif, la détermination a priori d'une matrice A plutôt que d'un scalaire n_0 nécessite une plus grande quantité d'information a priori. Puisque le recours aux lois conjuguées est caractéristique des cas où l'information a priori est rare et où la détermination des hyperparamètres est assez difficile, l'utilisation de la matrice de covariance $\sigma^2(X^tX)^{-1}/n_0$ évite une détermination probablement irréaliste de A.

Notons de nouveau que ces attaques contre les *G-priors* mentionnent à peine leur désavantage majeur, à savoir que leur choix n'est pas totalement fondé sur l'information a priori. Pour des applications des *G-priors* dans des problèmes de régression, voir Ghosh et Sen (1989) ou Blattberg et George (1991). Voir Bauwens et al. (1999, Chapitre 4) pour des alternatives aux lois a priori conjuguées pour les modèles linéaires, comme les lois a priori *poly-t* (voir la Note 4.7.5 ci-dessous).

4.5 Modèles dynamiques

4.5.1 Introduction

Les modèles dynamiques (ou *de séries temporelles*) apparaissent comme un modèle paramétrique où la distribution des variables observées $x_1, \ldots, x_T$ varie dans le temps, c'est-à-dire

$$f(x_1, \ldots, x_T|\theta) = \prod_{t=1}^{T} f_t(x_t|x_{1:(t-1)}, \theta)\,, \tag{4.17}$$

où $x_{1:(t-1)}$ indique le vecteur des variables précédentes $x_1, \ldots, x_{t-1}$, avec la convention que $x_{1:0}$ est soit vide, soit représente la valeur initiale x_0 d'une suite

d'observations (il est alors implicite dans le terme de gauche de (4.17)). Bien que la représentation (4.17) semble être inutilement restrictive, l'inclusion de composants non observés dans x_t fournit une perspective assez large pour ce modèle, comme cela sera expliqué dans le paragraphe sur les représentations par espace d'état.

Ces modèles sont évidemment des cas spéciaux de modèles paramétriques et, en tant que tels, peuvent donc être traités comme d'autres modèles paramétriques par les outils bayésiens, une fois la loi a priori choisie, suivant les indications fournies dans les sections précédentes. Ils sont isolés dans cette section pour plusieurs raisons : premièrement, il s'agit des modèles les plus couramment utilisés dans des applications allant de la Finance et l'Économie jusqu'aux expériences médicales et l'écologie. La plupart des modèles rencontrés dans la pratique présentent une dimension temporelle qui peut parfois être dissimulée, mais qui le plus souvent doit être prise en compte. C'est clairement le cas pour des données de pollution, comme les niveaux de concentration d'ozone, ou les cours d'action, pour lesquelles la valeur au temps t dépend de la valeur précédente et aussi des valeurs antérieures, par exemple à travers leur *tendance*.

Exemple 4.21. (Suite de l'Exemple 4.6) Le modèle autorégressif AR(1) est plus généralement défini par la loi de x_t conditionnellement à $x_{1:(t-1)}$ $(1 \leq t \leq T)$,

$$x_t = \mu + \varrho(x_{t-1} - \mu) + \epsilon_t \,, \tag{4.18}$$

où ϵ_t est indépendant de $x_{1:(t-1)}$ et suit, par exemple, une loi $\mathcal{N}(0, \sigma^2)$. La distribution de x_t sachant $x_{1:(t-1)}$ ne dépend que de x_{t-1}, ce qui prouve que (x_t) est une *chaîne de Markov* (Meyn et Tweedie, 1993).

La fonction de vraisemblance du modèle AR(1) est alors

$$\sigma^{-T} \exp\left\{ \frac{-1}{2\sigma^2} \sum_{i=1}^{T} (x_t - \mu + \varrho(x_{t-1} - \mu))^2 \right\}$$

et dépend donc de la condition initiale x_0. Soit x_0 est connu et le modèle est alors conditionnel à x_0, soit x_0 a été intégré en prenant pour loi a priori $\pi(x_0|\theta)$ et x_0 est alors un paramètre additionnel du modèle. Par exemple, si $x_0 = 0$, il est simple de voir que $\mathbb{E}_\theta[x_t] = 0$ et que $\text{var}(x_t) = \varrho^2 \text{var}(x_{t-1}) + \sigma^2$, donc, si $\varrho^2 \neq 1$,

$$\text{var}(x_t) = \frac{1 - \varrho^{2t}}{1 - \varrho^2} \sigma^2 \,, \tag{4.19}$$

ce qui implique que $\text{var}(x_t)$ converge vers $\sigma^2/(1 - \varrho^2)$ si $\varrho^2 < 1$ et tend vers $+\infty$ sinon. ‖

La seconde motivation pour étudier les modèles dynamiques est que ceux-ci représentent un plus grand défi que les modèles statiques étudiés

précédemment, de par les contraintes de *stationnarité*. Bien que nous ne puissions pas présenter une introduction rigoureuse de la notion de stationnarité pour les processus stochastiques (nous renvoyons les lecteurs à Meyn et Tweedie, 1993, pour une présentation générale des processus de Markov et à Box et Jenkins, 1976 ou Brockwell et Davis, 1998, pour le cas spécial des séries temporelles), rappelons ici qu'un processus (x_t) est *stationnaire* (ou strictement stationnaire) si la distribution de $(x_{t+1}, \ldots, x_{t+d})$ est la même que la distribution de $(x_1, \ldots, x_d)$ pour tout (t, d). Le problème de la stationnarité peut s'illustrer dans le cadre de l'Exemple 4.6 : lorsque $\varrho^2 \geq 1$, non seulement la variance $\text{var}(x_t)$ tend vers l'infini avec t, mais de plus le comportement limite de la chaîne (x_t) ne peut pas être caractérisé. Le processus (x_t) n'a pas de distribution limite, car la chaîne de Markov n'admet pas de distribution *stationnaire*, c'est-à-dire qu'il n'existe pas de densité f telle que, si $x_t \sim f$, $x_{t+1} \sim f$ (Exercice 4.51). Par exemple, si $\varrho = 1$, (x_t) est la *marche aléatoire* dans $\mathbb{R}$ et, en moyenne, elle prend un temps infini pour revenir à l'ensemble d'où elle est partie (Meyn et Tweedie, 1993).

Imposer la *stationnarité* d'un modèle est critiquable du fait que les données elles-mêmes devraient indiquer si le modèle sous-jacent est stationnaire. Cependant, pour des raisons allant de l'asymptotique à la causalité, en passant par l'identifiabilité (voir ci-dessous) et la pratique générale, il est courant d'imposer cette condition, même si l'inférence bayésienne d'un processus non stationnaire peut être conduite en principe (voir la Note 4.7.2). De telles contraintes se traduisent dans la distribution a priori par une restriction sur les valeurs de θ. Par exemple, pour le modèle AR(1) de l'Exemple 4.6, la contrainte est $|\varrho| < 1$. La difficulté pratique est que, pour des modèles plus complexes, les contraintes de stationnarité peuvent devenir beaucoup plus exigeantes et sont même inconnues dans certains cas, comme dans les modèles à seuil généraux (Tong, 1991).

Exemple 4.22. Le modèle AR(p) généralise le modèle AR(1) en augmentant la dépendance sur les valeurs passées, c'est-à-dire $(1 \leq t \leq T)$,

$$x_t - \mu = \sum_{i=1}^{p} \varrho_i (x_{t-i} - \mu) + \epsilon_t \,, \qquad \epsilon_t \sim \mathcal{N}(0, \sigma^2) \,. \tag{4.20}$$

Le processus stochastique défini par (4.20) est alors stationnaire si et seulement si les racines du polynôme

$$\mathcal{P}(x) = 1 - \sum_{i=1}^{p} \varrho_i x^i$$

sont toutes à l'extérieur du cercle unité dans le plan complexe (voir Brockwell et Davis, 1998, Section 3.1). Bien que cette condition soit clairement définie, elle est aussi *implicite* par rapport au vecteur $(\varrho_1, \ldots, \varrho_p)$: pour vérifier qu'un vecteur donné satisfait cette condition, il est nécessaire de trouver les racines

du polynôme $\mathscr{P}$ et de s'assurer qu'elles sont toutes de module plus grand que 1, ou de calculer les autocorrélations partielles (voir la Section 4.5.2) et d'appliquer le *lemme de Schur* pour vérifier qu'elles sont toutes entre -1 et 1. ‖

Exemple 4.23. Un modèle AR(p) à sauts (traduction de *switching AR*) est défini comme un modèle AR(p) dont les paramètres changent dans le temps selon un processus de Markov caché (ou non observé) à espace d'état fini, c'est-à-dire

$$x_t = \sum_{i=1}^{p} \varrho_i(z_t) x_{t-i} + \sigma(z_t)\epsilon_t\,, \quad \epsilon_t \sim \mathscr{N}(0,1)\,, \tag{4.21}$$

où (z_t) est la chaîne de Markov non observée,

$$P(z_t = i | z_{t-1} = j, z_{t-2}, \ldots) = \pi_{j,i}\,, \qquad i,j = 1,\ldots,K\,.$$

Ce modèle a été introduit par Hamilton (1989) comme une façon de représenter des séries avec des dynamiques variant dans le temps, comme la série de la Figure 4.2 qui est une transformation des cours de l'action IBM entre 1992 et 1997. Une difficulté avec le modèle (4.21) est qu'il n'existait pas de condition nécessaire et suffisante de stationnarité lorsque le nombre d'états K de la chaîne de Markov cachée (z_t) est plus grand que 2, jusqu'aux développements récents de Francq et Zakoïan (2001) et Yao et Attali (2000). ‖

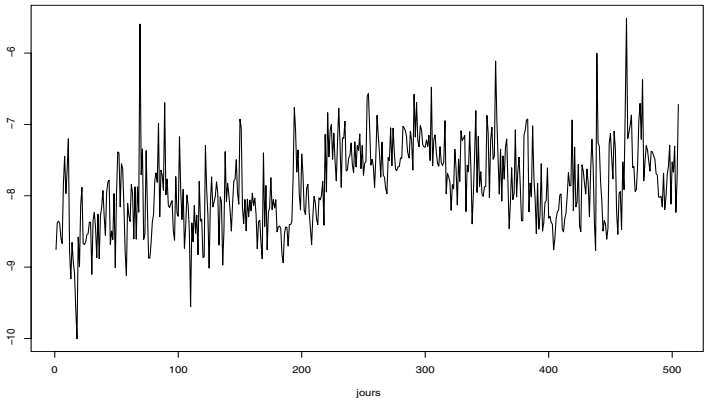

Fig. 4.2. Tracé du logarithme des cours de l'action IBM sur la période 1992-1997.

Nous développons dans les Sections 4.5.2-4.5.4 quelques caractéristiques des modèles dynamiques standard, à savoir, les modèles AR, MA et ARMA,

en nous concentrant sur les problèmes de représentation et de modélisation a priori sous condition de stationnarité. Les Notes 4.7.3 et 4.7.4 présentent deux autres modèles dynamiques souvent rencontrés dans la pratique. On pourra consulter West et Harrison (1998) pour une approche générale du traitement bayésien des séries temporelles et Bauwens *et al.* (1999) pour une monographie économétrique sur ce sujet.

4.5.2 Le modèle AR

Comme présenté dans l'Exemple 4.22, le modèle AR(p) exprime la distribution de x_t conditionnellement au passé $x_{1:(t-1)}$ comme une régression linéaire normale sur les p variables les plus récentes, c'est-à-dire ($t = 1, 2, \ldots$),

$$x_t \sim \mathcal{N}\left(\mu - \sum_{i=1}^{p} \varrho_i (x_{t-i} - \mu), \sigma^2\right) , \tag{4.22}$$

où le paramètre de position μ est introduit pour plus de généralité. Notons que ce modèle est markovien, car la distribution de x_t ne dépend que d'un nombre fixe de valeurs passées, $x_{(t-p):(t-1)}$, et qu'il peut s'exprimer comme une chaîne de Markov régulière en considérant le vecteur $\mathbf{z}_t = x_{t:(t-p+1)}$, c'est-à-dire

$$\mathbf{z}_t = (x_t, x_{t-1}, \ldots, x_{t-p+1}) ,$$

car

$$\mathbf{z}_t = \mu\mathbf{1} + B(\mathbf{z}_{t-1} - \mu\mathbf{1}) + \varepsilon_t , \tag{4.23}$$

où

$$\mathbf{1} = (1, \ldots, 1)^t , \quad B = \begin{pmatrix} \varrho_1 & \varrho_2 & \cdots & \varrho_p \\ 1 & 0 & \ldots & 0 \\ & & \ddots & \\ 0 & 0 & & 0 \end{pmatrix} \quad \text{et} \quad \varepsilon_t = (\epsilon_t, 0, \ldots, 0)^t .$$

Puisque la vraisemblance conditionnelle aux valeurs négatives du temps $x_0, \ldots, x_{-p+1}$ peut s'écrire

$$L(\mu, \varrho_1, \ldots, \varrho_p, \sigma | x_{1:T}, x_{0:(-p+1)}) = \sigma^{-T} \prod_{t=1}^{T} \exp\left\{ -\left(x_t - \mu + \sum_{i=1}^{p} \varrho_i (x_{t-i} - \mu) \right)^2 / 2\sigma^2 \right\} , \tag{4.24}$$

il est possible de trouver une loi a priori conjuguée naturelle pour le paramètre $\theta = (\mu, \varrho_1, \ldots, \varrho_p, \sigma^2)$, c'est-à-dire une distribution normale sur $(\mu, \varrho_1, \ldots, \varrho_p)$ et une loi inverse gamma sur σ^2. À la place de la loi a priori de Jeffreys, qui est controversée dans ce cadre (voir la Note 4.7.2), nous pouvons aussi proposer une loi a priori non informative plus courante comme $\pi(\mu, \sigma, \varrho) = 1/\sigma$.

Si nous imposons la contrainte de stationnarité que toutes les racines de $\mathscr{P}$ soient en dehors du cercle unité, l'espace des paramètres est trop complexe pour des valeurs de p plus grandes que 3 pour proposer comme loi a priori la loi conjuguée normale restreinte à cet espace : par exemple, simuler cette loi est trop coûteux. Une solution, appelée *récurrence de Durbin-Levinson* (voir Monahan, 1984), est de proposer une *reparamétrisation* des paramètres ϱ_i en les *autocorrélations partielles* ψ_i (Exercice 4.54) qui satisfont, sous la contrainte de stationnarité,

$$\psi_i \in (-1,1)\,, \quad i = 1, \cdots, p\,,$$

et permettent alors une loi a priori uniforme[36]. Le résultat suivant fournit une connexion constructive entre $(\varrho_1, \ldots, \varrho_p)$ et $(\psi_1, \ldots, \psi_p)$.

Lemme 4.24. *Sous la stationnarité du modèle (4.22), les coefficients ϱ_i se déduisent des coefficients ψ_i par l'algorithme suivant :*

ALGORITHME 4.1. **Récurrence de Durbin-Levinson**
0. Définir $\varphi^{ii} = \psi_i$ et $\varphi^{ij} = \varphi^{(i-1)j} - \psi_i \varphi^{(i-1)(i-j)}$, pour $i > 1$ et $j = 1, \cdots, i-1$.
1. Prendre $\varrho_i = \varphi^{pi}$ pour $i = 1, \cdots, p$.

Bien que les lois a priori et a posteriori de $(\varrho_1, \ldots, \varrho_p)$ résultantes ne soient pas explicites, au sens où le calcul de la loi a priori (ou a posteriori) pour une valeur donnée du paramètre est assez coûteuse en temps, cette représentation peut s'exploiter en simulation, comme dans le Chapitre 6 (voir aussi Barnett *et al.*, 1996), à cause de la linéarité de la relation entre les ϱ_j et un ψ_i donné, conditionnellement aux autres ψ_ℓ. Huerta et West (1999) proposent une approche différente reposant sur les racines réelles et complexes du polynôme $\mathscr{P}$, qui, inversées, sont aussi à l'intérieur de l'unité du cercle.

4.5.3 Le modèle MA

Un résultat fondamental en théorie des processus stochastiques est la *décomposition de Wold*, qui énonce que la plupart des processus stationnaires (x_t) peuvent se représenter sous la forme $(t = 1, 2, \ldots)$

$$x_t = \mu + \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i}\,, \tag{4.25}$$

où $\psi_0 = 1$ et (ϵ_t) est un *bruit blanc*, c'est-à-dire une séquence de variables aléatoires de moyenne nulle, de variance fixe et de covariance nulle ; voir Box et Jenkins (1976) pour des détails théoriques.

[36] Les autocorrélations partielles, dites aussi *coefficients de réflexion* dans la littérature de traitement du signal, peuvent s'utiliser pour tester la stationnarité, car, selon le lemme de Schur, elles doivent toutes être entre -1 et 1 pour que la chaîne (x_t) soit stationnaire.

Exemple 4.25. (Suite de l'Exemple 4.6) Si $x_t = \varrho x_{t-1} + \epsilon_t$, x_t peut s'écrire aussi

$$x_t = \epsilon_t + \varrho\epsilon_{t-1} + \varrho^2\epsilon_{t-2} + \dots$$

si $|\varrho| < 1$. ‖

Le *modèle MA(q)*, MA signifiant *moving average (moyenne mobile)*, est un cas spécial de (4.25) lorsque les ψ_i sont égaux à 0 pour $i > q$, c'est-à-dire

$$x_t = \mu + \epsilon_t - \sum_{j=1}^{q} \vartheta_j \epsilon_{t-j}\,, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2) \tag{4.26}$$

En contraste avec le modèle AR(1), où la covariance entre les termes de la série décroissent exponentiellement vers 0 mais sont toujours non nuls, le processus MA(q) est tel que les autocovariances

$$\gamma_s = \mathrm{cov}(x_t, x_{t+s})$$

sont égales à 0 pour $|s| > q$. Selon la décomposition de Wold, le processus MA(q) est stationnaire, quel que soit le vecteur $(\vartheta_1, \dots, \vartheta_q)$. Cependant, des considérations d'inversibilité et d'identifiabilité (voir l'Exercice 4.59) mènent à la condition que le polynôme

$$\mathcal{Q}(x) = 1 - \sum_{j=1}^{q} \vartheta_j x^j$$

doit avoir toutes ses racines en dehors du cercle unité.

Exemple 4.26. Dans le cas particulier du modèle MA(1), $x_t = \mu + \epsilon_t - \vartheta_1\epsilon_{t-1}$ et $\mathrm{var}(x_t) = (1 + \vartheta_1^2)\sigma^2$, avec $\gamma_1 = \vartheta_1\sigma^2$. Alors x_t peut aussi s'écrire comme

$$x_t = \mu + \tilde{\epsilon}_{t-1} - \frac{1}{\vartheta_1}\tilde{\epsilon}_t, \quad \tilde{\epsilon} \sim \mathcal{N}(0, \vartheta_1^2\sigma^2)\,,$$

ce qui montre que les couples (ϑ_1, σ) et $(1/\vartheta_1, \vartheta_1\sigma)$ mènent à deux représentations alternatives du même modèle. Ceci justifie en quelque sorte la restriction à $|\vartheta_1| < 1$. ‖

Contrairement au modèle AR(p), ce modèle n'est pas markovien *per se* (même s'il peut se représenter comme un processus de Markov, en utilisant la représentation à espace d'état introduite ci-dessous). Bien que le vecteur entier $x_{1:T}$ soit une variable aléatoire normale de moyenne constante μ et de matrice de covariance

$$\Sigma = \begin{pmatrix} \sigma^2 & \gamma_1 & \gamma_2 & \dots & \gamma_q & 0 & \dots & 0 & 0 \\ \gamma_1 & \sigma^2 & \gamma_1 & \dots & \gamma_{q-1} & \gamma_q & \dots & 0 & 0 \\ & & & \ddots & & & & & \\ 0 & 0 & 0 & \dots & 0 & 0 & \dots & \gamma_1 & \sigma^2 \end{pmatrix},$$

avec ($|s| \leq q$)

$$\gamma_s = \sigma^2 \sum_{i=0}^{q-|s|} \vartheta_i \vartheta_{i+|s|}\,, \tag{4.27}$$

et fournit donc une fonction de vraisemblance explicite, le calcul et évidemment l'intégration (ou la maximisation) de cette vraisemblance pour une valeur donnée du paramètre sont assez coûteux, car ils nécessitent d'inverser la matrice $n \times n$ Σ. Une représentation plus pratique est d'utiliser la vraisemblance de $x_{1:T}$ conditionnelle à $(\epsilon_0, \ldots, \epsilon_{-q+1})$,

$$L(\mu, \vartheta_1, \ldots, \vartheta_q, \sigma | x_{1:T}, \epsilon_0, \ldots, \epsilon_{-q+1}) = \tag{4.28}$$
$$\sigma^{-T} \prod_{t=1}^{T} \exp\left\{ -\left(x_t - \mu + \sum_{j=1}^{q} \vartheta_j \hat{\epsilon}_{t-j} \right)^2 / 2\sigma^2 \right\},$$

où ($t > 0$)

$$\hat{\epsilon}_t = x_t - \mu + \sum_{j=1}^{q} \vartheta_j \hat{\epsilon}_{t-j} \tag{4.29}$$

et $\hat{\epsilon}_0 = \epsilon_0$, …, $\hat{\epsilon}_{1-q} = \epsilon_{1-q}$. Cette définition récursive de la vraisemblance reste coûteuse, car elle implique T calculs de q termes. Néanmoins, même si le problème des valeurs de conditionnement $(\epsilon_0, \ldots, \epsilon_{-q+1})$ doit se traiter séparément, par exemple à travers une mise en œuvre de méthodes de Monte Carlo par chaînes de Markov (MCMC) (voir le Chapitre 6), la complexité de cette représentation est plus maniable que celle de la représentation donnée ci-dessus.

Une autre approche intéressante est d'utiliser la représentation dite *à espace d'état*, inspirée du *filtre de Kalman*, qui donne des formules linéaires récursives pour la *prédiction*, le lissage et le *filtrage*. Brockwell et Davis (1998, Chapitre 8) donnent une présentation générale de cette technique (voir aussi Cappé *et al.*, 2005), tandis que West et Harrison (1998) décrivent leur version bayésienne, mais l'idée générale est de représenter une série temporelle (x_t) comme un système de deux équations,

$$x_t = G_y \mathbf{y}_t + \varepsilon_t\,, \tag{4.30}$$
$$\mathbf{y}_{t+1} = F_t \mathbf{y}_t + \xi_t\,, \tag{4.31}$$

où les vecteurs ε_t et ξ_t sont des vecteurs multivariés normaux de matrices de covariance générales qui dépendent de t et $\mathbb{E}[\varepsilon_u \xi_v'] = 0$ pour tout (u, v). L'équation (4.30) est appelée *équation d'observation* et (4.31) est appelée *équation d'état*. Cette représentation projette le processus d'intérêt (x_t) dans un espace plus grand, l'*espace d'état*, où le processus $(\mathbf{y}_t)$ est markovien et linéaire. Par exemple, (4.23) est une *représentation à espace d'état* du modèle AR(p).

Le modèle MA(q) peut s'écrire de cette façon en définissant $\mathbf{y}_t = (\epsilon_{t-q}, \ldots, \epsilon_{t-1}, \epsilon_t)'$. L'équation d'état est alors

$$\mathbf{y}_{t+1} = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ & & \ldots & & \\ 0 & 0 & 0 & \ldots & 1 \\ 0 & 0 & 0 & \ldots & 0 \end{pmatrix} \mathbf{y}_t + \epsilon_{t+1} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \tag{4.32}$$

et l'équation d'observation est

$$x_t = \mu - \left(\vartheta_q \; \vartheta_{q-1} \; \ldots \; \vartheta_1 \; -1\right) \mathbf{y}_t \,.$$

Par conséquent, cette décomposition ne met pas en jeu un vecteur ε_t dans l'équation d'observation, tandis que ξ_t est dégénéré dans l'équation d'état. Ce phénomène de dégénérescence est assez commun dans les représentations à espace d'état, mais ceci n'est pas un obstacle à l'utilisation conditionnelle du modèle, comme dans les algorithmes MCMC du Chapitre 6. Notons aussi que la représentation à espace d'état d'un modèle n'est pas unique.

Exemple 4.27. (Suite de l'Exemple 4.26) Pour le modèle MA(1), l'équation d'observation peut aussi être $x_t = (1\ 0)\mathbf{y}_t$ avec $\mathbf{y}_t = (y_{1t}\ y_{2t})'$ associée à l'équation d'état

$$\mathbf{y}_{t+1} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \mathbf{y}_t + \epsilon_{t+1} \begin{pmatrix} 1 \\ \vartheta_1 \end{pmatrix} . \qquad \|$$

Quelle que soit la représentation choisie pour le modèle MA(q), la condition d'identifiabilité sur $\mathscr{Q}(x)$ impose que les ϑ_j varient dans un espace complexe, qui ne peut pas être décrit directement pour des valeurs de q plus grandes que 3. La reparamétrisation décrite dans le Lemme 4.24 s'applique aussi formellement dans ce cas, mais avec une interprétation différente pour les ψ_i, qui sont alors les *autocorrélations partielles inverses* (Jones, 1987). Une loi a priori uniforme pour les ψ_i peut s'utiliser pour l'estimation des ϑ_i, ce qui implique le recours à une méthode MCMC (voir Chapitre 6, Chib et Greenberg, 1994, Barnett *et al.*, 1996 et Billio *et al.*, 1999).

4.5.4 Le modèle ARMA

Une extension simple du modèle précédent est le modèle ARMA(p, q), où ($t = 1, 2, \ldots$)

$$x_t = \mu - \sum_{i=1}^{p} \varrho_i (x_{t-i} - \mu) + \epsilon_t - \sum_{j=1}^{q} \vartheta_j \epsilon_{t-j} \,, \tag{4.33}$$

où les ϵ_t's sont i.i.d. $\mathcal{N}(0, \sigma^2)$. Le but de tels modèles, relativement aux deux modèles AR et MA, est de permettre une plus forte parcimonie, c'est-à-dire d'utiliser des valeurs beaucoup plus petites de p et q que dans un modèle uniquement AR ou uniquement MA (voir la Note 6.6.6 pour des détails sur la notion de *parcimonie*).

Comme l'ont détaillé Box et Jenkins (1976), les conditions de stationnarité et d'identifiabilité correspondent de nouveau au fait que les racines des polynômes $\mathscr{P}$ et $\mathscr{Q}$ sont en dehors du cercle unité, avec comme condition supplémentaire que les deux polynômes n'aient pas de racine commune. (Mais ceci n'arrive presque sûrement pas sous une loi a priori continue pour les paramètres.) La reparamétrisation du Lemme 4.24 peut par conséquent s'appliquer à la fois aux ϑ_i et aux ϱ_j, nécessitant de nouveau un recours aux techniques MCMC, en raison de la complexité de la loi a posteriori.

Naturellement, des représentations à espace d'état existent également pour les modèles ARMA(p, q), une possibilité étant (Brockwell et Davis, 1998, Exemple 8.3.2)

$$x_t = \mu - \left(\vartheta_{r-1}\ \vartheta_{r-2}\ \ldots\ \vartheta_1\ -1\right)\mathbf{y}_t$$

pour l'équation d'observation et

$$\mathbf{y}_{t+1} = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ & & & \ldots & \\ 0 & 0 & 0 & \ldots & 1 \\ \varrho_r & \varrho_{r-1} & \varrho_{r-2} & \ldots & \varrho_1 \end{pmatrix} \mathbf{y}_t + \epsilon_{t+1} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}, \tag{4.34}$$

pour l'équation d'état, avec $r = \max(p, q+1)$ et la convention que $\varrho_t = 0$ si $t > p$ et $\vartheta_t = 0$ si $t > q$. Comme pour les modèles MA(q), cette représentation est pratique pour concevoir des algorithmes MCMC (voir le Chapitre 6) qui simulent la loi a posteriori des paramètres du modèle ARMA(p, q).

4.6 Exercices

Section 4.1

4.1 (Smith, 1984) Soit x, une variable aléatoire de moyenne μ, fonction de répartition F, et densité f. Les fonctions f et f' sont supposées bornées. Définir une suite de variables aléatoires y_n de fonction de répartition

$$G_n(y) = \left(1 - \frac{1}{n}\right) F(y) + \frac{1}{n} H_n(y),$$

satisfaisant
(i) $\mathbb{E}^{H_n}[y] = n^2$; et
(ii) $H_n' = h_n$ et h_n' sont bornés.

Montrer que $G_n \to F$, $G'_n = g_n \to f$, et $g'_n \to f'$, mais que $|\mu - \mathbb{E}[y_n]| \to \infty$.

4.2 Si $\psi(\theta|x)$ est une loi a posteriori associée à $f(x|\theta)$ et à la loi a priori π, éventuellement impropre, montrer que

$$\frac{\psi(\theta|x)}{f(x|\theta)} = k(x)\pi(\theta).$$

a. En déduire que, si f appartient à une famille exponentielle, la distribution a posteriori appartient elle aussi à une famille exponentielle, quelle que soit π.
b. Montrer que si ψ appartient à une famille exponentielle, f y appartient aussi.

4.3 *(Berger et Wolpert, 1988) Dans le cas suivant, Stein (1962b) met en avant certaines des limitations du principe de vraisemblance. Supposons qu'une valeur $\theta > 0$ puisse être évaluée soit par $x \sim \mathcal{N}(\theta, \sigma^2)$ (avec σ^2 connu), soit par

$$y \sim f(y|\theta) = cy^{-1} \exp\left\{-\frac{d^2}{2}\left(1 - \frac{\theta}{y}\right)^2\right\} \mathbb{I}_{[0,b\theta]}(y),$$

où b est très grand et d grand (disons 50).
a. Montrer que les deux estimateurs du maximum de vraisemblance de θ sont $\delta_1(x) = x$ et $\delta_2(y) = y$.
b. Considérer le cas particulier $x = y = \sigma d$. Expliquer pourquoi l'inférence sur θ devrait être la même dans les deux cas.
c. Expliquer pourquoi

$$[x - 1.96\,\sigma, x + 1.96\,\sigma]$$

pourrait être proposé comme intervalle de confiance à 95% pour θ.
d. En déduire que

$$[y - (1.96)(y/d), y + (1.96)(y/d)]$$

peut être utilisé comme intervalle de confiance si y est observé.
e. Montrer que

$$P(y - (1.96)(y/d) < \theta < y + (1.96)(y/d))$$

peut être rendu aussi petit que possible pour un choix idoine de b.
f. Conclure que l'intervalle de confiance ci-dessus n'est pas approprié pour de grandes valeurs de $x = y$ et de σ, et discuter de la pertinence des intervalles de confiance eu égard au principe de vraisemblance.
g. Étudier le même problème avec la loi a priori $\pi(\theta) = 1/\theta$.

4.4 Montrer que, si $p \in [0, 1]$, $\theta = p/(1-p)$ et si $\pi(\theta) = 1/\theta$, la loi a priori $\pi(p)$ est la distribution de Haldane.

4.5 Montrer que le phénomène opposé à celui de l'Exemple 4.2 peut avoir lieu, c'est-à-dire qu'il peut être tel que l'information a priori est négligeable. (*Indication :* Prendre $\pi(\theta)$ égal à $\mathcal{C}(\mu, 1)$ et $f(x|\theta) \propto \exp -|x - \theta|$, et montrer alors que l'estimateur MAP ne dépend pas de μ.)

4.6 Dans le cadre de l'Exemple 4.2, considérer $\pi(\theta) \propto \exp -a|\theta|$ et montrer que, pour a suffisamment petit, l'estimateur MAP n'est pas systématiquement égal à 0.

4.7 Montrer que le paradoxe d'un estimateur MAP constant exhibé dans l'Exemple 4.2 disparaît lorsque le nombre d'observations de la loi $\mathcal{C}(\theta, 1)$ augmente.

4.8 Un *tableau de contingence* est une matrice $k \times \ell$ telle que l'élément (i, j) est n_{ij}, le nombre d'occurrences simultanées de la i-ième modalité d'une première caractéristique et de la j-ième modalité d'une seconde caractéristique dans une population de n individus ($1 \leq i \leq k, 1 \leq j \leq \ell$). La probabilité de cette occurrence est notée p_{ij}.

a. Montrer que de telles lois appartiennent à une famille exponentielle.

b. Déterminer la loi des marges du tableau, c'est-à-dire de

$$n_{i\cdot} = n_{i1} + \ldots + n_{i\ell} \quad \text{et} \quad n_{\cdot j} = n_{1j} + \ldots + n_{kj}\,.$$

En déduire la loi de $(n_{1\cdot}, \ldots, n_{k\cdot})$ et de $(n_{\cdot 1}, \ldots, n_{\cdot \ell})$.

c. Donner les lois a priori conjuguées sur $p = (p_{ij})$ et la loi a priori de Jeffreys.

d. Dans le cas particulier où les deux variables sont *indépendantes*, les paramètres sont supposés satisfaire les relations $p_{ij} = p_{i\cdot}p_{\cdot j}$ où $(p_{1\cdot}, \ldots, p_{k\cdot})$ et $(p_{\cdot 1}, \ldots, p_{\cdot \ell})$ sont deux vecteurs de probabilités. Relier ces vecteurs aux lois obtenues en b. et construire les lois a priori conjuguées correspondantes.

e. Comparer les espérances a posteriori de p_{ij} pour les lois a priori conjuguées des questions c. et d. [*Note :* Voir Santner et Duffy, 1989, pour une présentation détaillée du traitement bayésien de ces modèles.]

4.9 Déterminer si les lois suivantes peuvent être des lois a posteriori :
(i) $\mathscr{T}_1(k, \mu(x), \tau^2(x))$ avec $x \sim \mathscr{N}(\theta, \sigma^2)$ et σ^2 connu ;
(ii) une distribution normale tronquée $\mathscr{N}(\mu(x), \tau^2(x))$ avec $x \sim \mathscr{P}(\theta)$; et
(iii) $\mathscr{P}a(\alpha(x), \mu(x))$ avec $x \sim \mathscr{B}(n, 1/\theta)$.

4.10 ***(Suite de l'Exercice 4.9)** Pour une distribution d'échantillonnage $f(x|\theta)$ et une distribution conditionnelle $g(\theta|x)$, donner une condition nécessaire et suffisante pour que $g(\theta|x)$ soit une loi a posteriori associée à $f(x|\theta)$ et à une loi a priori arbitraire $\pi(\theta)$.

4.11 Soit $(x_n)_n$ une chaîne de Markov à espace d'état fini $\{1, \ldots, p\}$ et de matrice de transition P.

a. Si l'échantillon est $x_1, \ldots, x_n$, exprimer la fonction de vraisemblance et calculer les lois a priori conjuguées des composantes de P.

b. La chaîne de Markov est désormais observée à des temps aléatoires $t_1 < \cdots < t_n$. Donner la fonction de vraisemblance $\ell(P|x_{t_1}, \ldots, x_{t_n})$, en supposant que la distribution des t_i ne dépend pas de P et déterminer si les lois a priori ci-dessus permettent toujours des calculs analytiques.

c. Une variable aléatoire y_t de distribution conditionnelle $f(y|\theta_{x_t})$ est observée pour $t = 1, \ldots, n$. On suppose que les y_t sont indépendants, conditionnellement aux x_t. Montrer que la distribution marginale des y_t est un mélange des distributions $f(y|\theta_k)$.

d. Si seulement les y_t sont observés, le modèle est une *chaîne de Markov cachée*. Lorsque $f(y|\theta)$ appartient à une famille exponentielle, donner la fonction de vraisemblance et les lois a priori conjuguées sur $(P, \theta_1, \ldots, \theta_p)$.

e. Considérer le cas particulier $p = 2$ et $f(y|\theta) = \theta \exp(-\theta y)\mathbb{I}_{\mathbb{R}^+}(y)$ afin d'établir si les lois a priori ci-dessus admettent une expression simple.

4.12 Soient $x \sim \mathscr{B}(m, p)$ et $p \sim \mathscr{B}e(1/2, 1/2)$.

a. Montrer que cette loi a priori est équivalente à la loi uniforme sur $\theta = \arcsin(\sqrt{p})$. Comment justifier cette transformation ? [*Note :* Voir Feller, 1970, pour plus de détails sur la loi de l'arcsinus.]

b. Soit $y \sim \mathscr{B}(n,q)$ une observation indépendante, avec $q \sim \mathscr{B}e(1/2,1/2)$. Utiliser l'approximation $\arcsin x \sim \mathscr{N}(\theta, 1/4m)$ afin d'obtenir une loi a posteriori approchée de $\arcsin(\sqrt{p}) - \arcsin(\sqrt{q})$.

c. En déduire une approximation de

$$\pi(|\arcsin(\sqrt{p}) - \arcsin(\sqrt{q})| < 0.1|x,y).$$

4.13 La *distribution logistique* est définie par la densité

$$e^{-(x-\theta)}/(1+e^{-(x-\theta)})^2$$

sur $\mathbb{R}$.

a. Montrer que la fonction ci-dessus est bien une densité de probabilité et calculer l'estimateur du maximum de vraisemblance de θ.

b. Montrer que cette loi n'appartient pas à une famille exponentielle (i) directement ; et (ii) en utilisant l'Exercice 3.20. En déduire qu'il n'existe pas de loi a priori conjuguée et proposer une loi a priori non informative.

c. Établir l'expression de l'estimateur du maximum de vraisemblance de θ pour un échantillon $x_1, \ldots, x_n$. Montrer par un exemple que la vraisemblance peut avoir plusieurs modes.

d. Relier la régression logistique et la loi logistique en exhibant des variables aléatoires logistiques latentes dans le modèle de régression logistique. Y a-t-il une contradiction entre la question b. et le fait que le modèle de régression logistique appartienne à une famille exponentielle, comme le montre l'Exemple 3.21 ?

4.14 Pour le modèle AR(1) de l'Exemple 4.6, montrer que la distribution a posteriori jointe $\pi(\varrho, \sigma^2|x_{1:(T-1)})$ admet une expression explicite pour la loi a priori conjuguée

$$\varrho \sim \mathscr{N}(0, \kappa\sigma^2), \qquad \sigma^2 \sim \mathscr{IG}(\alpha,\beta)\,.$$

En déduire la densité prédictive $\pi(x_T|x_{1:(T-1)})$.

Section 4.2

4.15 (Smith, 1988) Une justification usuelle des coûts quadratiques est qu'ils fournissent une approximation du second ordre des coûts symétriques. Soit la fonction de coût

$$\mathrm{L}(\theta,\delta) = 1 - e^{-(\delta-\theta)^2/2}$$

et $\pi(\theta|x) = (1/2)\{\varphi(\theta;8,1) + \varphi(\theta;-8,1)\}$, un mélange de deux distributions normales de moyennes respectives 8 et -8, et de variance 1.

a. Montrer que $\pi(\theta|x)$ peut en fait s'écrire comme une loi a posteriori.

b. Montrer que $\mathbb{E}^\pi[\theta|x]$ est un maximum local du coût a posteriori.

c. Relier le coût $\mathrm{L}(\theta,\delta)$ aux coûts intrinsèques de la Section 2.5.4.

4.16 Soient $x \sim \mathscr{P}(\lambda)$ et $\pi(\lambda) = e^{-\lambda}$. Le but de l'exercice est de comparer les estimateurs $\delta_c(x) = cx$ sous les coûts quadratiques $\mathrm{L}(\lambda,\delta) = (\delta-\lambda)^2$.

a. Calculer $R(\delta_c,\lambda)$ et montrer que δ_c n'est pas admissible pour $c > 1$.

b. Calculer $r(\pi,\delta_c)$ et en déduire le c^π optimal.

c. Montrer qu'il n'existe pas d'estimateur optimal δ_c au sens minimax.

d. Reprendre les questions précédentes pour la fonction de coût

$$\mathrm{L}'(\lambda,\theta) = \left(\frac{\delta}{\lambda} - 1\right)^2.$$

4.17 Montrer que l'estimateur de Bayes associé à un coût quadratique et une loi a priori propre ne peut pas être sans biais. Est-ce que ce résultat s'étend aux estimateurs de Bayes généralisés ? À d'autres coûts ?

4.18 Soient $x \sim \mathscr{B}(n,p)$ et $p \sim \mathscr{B}e(\alpha,\beta)$.

a. Calculer les distributions a posteriori et marginale. En déduire l'estimateur de Bayes sous le coût quadratique.

b. Si la loi a priori est $\pi(p) = [p(1-p)]^{-1}\mathbb{I}_{(0,1)}(p)$, donner l'estimateur de Bayes généralisé de p (lorsqu'il est défini).

c. Sous quelle condition sur (α,β), δ^π est-il sans biais ? S'agit-il d'une contradiction avec l'Exercice 4.17 ?

d. Donner l'estimateur de Bayes de p sous le coût

$$\mathrm{L}(p,\delta) = \frac{(\delta-p)^2}{p(1-p)}.$$

4.19 En utilisant les estimateurs du Tableau 4.1, montrer que les estimateurs correspondant à des lois a priori non informatives peuvent s'écrire comme des limites d'estimateurs conjuguées. Est-ce que cette convergence s'étend à d'autres quantités d'intérêt pour la même suite d'hyperparamètres conjugués ? Essayer d'établir un résultat général.

4.20 Soient $x \sim \mathscr{N}(\theta,1)$, $\theta \sim \mathscr{N}(0,1)$ et $\mathrm{L}(\theta,\delta) = \mathbb{I}_{\{\delta<\theta\}}$. Montrer qu'il n'existe pas d'estimateur de Bayes dans ce cas.

4.21 Soit $x \sim \mathscr{P}(\theta)$, avec $\Theta = \{\theta_1,\theta_2\}$ et $\mathscr{D} = \{d_1,d_2,d_3\}$. La fonction de coût est définie par la matrice

$$\mathrm{L} = \begin{pmatrix} 0 & 20 & 10 \\ 50 & 0 & 20 \end{pmatrix}$$

(où $\mathrm{L}_{ij} = \mathrm{L}(\theta_i,d_j)$, $i = 1,2$, $j = 1,2,3$). Montrer que les estimateurs de Bayes sont de la forme

$$\delta^\pi(x) = \begin{cases} d_1 & \text{si } x < k - \log_2 3, \\ d_2 & \text{si } x > k-1, \\ d_3 & \text{sinon,} \end{cases}$$

et définir k à partir de la loi a priori π.

4.22 (Ferguson, 1967) Soit x suivant la distribution négative binomiale renormalisée,

$$f(x|\theta) = \binom{r+x-1}{x}\theta^x(1+\theta)^{-(r+x)}, \qquad x = 0,1,\ldots, \quad \theta \in \mathbb{R}_+^*\,.$$

Montrer que $\mathbb{E}_\theta[x] = r\theta$ (d'où $\theta = p/(1-p)$). La fonction de coût est l'erreur quadratique pondérée

$$\mathrm{L}(\theta,\delta) = \frac{(\theta-\delta)^2}{\theta(1+\theta)}.$$

a. Donner l'estimateur du maximum de vraisemblance de θ.

b. Montrer que $\delta_0(x) = x/r$ admet une fonction de risque constante et est l'estimateur de Bayes généralisé pour $\pi(\theta) = 1$ si $r > 1$. Que se passe-t-il lorsque $r = 1$?

c. Montrer que

$$\delta_{\alpha,\beta}(x) = \frac{\alpha + x - 1}{\beta + r + 1}$$

est un estimateur de Bayes pour

$$\pi(\theta|\alpha,\beta) \propto \theta^{\alpha-1}(1+\theta)^{-(\alpha+\beta)}$$

et que cette loi est conjuguée pour $f(x|\theta)$.

d. En déduire que $\delta_1(x) = x/(r+1)$ est un estimateur minimax.

4.23 (Ferguson, 1967) Soient $\Theta = [0,1]$ et $\mathrm{L}(\theta,\delta) = \frac{(\theta-\delta)^2}{1-\theta}$, pour la loi géométrique

$$f(x|\theta) = \theta^x(1-\theta) \qquad (x \in \mathbb{N}).$$

a. Donner un développement en série entière de $R(\theta,\delta)$ comme fonction de θ.

b. Montrer que l'unique estimateur *non randomisé* de risque constant est δ_0 tel que

$$\delta_0(0) = 1/2, \qquad \delta_0(x) = 1 \text{ si } x \geq 1.$$

c. Montrer que, si δ^π est l'estimateur de Bayes associé à π, $\delta^\pi(n) = \mu_{n-1}/\mu_n$, où μ_i est le i-ième moment de π.

d. Montrer que δ_0 est minimax.

4.24 *(Casella et Strawderman, 1981) Soit $x \sim \mathcal{N}(\theta,1)$ avec $|\theta| \leq m$ $(m < 1)$.

a. Montrer que $\delta^m(x) = m\tanh(mx)$ est l'estimateur de Bayes associé à

$$\pi^m(\theta) = \frac{1}{2}\mathbb{I}_{\{-m,m\}}(\theta).$$

b. Montrer que, pour le coût quadratique, $r(\pi^m,\delta^m) = R(\delta^m,\pm m)$ et en déduire que δ^m est minimax. [*Note :* Il s'agit en fait de l'unique estimateur minimax dans ce cas.]

c. Comparer à l'estimateur δ^U associé à la loi a priori uniforme

$$\pi(\theta) = \frac{1}{2m}\mathbb{I}_{[-m,m]}(\theta),$$

en fonction de m. [*Note :* Gatsonis *et al.*, 1987, donnent une étude détaillée de la performance de δ^U en termes de minimaxité.]

4.25 (Casella et Berger, 2001) Soient $x \sim \mathcal{U}_{\{1,2,\ldots,\theta\}}$ et $\theta \in \Theta = \mathbb{N}^*$.

a. Si $\mathcal{D} = \Theta$, montrer que, sous le coût quadratique, $\mathbb{E}^\pi[\theta|x]$ n'est pas forcément l'estimateur de Bayes.

b. Si $\mathcal{D} = [1,+\infty)$, montrer que $\mathbb{E}^\pi[\theta|x]$ est l'estimateur de Bayes (lorsqu'il existe).

c. Montrer que $\delta_0(x) = x$ est admissible pour tout choix de $\mathcal{D}$. (*Indication :* Commencer par $R(1,\delta_0)$.)

d. Montrer que δ_0 est un estimateur de Bayes et qu'il existe d'autres estimateurs de Bayes pour cette loi a priori, de fonctions de risque différentes.

4.26 Soient x_1, x_2 i.i.d. de distribution $f(x|\theta) = (1/2)\exp(-|x-\theta|)$ et $\pi(\theta) = 1$. Déterminer les estimateurs de Bayes associés aux coûts absolus et quadratiques. Même question pour une observation additionnelle. [*Note :* Voir l'Exemple 1.12 pour une motivation historique.]

Section 4.3.1

4.27 Chrystal (1891) écrit : *"Personne ne dira que, si vous mettez simplement deux boules blanches dans un sac contenant une boule de couleur inconnue, avec une même chance qu'elle soit noire ou blanche, cette action accroît le rapport des chances que la boule inconnue soit blanche de un contre un à trois contre un"*, comme un argument contre la règle de succession de Laplace. Considérez-vous que cette critique est valable ? (Voir Zabell, 1989.)

4.28 (Jeffreys, 1961)

a. Montrer que

$$\sum_{i=1}^{N} \binom{i}{x_1}\binom{N-i}{x-x_1} = \binom{N+1}{x+1}$$

(i) par des calculs algébriques ; et
(ii) en utilisant le calcul combinatoire.

b. Si l'échantillon contient $x = x_1 + x_2$ individus, montrer que la probabilité que les $y = y_1 + y_2$ tirages suivants contiendront y_1 individus de la première population et y_2 de la seconde, est

$$P(y_1, y_2|x_1, x_2) = \frac{y!}{y_1!\,y_2!}\,\frac{(x_1+1)\ldots(x_1+y_1)(x_2+1)\ldots(x_2+y_2)}{(x+2)\ldots(x+y+1)}.$$

c. Pour $x = x_1$, en déduire que la probabilité que les y tirages suivants sont du même type est

$$\frac{x+1}{x+y+1}.$$

4.29 Généraliser la règle de succession de Laplace au modèle multinomial.

Certains problèmes similaires à la règle de succession de Laplace ont été considérés par Lewis Carroll dans son livre Pillow Problems. *Seneta (1993) donne un commentaire détaillé sur ces problèmes, dont deux sont donnés ci-dessous.*

4.30 Soient deux sacs, H et K, contenant deux boules chacun. Chaque boule est soit blanche, soit noire. Une boule blanche est ajoutée au sac H et une boule est choisie au hasard dans le sac H et transférée dans le sac K, sans qu'on regarde sa couleur.

a. Quelle est la probabilité de tirer une boule blanche du sac K ?
b. Une boule blanche est ensuite ajoutée au sac K et on transfère de nouveau du sac K au sac H une boule prise au hasard sans la regarder. Quelle est désormais la probabilité de tirer une boule blanche du sac H ?

4.31 *"Pour une infinité de baguettes cassées, établir la probabilité qu'une d'entre elles au moins soit cassée au milieu."* Bien que cette question soit mal formulée, puisque le milieu est de mesure zéro, une solution discrète est proposée ici.

a. Supposons que chaque baguette a $2m + 1$ points de rupture et qu'il y a exactement $2m + 1$ baguettes. Donner la probabilité qu'aucune baguette ne casse au milieu et calculer la valeur limite de cette probabilité lorsque m tend vers l'infini.
b. Étudier la dépendance de cette limite à l'hypothèse que le nombre m de points de rupture est égal au nombre de baguettes.

Section 4.3.2

4.32 Dans le cadre de l'Exemple 4.17, développer un modèle bayésien pour la distribution de $(t_2 - t_1)$. Étendre au problème suivant : Étant donné qu'un feu est au rouge depuis une minute, quelle est la probabilité qu'il passe au vert la minute suivante ?

4.33 Montrer que, pour le problème du tramway, l'estimateur du maximum de vraisemblance $\hat{N} = T$ est admissible pour toute fonction de coût de la forme $\mathrm{L}(|\hat{N} - N|)$, avec L fonction strictement croissante. (*Indication :* Considérer d'abord le cas $N = 1$.)

Section 4.3.3

4.34 Pendant le lancement d'un nouveau journal étudiant, $n_1 = 220$ et $n_2 = 570$ personnes ont acheté les numéros tests -1 et 0. Le nombre de personnes qui ont acheté les deux numéros est $n_{11} = 180$. Donner un estimateur de Bayes de N, le nombre total de lecteurs, en supposant qu'un modèle de capture-recapture s'applique et que $\pi(N)$ est $\mathscr{P}(1000)$.

4.35 (Castledine, 1981) Pour le modèle de Wolter introduit en Section 4.3.3, c'est-à-dire lorsque n_1 et n_2 sont des variables aléatoires, le *modèle temporel* considère le cas où tous les individus ont la même probabilité de capture pour une expérience donnée, mais où cette probabilité varie entre la première et la seconde capture. Ces deux probabilités sont notées p_1 et p_2 .

a. Donner la vraisemblance et l'estimateur du maximum de vraisemblance associés à ce modèle lorsque p_1 et p_2 sont *connus*.

b. Montrer que la loi a posteriori de N sachant p_1 et p_2 ne dépend que de $n_+ = n_1 + n_2 - n_{11}$ et $\mu = 1 - (1 - p_1)(1 - p_2)$. Lorsque la loi a priori de N est $\pi(N) = 1$, montrer que $\pi(N|n_+, \mu)$ est la loi $\mathscr{N}eg(n_+, \mu)$.

c. Donner la distribution marginale a posteriori de N lorsque $p_1 \sim \mathscr{B}(\alpha, \beta)$ et $p_2 \sim \mathscr{B}(\alpha, \beta)$.

d. Montrer que, si $\alpha = 0$, $\beta = 1$, nous retrouvons le modèle de Darroch comme distribution marginale de N. Cette décomposition facilite-t-elle le calcul de l'estimateur de Bayes ?

Section 4.4.1

4.36 *(Robert, 1990) La fonction de Bessel modifiée I_ν ($\nu \geq 0$) est une solution de l'équation différentielle $z^2 f'' + z f' - (z^2 + \nu^2) f(z) = 0$ et peut être représentée par un développement en séries limitées

$$I_\nu(z) = \left(\frac{z}{2}\right)^\nu \sum_{k=0}^{\infty} \frac{(z/2)^{2k}}{k!\ \Gamma(\nu + k + 1)}.$$

a. Montrer que les séries ci-dessus convergent dans $\mathbb{R}$ quel que soit $\nu \geq 0$.

b. En développant

$$\int_0^\pi e^{z\cos(\theta)} \sin^{2\nu}(\theta)\, d\theta$$

en série entière, montrer que I_ν peut s'écrire

$$I_\nu(z) = \frac{(z/2)^\nu}{\pi^{1/2}\Gamma(\nu + \frac{1}{2})} \int_0^\pi e^{z\cos(\theta)} \sin^{2\nu}(\theta)\, d\theta. \tag{4.35}$$

c. Établir les formules de récurrence suivantes :

$$\begin{cases} I_{\nu+1}(z) = I_{\nu-1}(z) - (2\nu/z)I_\nu(z), \\ I'_\nu(z) = I_{\nu-1}(z) - (\nu/z)I_\nu(z). \end{cases}$$

d. Établir à partir de la représentation (4.35) et par une intégration par parties que, pour $z > 0$,

$$I_{\nu+1}(z) \leq I_\nu(z).$$

e. Déduire du développement en série entière de I_ν que $t^{-\nu}I_\nu(t)$ croît en t. Si on définit r_ν comme

$$r_\nu(t) = \frac{I_{\nu+1}(t)}{I_\nu(t)},$$

montrer que r_ν est une fonction croissante et concave, et que $r_\nu(t)/t$ décroît.

f. Montrer que

$$\lim_{t\to 0} r_\nu(t) = 1, \qquad \lim_{t\to\infty} \frac{r_\nu(t)}{t} = \frac{1}{2(\nu+1)},$$

et que

$$r'_\nu(t) = 1 - \frac{2\nu+1}{t} r_\nu(t) - r^2_\nu(t).$$

g. Montrer que la densité d'une loi du khi deux décentré de paramètre de décentrage λ et à ν degrés de liberté peut s'exprimer comme une fonction de Bessel modifiée, soit,

$$p_{\lambda,\nu}(x) = \frac{1}{2}\left(\frac{x}{\lambda}\right)^{\frac{\nu-2}{4}} I_{\frac{\nu-2}{2}}(\sqrt{\lambda x}) e^{-\frac{x+\lambda}{2}}.$$

4.37 *(Bock et Robert, 1985) Sur $\mathbb{R}^p$, la sphère de rayon c est définie par

$$S_c = \left\{ z \in \mathbb{R}^p;\ ||z||^2 = c \right\}.$$

a. Si $x \sim \mathcal{N}_p(\theta, I_p)$, avec $p \geq 3$, et si θ a pour loi a priori π_c, la loi uniforme sur S_c, montrer que la densité marginale de x est proportionnelle à

$$m_c(x) = e^{-||x||^2/2} e^{-c^2/2} \frac{I_{\frac{p-2}{2}}(||x||c)}{(c||x||)^{\frac{p-2}{2}}}.$$

b. Montrer que le coefficient de proportionnalité est indépendant de c et rappeler pourquoi il n'apparaît pas dans la loi a posteriori.

c. Déduire de la question a. l'espérance a posteriori δ_c par une dérivation. (*Indication :* Voir le Lemme 4.8.)

d. Montrer que, si $c \geq \sqrt{p}$, δ_c est un estimateur à rétrécisseur en dehors de la boule $\{x;\ ||x|| \leq \varrho\}$ et à "agrandisseur" à l'intérieur. Déterminer la valeur seuil ϱ.

e. Montrer que δ_c ne peut pas être minimax. Cet estimateur est-il admissible ?

f. Expliquer pourquoi δ_c n'est jamais à l'intérieur de S_c alors que π_c se concentre sur S_c. Est-ce que δ_c est le "vrai" estimateur de Bayes ?

g. En utilisant les relations de récurrence de l'Exercice 4.36, montrer que

$$\delta_c(x) = \left(1 - \frac{p-2}{||x||^2}\right) x + h_c(||x||^2)x,$$

où $h_c(t) > 0$ lorsque $t \leq \max(c^2, p-2)$. Proposer un estimateur plus intéressant.

4.38 Soit $x_1, \ldots, x_{10}$ i.i.d. $\mathcal{N}(\theta, \theta^2)$, avec $\theta > 0$, représentant dix observations de la vitesse d'une étoile. Justifier le choix $\pi(\theta) = 1/\theta$ et déterminer l'estimateur de Bayes généralisé associé à un coût invariant

$$\mathrm{L}(\theta, \delta) = \left(\frac{\delta}{\theta} - 1\right)^2 .$$

(*Indication :* Utiliser l'Exercice 3.33.)

4.39 *(Lindley, 1965) Soit $x_1, \ldots, x_n$ un échantillon de $\mathcal{N}(\theta, \sigma^2)$, avec σ^2 connu. La densité a priori $\pi(\theta)$ est telle qu'il existe ϵ, M et c tels que $c(1-\epsilon) \leq \pi(\theta) \leq c(1+\epsilon)$ pour $\theta \in I = [\bar{x} - 1.96\,\sigma/\sqrt{n}, \bar{x} + 1.96\,\sigma/\sqrt{n}]$ et $\pi(\theta) \leq Mc$ sinon.

a. Montrer que ces contraintes sont compatibles, c'est-à-dire qu'une telle loi a priori existe.

b. Montrer que

$$(1-\epsilon)[0.95(1+\epsilon) + 0.05M]^{-1} \frac{e^{-(x-\theta)^2 n/2\sigma^2}}{\sqrt{2\pi\sigma^2/n}} \leq \pi(\theta|x)$$
$$\leq (1+\epsilon)[(1-\epsilon)0.95]^{-1} \frac{e^{-(x-\theta)^2 n/2\sigma^2}}{\sqrt{2\pi\sigma^2/n}}$$

si $\theta \in I$ et

$$\pi(\theta|x) \leq \frac{M}{0.95(1-\epsilon)} \frac{e^{-1.96^2/2}}{\sqrt{2\pi\sigma^2/n}}$$

sinon.

c. Discuter de l'intérêt de ces approximations pour $\theta \in I$ et $\theta \notin I$. Pouvez-vous obtenir une région de confiance conservatrice ?

4.40 Soient une variable aléatoire normale, $x \sim \mathcal{N}(\theta, 1)$ et une transformation bijective $\eta = \sinh(\theta)$.

a. Lorsque $\pi(\eta) = 1$, montrer que la distribution a posteriori résultante sur θ est

$$\pi(\theta|x) \propto e^x \mathcal{N}(x+1, 1) + e^{-x} \mathcal{N}(x-1, 1).$$

b. Comparer le comportement de cette loi a posteriori avec celui de la loi a posteriori de Jeffreys $\mathcal{N}(x, 1)$ en calculant les variance, quantiles et modes a posteriori. En particulier, déterminer les valeurs de x pour lesquelles la loi a posteriori est bimodale et celles pour lesquelles il y a deux maxima globaux.

c. Considérer le comportement de $\pi(\theta|x)$ pour de grandes valeurs de x et conclure que la loi a priori $\pi(\eta) = 1$ n'est pas un choix raisonnable.

Section 4.4.2

4.41 (Jeffreys, 1961) Soient $x_1, \ldots, x_{n_1}$ i.i.d. de loi $\mathcal{N}(\theta, \sigma^2)$ et $\bar{x}_1$, s_1^2 les statistiques associées. Pour un second échantillon d'observations de même taille, donner la distribution prédictive de $(\bar{x}_2, s_2^2)$ sous la loi non informative $\pi(\theta, \sigma) = \frac{1}{\sigma}$. Si $s_2^2 = s_1^2/y$ et $y = e^z$, en déduire que z suit la loi de Fischer.

4.42 Montrer que, si $x \sim \mathcal{G}(\alpha, \beta)$, $1/x \sim \mathcal{IG}(\alpha, \beta)$ comme défini dans (4.12).

4.43 *(Ghosh et Yang, 1996) Comme dans l'Exercice 3.47, considérer $x_{11}, \ldots, x_{1n_1}$ et $x_{21}, \ldots, x_{2n_2}$, deux échantillons indépendants avec $x_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$.

a. Montrer que la matrice d'information de Fisher est

$$\mathbf{I}(\mu_1,\mu_2,\sigma) = \sigma^{-2}\begin{pmatrix} n_1 & 0 & 0 \\ 0 & n_2 & 0 \\ 0 & 0 & 2(n_1+n_2) \end{pmatrix}.$$

b. La loi a priori coïncidente de Welch et Peers (1963) (voir la Section 3.5.5) pour la quantité d'intérêt $\theta = (\mu_1 - \mu_2)/\sigma$ est solution de l'équation différentielle

$$\frac{\partial}{\partial\mu_1}(\eta_1\pi) + \frac{\partial}{\partial\mu_2}(\eta_2\pi) + \frac{\partial}{\partial\sigma}(\eta_3\pi) = 0\,, \tag{4.36}$$

où

$$(\eta_1,\eta_2,\eta_3) = \mathbf{I}^{-1}\nabla\theta/(\nabla\theta^t\mathbf{I}^{-1}\nabla\theta)^{1/2}\,.$$

Montrer qu'une classe de solutions à (4.36) est de la forme

$$\left[n_1^{-1} + n_2^{-1} + \frac{1}{2}(\mu_1-\mu_2)^2/\{(n_1+n_2)\sigma^2\}\right]^{1/2} g(\mu_1,\mu_2,\sigma) \tag{4.37}$$

où

$$g(\mu_1,\mu_2,\sigma) \propto \left[d_1(\mu_1-\mu_2)^2 + d_2(n_1\mu_1^2 + n_2\mu_2^2) + d_3\sigma^2\right]^c\,,$$

c est une constante arbitraire et (d_1,d_2,d_3) satisfont

$$d_1(n_1^{-1}+n_2^{-1}) + d_2 = \frac{1}{2}d_3(n_1+n_2)^{-1}\,.$$

c. En déduire que la loi a priori coïncidente pour (θ,μ_2,σ) est

$$\pi(\theta,\mu_2,\sigma) \propto \sigma^{2c+1}\left[n_1^{-1}+n_2^{-1}+\frac{1}{2}\theta^2(n_1+n_2)^{-1}\right]^{c+1/2}\,.$$

d. Montrer que

$$\pi(\theta|\bar{x}_1,\bar{x}_2,s) \propto \left[n_1^{-1}+n_2^{-1}+\frac{1}{2}\theta^2(n_1+n_2)^{-1}\right]^{c+1/2} \times \int_0^\infty v^{n_1+n_2-2c-4}\exp\left\{\frac{-1}{2}\left(v^2 + \frac{n_1n_2}{n_1+n_2}(vz-\theta)^2\right)\right\}dv$$

où $z = (\bar{x}_1-\bar{x}_2)/s$.

e. Montrer que la distribution de z ne dépend que de θ.

f. Montrer que le choix unique de c qui évite le paradoxe de marginalisation des Exercices 3.45-3.51 est $c = -1$.

Section 4.4.3

4.44 a. Si $x \sim \mathcal{N}_p(\theta,\Sigma)$, montrer que, pour toute loi a priori π,

$$\delta^\pi(x) = x + \Sigma\nabla\log m_\pi(x).$$

b. (Bock, 1988) Les *pseudo-estimateurs de Bayes* sont définis comme les estimateurs de la forme

$$\delta(x) = x + \nabla\log m(x)$$

où $x \sim \mathcal{N}_p(\theta, I_p)$. Montrer que l'estimateur de James-Stein tronqué donné dans l'Exemple 4.9 est un pseudo-estimateur de Bayes (c'est-à-dire donner la valeur correspondante de m). Peut-il s'agir d'un estimateur de Bayes ?

4.45 *Pour un modèle normal $\mathcal{N}_k(X\beta, \Sigma)$ où la matrice de covariance Σ est complètement inconnue, donner la loi a priori non informative de Jeffreys.

a. Montrer que la loi a posteriori de Σ, conditionnelle à β, est une distribution de Wishart et en déduire qu'il n'existe pas de loi marginale a posteriori propre sur β lorsque le nombre d'observations est inférieur à k.

b. Expliquer alors pourquoi il n'est pas possible de construire une loi conjuguée. Considérer le cas particulier où Σ suit une loi de Wishart.

c. Quelle est la raison fondamentale pour laquelle ce qui était possible dans la Section 4.4.2 ne l'est plus pour ce modèle ?

4.46 *Soit le problème de la prédiction pour un modèle de régression linéaire, avec $y = X\beta + \epsilon$ observé, $\beta \in \mathbb{R}^k$, $\epsilon \sim \mathcal{N}_p(0, \Sigma)$. On cherche à prédire $z = T\beta + \epsilon'$, avec T connu et $\epsilon' \sim \mathcal{N}_p(0, \Sigma)$ indépendant de ϵ.

a. Si δ est le prédicteur considéré et si l'erreur de prédiction est évaluée par la fonction de coût $\mathrm{L}(z, \delta) = ||z - \delta||^2$, montrer que l'erreur moyenne est

$$\mathbb{E}^{z,x}[\mathrm{L}(z, \delta(x))] = \mathrm{tr}(\Sigma) + \mathbb{E}^x[||\delta(x) - T\beta||^2].$$

b. Montrer que ce problème est équivalent à celui de l'estimation de β sous le coût quadratique associé à $Q = T^tT$. (*Indication :* Montrer auparavant que $\delta(x)$ est forcément de la forme $T\gamma(x)$, avec $\gamma(x) \in \mathbb{R}^k$, ou qu'il est dominé par un tel estimateur.)

c. Déduire du fait que Q est dégénérée et admet une seule valeur propre non nulle qu'un effet Stein ne peut pas avoir lieu dans un tel cas.

d. Considérer maintenant que T est une matrice aléatoire, de moyenne 0 et telle que $\mathbb{E}[T^tT] = M$. Montrer que, lorsque $\delta(x) = T\gamma(x)$, le risque fréquentiste est

$$\mathbb{E}^{z,x,T}[\mathrm{L}(z, \delta(x))] = \mathrm{tr}(\Sigma) + \mathbb{E}^x[(\gamma(x) - \beta)^t M(\gamma(x) - \beta)]\,,$$

et donc qu'un effet Stein est possible lorsque M a trois valeurs propres non nulles ou plus. [*Note :* Ce phénomène est relié aux paradoxes de statistiques libres développés par Brown, 1986a ; voir aussi Foster et George, 1998.]

e. Soit $\beta \sim \mathcal{N}_k(0, \sigma^2 I_k)$. Calculer le prédicteur de Bayes de z lorsque T est fixé et lorsque T est aléatoire. Conclure.

4.47 Les *modèles tobit* sont utilisés en Économétrie (voir Gouriéroux et Monfort, 1996) pour représenter des phénomènes tronqués. Soit $y|x \sim \mathcal{N}(\beta^t x, \sigma^2)$, qui n'est observé que s'il est strictement positif, x étant une variable explicative dans $\mathbb{R}^p$.

a. Montrer que les modèles tobit sont des mélanges de modèles probit (pour $y < 0$) et de modèles de régression standard (pour $y \geq 0$).

b. Donner la fonction de vraisemblance $\ell(\beta, \sigma^2|y_1, \ldots, y_n)$ associée à l'échantillon $y_1, \ldots, y_n, x_1, \ldots, x_n$ et calculer une statistique exhaustive pour ce modèle.

c. Conditionnellement à $(x_1, \ldots, x_n)$, montrer que ce modèle appartient à une famille exponentielle et proposer une loi a priori conjuguée pour (β, σ). Est-ce que cette loi permet des calculs analytiques ?

4.48 *Le modèle de *régression inverse* (ou *calibration*) est donné par

$$y \sim \mathcal{N}_p(\beta, \sigma^2 I_p), \qquad z \sim \mathcal{N}_p(\lambda_0\beta, \sigma^2 I_p), \qquad s^2 \sim \sigma^2\chi^2_q,$$

avec $\beta \in \mathbb{R}^p$, $\lambda_0 \in \mathbb{R}$.

a. Donner l'estimateur du maximum de vraisemblance de λ et montrer que son risque quadratique peut être infini.

b. Calculer la loi a priori de Jeffreys pour $(\beta, \sigma^2, \lambda_0)$ et montrer que l'espérance a posteriori correspondante de λ_0 est l'*estimateur de régression inverse*, $\delta^I(y,z,s) = y^t z/(s + ||y||^2)$.

c. En recourant à la technique des *lois a priori de référence* introduite dans la Section 3.5, proposer une loi a priori alternative $\pi(\{\lambda_0, (\beta, \sigma^2)\})$ lorsque (β, σ^2) est considérée comme un paramètre de nuisance. Calculer l'espérance a posteriori correspondante de λ_0, $\delta^R(y,z,s)$.

d. Montrer que, lorsque q tend vers l'infini, δ^I converge presque sûrement vers 0, mais que δ^R ne souffre pas de cette incohérence. [*Note :* Voir Osborne, 1991, pour une revue des modèles de calibration, et Kubokawa et Robert, 1994, pour des considérations décisionnelles sur ces estimateurs.]

Section 4.5.1

4.49 Pour le modèle AR(1) donné par (4.18), donner la matrice de covariance de $(x_1, \ldots, x_T)$.

4.50 (Suite de l'Exercice 4.49)

a. Montrer que la variance de x_t est donnée par (4.19).

b. Que se passe-t-il dans le cas où $\varrho = 1$, où (4.19) n'a pas de sens ?

c. Étendre au cas où x_0 est une valeur arbitraire.

4.51 ***(Suite de l'Exercice 4.50)** On souhaite établir qu'il n'existe pas de loi stationnaire pour le modèle AR(1) lorsque $|\varrho| \geq 1$, c'est-à-dire pas de densité f telle que, si $x_t \sim f$, alors $x_{t+1} \sim f$.

a. Montrer que, lorsque $|\varrho| < 1$, la loi stationnaire est la distribution normale $\mathcal{N}(0, \sigma^2/(1-\varrho^2))$.

b. Dans le cas où $|\varrho| = 1$, montrer que la mesure de Lebesgue est la mesure stationnaire de la chaîne (x_t), c'est-à-dire pour tout ensemble mesurable A,

$$\int_A dx = \int_A \int f(y|x) dx dy\,,$$

où $f(y|x)$ est la loi conditionnelle de x_t sachant x_{t-1}, soit $\mathcal{N}(x_{t-1}, \sigma^2)$ dans ce cas. Déduire de l'unicité de la mesure stationnaire la non-existence d'une loi de probabilité stationnaire.

c. Étendre au cas $|\varrho| \geq 1$, en écrivant x_t comme

$$x_t = \sum_{i=0}^{t-1} \varrho^i \epsilon_{t-i} + \varrho^t x_0$$

et en déduisant que x_t est infini presque sûrement lorsque t tend vers l'infini. (*Indication :* Pour $x_0 = 0$, remplacer la décomposition ci-dessus avec la décomposition correspondante conditionnellement à x_1.)

Section 4.5.2

4.52 (Bernardo et Smith, 1994) Montrer que, pour un vecteur bidimensionnel,

$$(x_1\ x_2)^t \sim \mathcal{N}_2\left((\mu_1\ \mu_2)^t, \begin{bmatrix} \sigma_1^2 & \varrho\sigma_1\sigma_2 \\ \varrho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right),$$

la loi a priori de Jeffreys est $\pi(\theta) \propto (1-\varrho^2)^{-1}/\sigma_1\sigma_2$.

4.53 (Bauwens *et al.*, 1999) Pour le modèle AR(1) donné par (4.18),

a. Montrer que μ est un paramètre de position et, donc, qu'il n'apparaît pas dans la loi a priori de Jeffreys.

b. Montrer que

$$\mathbb{E}\left[\frac{\partial^2 \log L(\theta|x_{1:T})}{\partial\sigma^2}\right] = \frac{-T}{2\sigma^4}\,, \quad \mathbb{E}\left[\frac{\partial^2 \log L(\theta|x_{1:T})}{\partial\varrho^2}\right] = \frac{-1}{\sigma^2}\mathbb{E}\left[\sum_{t=0}^{T-1} x_t^2\right]\,.$$

c. En utilisant la loi stationnaire des y_t, déduire de $\mathbb{E}[y_t^2] = \sigma^2/(1-\varrho^2)$ la loi a priori de Jeffreys $\pi_1^J(\sigma^2,\varrho) = 1/\sigma^2\sqrt{1-\varrho^2}$.

4.54 *(Brockwell et Davis, 1998) L'algorithme de Durbin-Levinson calcule les autocorrélations partielles comme suit : soit $\phi_{n1},\ldots,\phi_{nn}$ défini récursivement à partir des autocovariances $\gamma(s)$ par

$$\phi_{nn} = \left(\gamma(n) - \sum_{j=1}^{n-1}\phi_{(n-1)j}\gamma(n_j)\right) v_{n-1}^{-1}$$

et

$$\begin{pmatrix}\phi_{n1}\\ \vdots \\ \phi_{n(n-1)}\end{pmatrix} = \begin{pmatrix}\phi_{(n-1)1}\\ \vdots \\ \phi_{(n-1)(n-1)}\end{pmatrix} - \phi_{nn}\begin{pmatrix}\phi_{(n-1)(n-1)}\\ \vdots \\ \phi_{(n-1)1}\end{pmatrix}\,,$$

où $v_n = v_{n-1}(1-\phi_{nn})^2$, $\phi_{11} = \gamma(1)/\gamma(0)$ et $v_0 = \gamma(0)$.

a. Montrer que, si $\psi_n = \phi_{nn}$, l'inverse de l'algorithme de Durbin-Levinson s'obtient à partir du Lemme 4.24.

b. Montrer que les autocorrélations partielles ψ_n d'un processus MA(q) sont nulles pour $n > q$.

c. Montrer que les autocorrélations partielles ψ_n d'un processus AR(1) sont données par $\psi_n = (-1)^{n+1}\vartheta_1^n/(1+\vartheta_1^2+\ldots+\vartheta_1^{2n})$.

4.55 (Bauwens *et al.*, 1999) Pour le modèle AR(1) représenté dans (4.18),

a. En utilisant la décomposition de Wold (4.25) obtenue dans l'Exemple 4.25, montrer que

$$\mathbb{E}\left[x_t^2\right] = \mathbb{E}\left[\left(\varrho^t x_0 + \sum_{i=0}^{t-1}\varrho^i\epsilon_{t-i}\right)^2\right] = \frac{1-\varrho^{2t}}{1-\varrho^2}\sigma^2$$

avec $x_0 = 0$.

b. En déduire la loi a priori de Jeffreys π_2^J.

Section 4.5.3

4.56 *Donner la décomposition de Wold pour le modèle stationnaire AR(p). (*Indication :* Utiliser la *représentation par polynôme retard* du modèle AR(p), soit, $\mathscr{P}(B)x_t = \epsilon_t$, où $B^d x_t = x_{t-d}$.)

4.57 Montrer que les autocorrélations γ_s du modèle MA(q) sont données par (4.27).

4.58 Établir la représentation (4.32). Généraliser la représentation de l'Exemple 4.27 au modèle général MA(q).

Section 4.5.4

4.59 *Un modèle ARMA(p,q)

$$x_t - \mu = \sum_{i=1}^{p} \varrho_i (x_{t-i} - \mu) + \sum_{j=1}^{q} \vartheta_j \epsilon_{t-j} + \epsilon_t \,,$$

est *inversible* (Brockwell et Davis, 1998, Section 3.1) s'il existe une suite $(\varpi_j)_j$ telle que

$$\sum_i |\varpi_j| < \infty \quad \text{et} \quad \epsilon_t = \sum_{j=0}^{\infty} \varpi_j x_{t-j} \,.$$

Montrer que l'inversibilité est équivalente à la condition que $\mathscr{Q}(x)$ ait ses racines hors du cercle unité. (*Indication :* Utiliser la *représentation polynôme retard* du modèle ARMA(p,q), c'est-à-dire $\mathscr{P}(B)x_t = \mathscr{Q}(B)\epsilon_t$, avec $B^d x_t = x_{t-d}$.)

4.60 *Un modèle ARMA(p,q) est dit *causal* (Brockwell et Davis, 1998, Section 3.1) s'il existe une suite $(\varphi_j)_j$ telle que

$$\sum_i |\varphi_j| < \infty \quad \text{et} \quad x_t = \sum_{j=0}^{\infty} \varphi_j \epsilon_{t-j} \,.$$

Montrer que la causalité est équivalente à la condition que $\mathscr{P}(x)$ ait ses racines hors du cercle unité. (*Indication :* Utiliser la représentation polynôme retard de l'Exercice 4.59.)

4.61 Montrer que la représentation (4.34) est vérifiée. Proposer une représentation alternative.

4.62 Proposer une représentation à espace d'états similaire à (4.34) pour le *modèle ARIMA*

$$z_t - \mu = \sum_{i=1}^{p} \varrho_i (z_{t-i} - \mu) + \sum_{j=1}^{q} \vartheta_j \epsilon_{t-j} + \epsilon_t \,, \tag{4.38}$$

où z_t est la série différenciée, $z_t = x_t - x_{t-d}$, $d \in \mathbb{N}^*$. [*Note :* Comme Brockwell et Davis, 1998, Section 6.5 le détaille, le modèle général ARIMA(p,d,q) est donné par un modèle ARMA(p,q) sur les séries différenciées $x_t - \Psi_1 x_{t-d} - \ldots - \Psi_P x_{t-Pd}$.]

Note 4.7.1

4.63 (Deely et Gupta, 1968) Soient $x_1 \sim \mathscr{N}(\theta_1, \sigma_1^2), \ldots, x_k \sim \mathscr{N}(\theta_k, \sigma_k^2)$ où la quantité d'intérêt est $\theta_{[k]}$, la plus grande des moyennes $\theta_1, \ldots, \theta_k$. La fonction de coût est $\mathrm{L}(\theta, \varphi) = \theta_{[k]} - \varphi$.

a. Montrer que, si $\sigma_1 = \ldots = \sigma_k$ sont connues et $\pi(\theta_1) = \ldots = \pi(\theta_k) = 1$, l'estimateur de Bayes sélectionne la population comportant la plus grande observation.

b. Généraliser au cas où les θ_i ont une loi a priori échangeable $\mathscr{N}(0, \tau^2)$.

4.64 *(Goel et Rubin, 1977) Montrer que les ensembles s_j^* constituent véritablement une classe complète lorsque la loi a priori sur $\theta = (\theta_1, \ldots, \theta_k)$ est symétrique. (*Indication :* Montrer que les s_j^* sont optimaux parmi les sous-ensembles de taille $|s_j^*|$.)

4.65 (Suite de l'Exercice 4.64) Étendre ce résultat aux lois $f(x|\theta)$ à rapport de vraisemblance monotone en θ.

4.66 (Chernoff et Yahav, 1977) Étendre le résultat de classe complète de l'Exercice 4.64 à la fonction de coût

$$\mathrm{L}(\theta, s) = c(\theta_{[k]} - \theta_s) - \frac{1}{s}\sum_{j\in s}\theta_j.$$

(*Indication :* Montrer que, si $\theta_{i_1} \leq \ldots \leq \theta_{i_j}$, $s = \{i_1, \ldots, i_j\}$ est dominé par l'ensemble $\{i_j\}$.)

Note 4.7.2

4.67 *Pour le modèle AR(1) donné par (4.18), supposons que la quantité d'intérêt soit x_0, la valeur de départ de la chaîne. Calculer la loi a priori de référence pour l'ordre $\{x_0, (\varrho, \sigma^2)\}$ et calculer un estimateur de x_0 sous le coût quadratique.

Note 4.7.3

4.68 Soit le modèle à facteurs ($t = 1, \ldots, T$),

$$\begin{cases} y_t^* = [\alpha + \beta(y_{t-1}^*)^2]^{1/2}\epsilon_t^* \\ y_t = y_t^*\mu + \sigma\epsilon_t, \end{cases} \tag{4.39}$$

avec $\epsilon_t^* \sim \mathcal{N}(0,1)$, et où seuls les $y_t \in \mathbb{R}^p$ sont observés.

a. Écrire la vraisemblance (complète) associée aux couples (y_t, y_t^*).

b. Montrer que les y_t^* ne peuvent pas être marginalisés analytiquement.

c. En déduire que le modèle à facteurs ne peut pas s'exprimer comme un cas particulier de modèle ARCH donné par (4.40).

4.69 (Bauwens *et al.*, 1999) Montrer que le modèle ARCH(p) est sans intérêt lorsque $\alpha = 0$. (*Indication :* Montrer que $\mathrm{var}(y_t) = 0$.)

4.7 Notes

4.7.1 Classement et sélection

Beaucoup d'efforts ont été consacrés au problème d'estimation et de comparaison de plusieurs moyennes normales. Nous mentionnons brièvement ici quelques approches proposées, afin d'illustrer l'intérêt d'un traitement bayésien, et renvoyons les lecteurs à la littérature pour une discussion plus détaillée ; voir, par exemple, Gibbons *et al.* (1977) Gupta and Panchapakesan (1979) et Dudewicz et Koo (1982), à la suite des articles introductifs de Bechofer (1954) et Gupta (1965). Comme le décrivent Berger et Deely (1988), les techniques de classement et de sélection apparaissent aussi comme des substituts de l'*analyse de la variance* (Chapitre 10).

Pour $x_1 \sim \mathcal{N}(\theta_1, \sigma_1^2)$, ..., $x_k \sim \mathcal{N}(\theta_k, \sigma_k^2)$ donnés, on cherche à sélectionner la population d'espérance la plus élevée, $\theta_{[k]}$. Les variances $\sigma_1^2, \ldots, \sigma_k^2$ sont ici supposées connues, mais le cadre plus général où elles sont estimées par $\hat\sigma_1^2, \ldots, \hat\sigma_k^2$ peut aussi être traité par le paradigme bayésien. Berger et Deely (1988) reformulent ce problème pour répondre aux questions suivantes : (a) Peut-on accepter l'hypothèse $H_0 : \theta_1 = \cdots = \theta_k$? (b) Dans le cas d'une réponse négative, quelle est la moyenne la plus élevée ? Ils résolvent ce problème en calculant d'abord le facteur de Bayes contre H_0, puis les probabilités a posteriori p_j que θ_j soit

la moyenne la plus élevée ($1 \leq j \leq k$). (Le Chapitre 5 traite de la définition et du calcul de ces quantités.) Pour ce faire, ils recourent à des lois a priori hiérarchiques (Chapitre 10)

$$\theta_i|\beta, \sigma_\pi^2 \sim \mathcal{N}(\beta, \sigma_\pi^2), \quad \beta \sim \mathcal{N}(\beta^0, A) \quad \text{et} \quad \sigma_\pi^2 \sim \gamma \mathbb{I}_0(\sigma_\pi^2) + (1-\gamma)\pi_{22}^*(\sigma_\pi^2).$$

La structure particulière de la loi a priori sur σ_π^2 est due à la nécessité de tester le fait que les θ_i sont identiques. Pour π_{22}^*, Berger et Deely (1988) proposent la loi *informative*

$$\pi_{22}^*(\sigma_\pi^2) = (m-1)C(1+C\sigma_\pi^2)^{-m},$$

où C et m peuvent s'obtenir à partir de quantiles a priori. Pour une loi a priori *non informative*, des choix possibles sont $\pi_{22}^*(\sigma_\pi^2) = 1$ et

$$\pi_{22}^*(\sigma_\pi^2) = \prod_{i=1}^{k} (\sigma_i^2 + \sigma_\pi^2)^{-1/k},$$

bien que ces lois a priori puissent rendre difficile le calcul de la probabilité a posteriori de H_0 (voir le Chapitre 5).

Goel et Rubin (1977) adoptent une perspective plus décisionnelle, considérant comme un espace de décision $\mathscr{D}$ l'ensemble de toutes les sous-parties non vides de $\{1, \ldots, k\}$, noté $\{s_1, s_2, \ldots, s_K\}$ avec $K = 2^k - 1$. Ils introduisent la fonction de coût

$$\mathrm{L}(\theta, s) = c|s| + \theta_{[k]} - \theta_s,$$

où $|s|$ est le cardinal de s et $\theta_s = \max_{j \in s} \theta_j$. Ce coût comprend une pénalité c pour toute population comprise dans l'ensemble de décision s. Ce qui est plutôt logique, puisque, par souci de parcimonie, l'ensemble de décision s doit être choisi aussi petit que possible, le cas idéal étant $|s| = 1$. Goel et Rubin (1977) ont montré d'abord qu'une règle bayésienne associée à cette fonction de coût et à une loi a priori symétrique doit être choisie parmi les ensembles $s_j^* = \{\omega_k, \ldots, \omega_{k-j+1}\}$ ($1 \leq j \leq k$), où ω_j est la population des $x_{(j)}$. La règle bayésienne s^π est alors solution de

$$\varrho(\pi, s^\pi|x) = \min_{j=1,\ldots,k} \varrho(\pi, s_j^*|x),$$

où $\varrho(\pi, s|x) = c|s| + \mathbb{E}^\pi[\theta_{[k]} - \theta_s|x]$. Introduisant

$$\Delta_m = \varrho(\pi, s_{m+1}^*|x) - \varrho(\pi, s_m^*|x) \qquad (1 \leq m \leq k-1),$$

la règle bayésienne vaut s_k^* si $A = \{j;\, \Delta_j \geq 0\}$ est vide, s_m^* sinon, avec $m = \min(A)$. Un point délicat dans l'obtention de s^π est bien entendu le calcul des espérances a posteriori $\mathbb{E}^\pi[\theta_{[k]} - \theta_s|x]$. Ces auteurs détaillent le cas particulier d'une loi a priori normale échangeable pour les θ_j, qui reste dépendante de la fonction

$$t_m(z) = \int_{-\infty}^{+\infty} \Phi^m(z+x)\Phi(-x)\, dx.$$

Cependant, ils montrent que, dans le cas non informatif, pour $\sigma_1 = \cdots = \sigma_k$, la règle bayésienne est s_1^* lorsque $c/\sigma_1 \geq 1/\pi^2$.

4.7.2 Loi de Jeffreys pour un modèle AR(1)

La loi a priori de Jeffreys porte à controverse dans ce cas, à cause du débat sur la prise en compte ou non de la condition de stationnarité et des différences qui en résulte. Si nous supposons $x_t = \mu + \varrho(x_{t-1} - \mu) + \epsilon_t$ avec $x_0 = 0$, l'a priori de Jeffreys associé à cette représentation stationnaire est (Exercice 4.53)

$$\pi_1^J(\mu, \sigma^2, \varrho) \propto \frac{1}{\sigma^2} \frac{1}{\sqrt{1-\varrho^2}}\,.$$

Lorsque la région de non stationnarité $|\varrho| > 1$ est incluse, Phillips (1991) montre que l'a priori de Jeffreys est alors (Exercice 4.55)

$$\pi_2^J(\mu, \sigma^2, \varrho) \propto \frac{1}{\sigma^2} \frac{1}{\sqrt{|1-\varrho^2|}} \sqrt{\left|1 - \frac{1-\varrho^{2T}}{T(1-\varrho^2)}\right|}\,.$$

Bien que $\pi_2^J(\mu, \sigma^2, \varrho)$ soit équivalent à $\pi_1^J(\mu, \sigma^2, \varrho)$ pour des valeurs élevées de T et $|\varrho| < 1$, la partie dominante de la loi a priori correspond à la région de non stationnarité, puisqu'elle est équivalente à ϱ^{2T} (Bauwens *et al.*, 1999). Berger et Yang (1994) ont aussi montré que la loi a priori de référence est π_1^J et qu'elle n'est définie que lorsque la contrainte de stationnarité est vérifiée. Ils suggèrent alors de symétriser cette loi a priori sur la région $|\varrho| > 1$, posant

$$\pi^B(\mu, \sigma^2, \varrho) \propto \frac{1}{\sigma^2} \begin{cases} 1/\sqrt{1-\varrho^2} & \text{si } |\varrho| < 1, \\ 1/|\varrho|\sqrt{\varrho^2-1} & \text{si } |\varrho| > 1, \end{cases}$$

qui a une forme plus raisonnable que π_2^J, comme le montre la Figure 4.3.

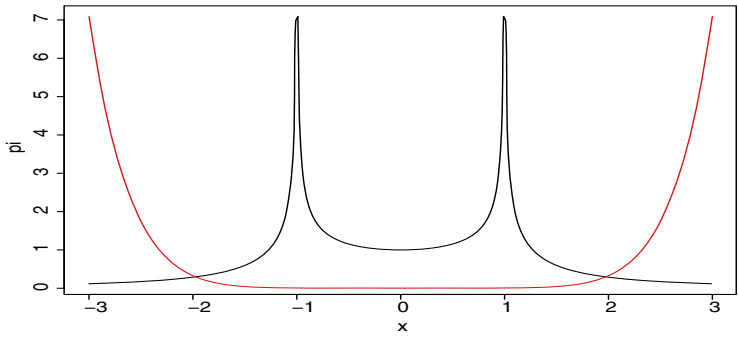

Fig. 4.3. Graphes des lois a priori $\pi_1^J(\varrho)$ et $\pi^B(\varrho)$ pour $T = 10$.

Comme le détaillent Bauwens et al. (1999, Section 6.8), il est aussi possible de construire des lois a priori de Jeffreys dans les cas stationnaire et non stationnaire lorsqu'on prend en compte la loi de la valeur initiale x_0, ce qui donne des lois similaires à π_1^J et π_2^J.

4.7.3 Modèles ARCH

Les modèles ARCH, introduits par Engle (1982), sont utilisés, notamment en Finance, pour représenter des processus dont les termes d'erreur sont indépendants et de variance non constante dans le temps ; un processus ARCH(p) par exemple se définit comme

$$x_t = \sigma_t \epsilon_t \,, \qquad \sigma_t^2 = \alpha + \sum_{i=1}^{p} \beta_i x_{t-i}^2 \,, \tag{4.40}$$

où les ϵ_t sont i.i.d. $\mathcal{N}(0,1)$. L'acronyme ARCH signifie *autoregressive conditional heterocedasticity*, le dernier terme étant utilisé par les économètres pour qualifier les modèles à variance non constante. Gouriéroux (1997) décrit ces modèles en détail, ainsi que les méthodes inférentielles classiques correspondantes ; voir Bauwens et al. (1999, Section 7.4) pour des extensions bayésiennes aux processus GARCH (pour *generalised ARCH*).

Comme le montrent Nelson (1990) et Kleibergen et Van Dijk (1993), une condition de stationnarité pour un modèle ARCH(1) est que $\mathbb{E}[\log(\beta_1 \epsilon_t^2)] < 0$, ce qui est équivalent à $\beta_1 < 3.4$.

Au contraire des modèles à volatilité stochastique de la Note 4.7.4, les modèles ARCH(p) bénéficient de fonctions de vraisemblance exprimables analytiquement, conditionnellement aux valeurs initiales $x_1, \ldots, x_p$. Les non-linéarités dans les termes de variance requièrent cependant l'utilisation de méthodes d'approximation comme celles du Chapitre 6.

4.7.4 Modèles à volatilité stochastique

Les modèles à volatilité stochastique s'appliquent à décrire la *volatilité*, $\log(\sigma_t^2)$, d'une série x_t d'une variable aléatoire. Bien que de tels modèles soient plus complexes à étudier que leurs contreparties ARCH, ils sont souvent utilisés en Finance pour modéliser des séries présentant des variations d'échelle brusques (voir, par exemple, Jacquier *et al.*, 1994).

Une illustration simple de ces modèles est le cas SV(1), où ($t = 1, \ldots, T$)

$$\begin{cases} y_t^* = \alpha + \varrho y_{t-1}^* + \sigma \epsilon_{t-1}^* \,, \\ y_t = e^{y_t^*/2} \epsilon_t \,, \end{cases} \tag{4.41}$$

et où les ϵ_t et ϵ_t^*'s sont i.i.d. $\mathcal{N}(0,1)$. La quantité non observée (y_t^*) représente donc la *volatilité*. (Une hypothèse courante sur la condition initiale est que $y_0^* \sim \mathcal{N}(\alpha, \sigma^2)$.) La Figure 4.4 représente une série simulée de volatilités stochastiques pour $\sigma = 1$ et $\varrho = .9$.

La difficulté avec ce modèle est que l'information relative aux paramètres $(\alpha, \varrho, \sigma)$ est contenue dans les volatilités non observées. En effet, conditionnellement à y_t^*, ces volatilités sont indépendantes de y_t. (Bien entendu, les paramètres dépendent bien des données, au moins marginalement.) De plus, la vraisemblance observée $L(\alpha, \varrho, \sigma | y_0, \ldots, y_T)$ n'admet pas d'expression analytique, puisque les y_t^* ne peuvent pas être marginalisés explicitement. En revanche, la vraisemblance complète est explicite, soit

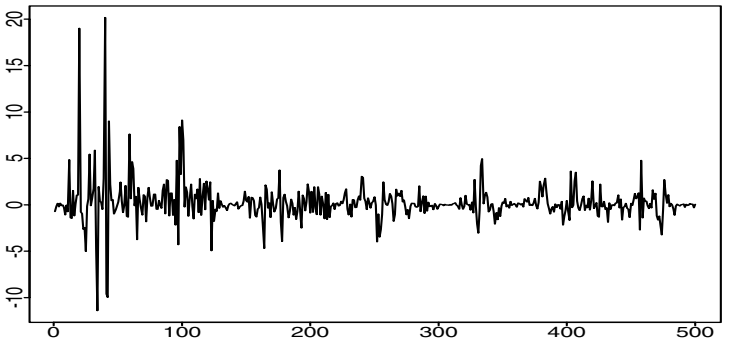

Fig. 4.4. Échantillon simulé du modèle de volatilité stochastique (4.41) avec $\sigma = 1$ et $\varrho = .9$. (*Source :* Robert et Casella, 1999.)

$$L^c(\alpha, \varrho, \sigma | y_0, y_0^* \ldots, y_T, y_T^*) \propto \tag{4.42}$$
$$\sigma^{-T+1} \exp - \left\{ (y_0^* - \alpha)^2 + \sum_{t=1}^{T} (y_t^* - \alpha - \varrho y_{t-1}^*)^2 \right\} \Big/ 2\sigma^2$$
$$\exp - \sum_{t=0}^{T} \left\{ y_t^2 e^{-y_t^*} + y_t^* \right\} /2 \, .$$

Ceci peut alors être utilisé dans des méthodes simulées (Chapitre 6), en alternance avec la simulation des volatilités non observées y_t^*. La Figure 4.5 illustre une telle simulation pour le jeu de données simulé de la Figure 4.4, tel que les valeurs des y_t^* sont connues. (L'image floue au-dessus du graphe est appelée *carte d'allocation* et représente les valeurs successives des y_t^* comme des niveaux de gris correspondants aux itérations de la méthode simulée utilisée.)

4.7.5 Lois a priori poly-t

Les lois a priori *poly-t* ont été proposées par Drèze (1976b) et Richard et Tompa (1980) comme une alternative robuste aux lois conjuguées pour les modèles de régression linéaire. Leur motivation est donnée par l'exemple suivant, développé par Bauwens et al. (1999, Section 4.5). Considérons deux régressions indépendantes,

$$y_1 = X_1\beta + \sigma_1\varepsilon_1, \; y_2 = X_2\beta + \sigma_2\varepsilon_2, \; \varepsilon_1 \sim \mathscr{N}_{T_1}(0, I_{T_1}), \; \varepsilon_2 \sim \mathscr{N}_{T_2}(0, I_{T_2}).$$

Si $\pi(\beta, \sigma_1, \sigma_2) = 1/\sigma_1\sigma_2$, l'intégration des variances σ_i donne la loi a posteriori marginale dite $2 - 0$ *poly-t*

$$\pi(\beta | y_1, y_2) \propto [S_1 + (\beta - \hat{\beta}_1)^t M_1 (\beta - \hat{\beta}_1)]^{-T_1/2}$$
$$\times [S_2 + (\beta - \hat{\beta}_2)^t M_2 (\beta - \hat{\beta}_2)]^{-T_2/2} \, ,$$

où $\hat{\beta}_i$ est l'estimateur des moindres carrés ordinaires $(X_i^t X_i)^{-1} X_i y_i$, $M_i = (X_i^t X)$ et $S_i = ||y_i - X_i - \hat{\beta}_i||^2$ $(i = 1, 2)$.

En général, une *loi* $m - n$ *poly-t* est définie comme le produit de m densités de Student, divisé par n densités du même type,

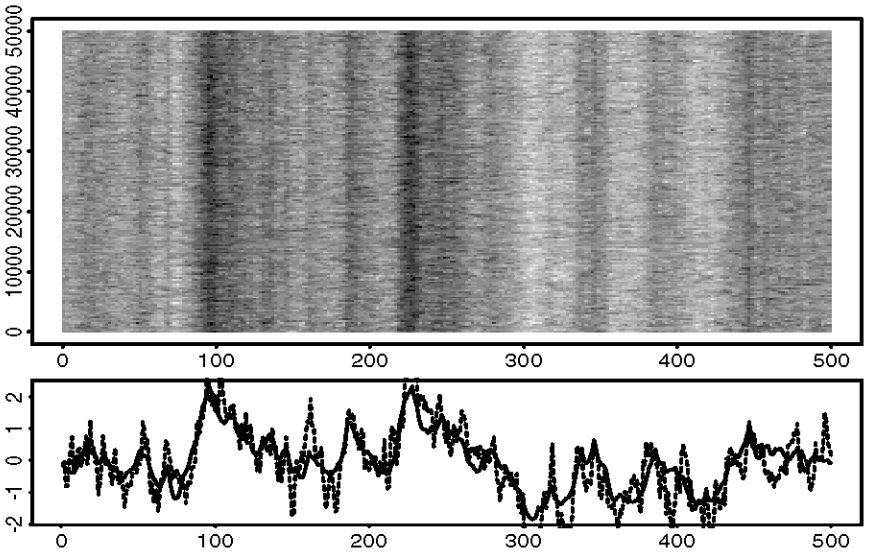

Fig. 4.5. Carte d'allocation (*haut*) et allocations moyennes par vraies volatilités (*bas*) pour le modèle (4.41). Les vraies volatilités sont représentées par des tirets. (*Source :* Mengersen *et al.*, 1999)

$$\varphi_{m,n}(x) \propto \prod_{j=1}^{n} \left[1 + (x-\mu_j^0)^t P_j^0 (x-\mu_j^0)\right]^{\nu_j^0/2} \Big/ \prod_{j=1}^{m} \left[1 + (x-\mu_j^1)^t P_j^1 (x-\mu_j^1)\right]^{\nu_j^1/2} .$$

Comme le montrent Bauwens et al. (1999, Théorème A.21), les densités $\varphi_{m,0}$ peuvent s'exprimer comme un mélange (continu) de densités régulières de Student par $(m-1)$ variables auxiliaires, une propriété qui peut s'utiliser soit pour une simulation directe, comme dans Bauwens (1984), soit pour une mise en œuvre MCMC (Chapitre 6), puisque le calcul direct de la constante de normalisation de $\varphi_{m,n}$, ou de l'espérance a posteriori correspondante, n'est pas possible. Une difficulté supplémentaire avec les lois a priori poly-t est que, relativement aux lois conjuguées, elles nécessitent la détermination d'un nombre beaucoup plus grand d'hyperparamètres.

5

Tests et régions de confiance

> "Twenty-six more tests were going to take the rest of daylight, maybe more. Heat or no heat, the days still grew shorter as if winter really was coming on, and a failed test would take a few minutes longer that one passed, just to make certain."
>
> **Robert Jordan**, *Lord of Chaos.*

5.1 Introduction

Bien que la théorie des tests puisse être envisagée comme cas particulier de la Théorie de la Décision pour un espace de décision restreint (et même comme un problème d'estimation), nous considérons l'inférence sur les tests dans un chapitre séparé, car il y a beaucoup plus d'ambiguïté dans la définition des buts inférentiels pour les tests que pour l'estimation d'une fonction régulière du paramètre. En effet, cette partie de l'inférence statistique bayésienne est encore incomplète, dans le sens où plusieurs autres réponses ont été avancées, mais aucune n'est entièrement satisfaisante. En particulier, il existe des différences notoires entre la théorie des tests fréquentistes et celle des tests bayésiens. De ce point de vue, le cadre des tests rend l'approche bayésienne plutôt attrayante, car la notion de probabilité d'une hypothèse, $\pi(\theta \in \Theta_0|x)$, ne peut être définie qu'à travers cette approche.

En réalité, certains bayésiens pensent que les tests ne devraient pas exister, ou, du moins, les tests d'une hypothèse nulle ponctuelle (voir, par exemple, Gelfand *et al.*, 1992) ; nous verrons dans ce chapitre plusieurs raisons philosophiques qui d'une manière ou d'une autre, plaident pour cette perspective

radicale. Ces raisons vont de l'aspect réducteur de la notion de modèle *(aucun modèle n'est correct, mais certains modèles sont moins faux* [ou plus utiles] *que d'autres)*, aux modifications artificielles de la loi a priori imposée par l'hypothèse nulle ponctuelle, au manque de structure décisionnelle du problème donné, à l'utilisation subséquente de fonctions de coût rudimentaires $0-1$ et de niveaux d'acceptation conventionnels, à l'impossibilité d'utiliser des lois a priori impropres dans les cas d'hypothèses ponctuelles et dans le cadre du choix de modèle (Chapitre 7). Mais les considérations pragmatiques sont telles que la boîte à outils bayésienne se doit d'inclure aussi des techniques de tests, ne serait-ce que parce que les utilisateurs de la Statistique ont été formés et habitués à traduire leurs problèmes en termes de tests, étant donné leur forte inclination à prendre cette formulation au pied de la lettre.

Nous considérerons d'abord dans la Section 5.2 l'approche bayésienne standard des tests, qui repose sur une évaluation des décisions par des coûts $0-1$ et comparerons les procédures bayésiennes avec leurs homologues fréquentistes dans la Section 5.3. Nous proposerons ensuite dans la Section 5.4 une alternative à l'approche décisionnelle fondée sur des coûts plus adaptés qui mettent en avant l'évaluation *ex post* pour des procédures de tests (par opposition aux procédures de Neyman-Pearson pour lesquelles l'évaluation fonctionne dans un esprit *ex ante*).

Ce chapitre exhibe un fort contraste entre les approches bayésienne et fréquentiste, et ce de diverses perspectives. Cette opposition est révélatrice du caractère incomplet de la modélisation classique, qui nécessite des concepts artificiels pour construire ses procédures optimales. Contrairement au cadre de l'estimation ponctuelle, ces procédures fréquentistes optimales ne sont plus des limites de procédures bayésiennes et elles en diffèrent numériquement. Cependant, nous modérons ce rejet dans la Section 5.3 en montrant que les procédures classiques et bayésiennes non informatives peuvent parfois mener à des conclusions similaires. Le Chapitre 7 traite du *choix de modèle*, qui peut être vu comme un cas particulier de tests d'hypothèses nulles ponctuelles, mais il présente assez de spécificités et de difficultés propres pour mériter un chapitre à lui seul (sans même prendre en compte le fait qu'il requière l'utilisation quasi systématique des méthodes numériques présentées dans le Chapitre 6).

5.2 Une première approche de la théorie des tests

5.2.1 Tests décisionnels

Soit un modèle statistique $f(x|\theta)$ avec $\theta \in \Theta$. Étant donné un sous-ensemble d'intérêt de Θ, Θ_0, qui se réduit parfois à un singleton $\{\theta_0\}$, la question posée est : *la vraie valeur du paramètre θ appartient-elle à Θ_0,* ce

qu'on appelle *tester* l'hypothèse[37]

$$H_0 : \theta \in \Theta_0,$$

souvent appelée *hypothèse nulle*. Pour les modèles linéaires, Θ_0 peut être un *sous-espace* de l'espace du vecteur Θ et le problème de test est alors un cas particulier du problème générique du *choix de modèle*, problème auquel le Chapitre 7 est consacré.

Exemple 5.1. Soit un modèle de *régression logistique*,

$$P_\alpha(y = 1) = 1 - P_\alpha(y = 0) = \exp(\alpha^t x)/(1 + \exp(\alpha^t x)), \qquad \alpha, x \in \mathbb{R}^p,$$

qui modélise la probabilité de développer un cancer de la prostate dans sa vie en fonction de variables explicatives $x = (x_1, \ldots, x_p)$. On s'intéresse particulièrement aux variables liées à l'environnement de travail comme la concentration d'amiante x_{i_0} ; un syndicat peut par exemple vouloir tester si le coefficient α_{i_0} correspondant à x_{i_0} est nul ou pas. ‖

Dans la perspective de Neyman-Pearson (Section 5.3), le problème de test est formalisé à l'aide d'un espace de décision $\mathscr{D}$ restreint à {oui, non} ou, d'une manière équivalente, à $\{1, 0\}$. En effet, il est logique de comprendre un problème de test comme une inférence sur la fonction indicatrice $\mathbb{I}_{\Theta_0}(\theta)$ et, par conséquent, de proposer des réponses dans $\mathbb{I}_{\Theta_0}(\Theta) = \{0, 1\}$. Bien entendu, la pertinence d'une telle restriction est moins évidente lorsque l'on considère que les tests apparaissent souvent comme composantes (ou comme étapes préliminaires) de structures inférentielles plus complexes et, en particulier, que la réponse à la question testée a aussi des conséquences en terme d'erreurs d'estimation (standard). Il serait alors plus intéressant de proposer des procédures prenant des valeurs dans $[0, 1]$. (Nous examinerons cette approche dans la Section 5.4.)

Dans certains cas, on dispose d'une information additionnelle sur le support de θ, à savoir que $\theta \in \Theta_0 \cup \Theta_1 \neq \Theta$. Dans ce cas, on définit l'hypothèse *alternative* contre laquelle nous testons H_0 comme

$$H_1 : \theta \in \Theta_1.$$

Dans cette formalisation, toute procédure de test φ apparaît comme un estimateur de $\mathbb{I}_{\Theta_0}(\theta)$ et nous n'avons besoin que d'une fonction de coût $\mathrm{L}(\theta, \varphi)$ pour construire des estimateurs de Bayes. Par exemple, la fonction de coût proposée par Neyman et Pearson est le coût $0 - 1$

$$\mathrm{L}(\theta, \varphi) = \begin{cases} 1 & \text{si } \varphi \neq \mathbb{I}_{\Theta_0}(\theta), \\ 0 & \text{sinon,} \end{cases}$$

[37] Il y a une certaine ambiguïté dans la terminologie : le mot *test* couvre simultanément la question et la procédure utilisée pour répondre à la question.

présentée dans le Chapitre 2. Pour ce coût, la solution bayésienne est

$$\varphi^\pi(x) = \begin{cases} 1 & \text{si } P^\pi(\theta \in \Theta_0|x) > P^\pi(\theta \in \Theta_0^c|x), \\ 0 & \text{sinon.} \end{cases}$$

Cet estimateur se justifie aisément en termes intuitifs, car il choisit l'hypothèse avec la probabilité a posteriori la plus grande. Une généralisation du coût ci-dessus est de pénaliser différemment les erreurs suivant que l'hypothèse nulle est vraie ou fausse. Les coûts pondérés $0-1$

$$\mathrm{L}(\theta, \varphi) = \begin{cases} 0 & \text{si } \varphi = \mathbb{I}_{\Theta_0}(\theta), \\ a_0 & \text{si } \theta \in \Theta_0 \text{ et } \varphi = 0, \\ a_1 & \text{si } \theta \notin \Theta_0 \text{ et } \varphi = 1, \end{cases} \tag{5.1}$$

sont appelés "$a_0 - a_1$" pour des raisons évidentes. L'estimateur de Bayes associé est alors donné par le résultat suivant.

Proposition 5.2. *Sous le coût* (5.1)*, l'estimateur de Bayes associé à la loi a priori* π *est*

$$\varphi^\pi(x) = \begin{cases} 1 & \textit{si } P^\pi(\theta \in \Theta_0|x) > \dfrac{a_1}{a_0 + a_1}, \\ 0 & \textit{sinon.} \end{cases}$$

Preuve. Puisque le coût a posteriori est

$$\begin{aligned} \mathrm{L}(\pi, \varphi|x) &= \int_\Theta \mathrm{L}(\theta, \varphi)\pi(\theta|x)d\theta \\ &= a_0 P^\pi(\theta \in \Theta_0|x)\mathbb{I}_{\{0\}}(\varphi) + a_1 P^\pi(\theta \notin \Theta_0|x)\mathbb{I}_{\{1\}}(\varphi), \end{aligned}$$

l'estimateur de Bayes peut être calculé directement. □

Pour ce type de coût, l'hypothèse nulle H_0 est rejetée quand la probabilité a posteriori de H_0 est trop petite, le *niveau d'acceptation* $a_1/(a_0 + a_1)$ étant déterminé par le choix de la fonction de perte. Notons que φ^π ne dépend que de a_0/a_1 et que, plus a_0/a_1 est grand, c'est-à-dire plus une réponse incorrecte est pénalisée sous H_0 relativement à H_1, plus la probabilité a posteriori de H_0 doit être petite pour être rejetée.

Exemple 5.3. Soient $x \sim \mathscr{B}(n, p)$ et $\Theta_0 = [0, 1/2]$. Pour la loi a priori uniforme $\pi(p) = 1$, la probabilité a posteriori de H_0 est

$$\begin{aligned} P^\pi(p \le 1/2|x) &= \frac{\int_0^{1/2} p^x(1-p)^{n-x}dp}{B(x+1, n-x+1)} \\ &= \frac{(1/2)^{n+1}}{B(x+1, n-x+1)} \left\{ \frac{1}{x+1} + \frac{n-x}{(x+1)(x+2)} + \ldots + \frac{(n-x)!x!}{(n+1)!} \right\} \end{aligned}$$

qui peut se calculer facilement et être comparée au niveau d'acceptation. ‖

Exemple 5.4. Soient $x \sim \mathcal{N}(\theta, \sigma^2)$ et $\theta \sim \mathcal{N}(\mu, \tau^2)$. Alors $\pi(\theta|x)$ est la loi normale $\mathcal{N}(\mu(x), \omega^2)$ avec

$$\mu(x) = \frac{\sigma^2\mu + \tau^2 x}{\sigma^2 + \tau^2} \quad \text{et} \quad \omega^2 = \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}.$$

Pour tester $H_0 : \theta < 0$, nous calculons

$$\begin{aligned} P^\pi(\theta < 0|x) &= P^\pi\left(\frac{\theta - \mu(x)}{\omega} < \frac{-\mu(x)}{\omega}\right) \\ &= \Phi\left(-\mu(x)/\omega\right). \end{aligned}$$

Si z_{a_0,a_1} est le quantile $a_1/(a_0+a_1)$, donc s'il satisfait $\Phi(z_{a_0,a_1}) = a_1/(a_0+a_1)$, H_0 est acceptée lorsque

$$-\mu(x) > z_{a_0,a_1}\omega,$$

la borne supérieure d'acceptation étant alors

$$-\frac{\sigma^2}{\tau^2}\mu - (1 + \frac{\sigma^2}{\tau^2})\omega z_{a_0,a_1}.$$

‖

Notons de nouveau que, d'un point de vue bayésien, il semble naturel de fonder la décision sur la probabilité a posteriori que l'hypothèse soit vraie. Dans la Section 5.4, nous montrons qu'une approche décisionnelle alternative mène à cette probabilité a posteriori en tant qu'estimateur de Bayes et évite ainsi la comparaison à un niveau d'acceptation prédéterminé. En fait, une difficulté liée aux coûts (5.1) est le choix des poids a_0 et a_1, car ils sont choisis le plus souvent de manière automatique plutôt que déterminés par des considérations d'utilité.

5.2.2 Le facteur de Bayes

Bien que, d'un point de vue décisionnel, le *facteur de Bayes* ne soit qu'une transformation bijective de la probabilité a posteriori, il a fini par être considéré comme réponse en soi en théorie des tests bayésiens, sous l'impulsion de Jeffreys (1939).

Définition 5.5. *Le* facteur de Bayes *est le rapport des probabilités a posteriori des hypothèses nulle et alternative sur le rapport des probabilités a priori de ces mêmes hypothèses, soit*

$$B_{01}^\pi(x) = \frac{P(\theta \in \Theta_0 \mid x)}{P(\theta \in \Theta_1 \mid x)} \Big/ \frac{\pi(\theta \in \Theta_0)}{\pi(\theta \in \Theta_1)}.$$

Ce rapport évalue la modification de la vraisemblance de l'ensemble Θ_0 par rapport à celle de l'ensemble Θ_1 due à l'observation et peut se comparer naturellement à 1, bien qu'une échelle de comparaison exacte doive être fondée sur une fonction de coût. Dans le cas particulier où $\Theta_0 = \{\theta_0\}$ et $\Theta_1 = \{\theta_1\}$, le facteur de Bayes se simplifie et devient le *rapport de vraisemblance* classique

$$B_{01}^{\pi}(x) = \frac{f(x|\theta_0)}{f(x|\theta_1)}.$$

En général, le facteur de Bayes dépend de l'information a priori, mais il est souvent proposé comme réponse bayésienne "objective", car il élimine partiellement l'influence du modèle a priori et souligne le rôle des observations. De fait, il peut être perçu comme un rapport de vraisemblance bayésien, car, si π_0 est la loi a priori sous H_0 et π_1, la loi a priori sous H_1, $B_{01}^{\pi}(x)$ peut s'écrire

$$B_{01}^{\pi}(x) = \frac{\int_{\Theta_0} f(x|\theta_0)\pi_0(\theta)\,d\theta}{\int_{\Theta_1} f(x|\theta_1)\pi_1(\theta)\,d\theta} = \frac{m_0(x)}{m_1(x)}\,, \tag{5.2}$$

ce qui revient donc à remplacer les vraisemblances par des marginales sous les deux hypothèses.

Comme nous l'avons indiqué ci-dessus, le facteur de Bayes est, d'un point de vue décisionnel, complètement équivalent à la probabilité a posteriori de l'hypothèse nulle puisque, sous (5.1), H_0 est accepté lorsque

$$B_{01}^{\pi}(x) > \frac{a_1}{a_0} \Big/ \frac{\varrho_0}{\varrho_1} = \frac{a_1\varrho_1}{a_0\varrho_0}, \tag{5.3}$$

où

$$\varrho_0 = \pi(\theta \in \Theta_0) \quad \text{et} \quad \varrho_1 = \pi(\theta \in \Theta_1) = 1 - \varrho_0. \tag{5.4}$$

Cette version alternative de la Proposition 5.2 fournit ainsi une illustration de la dualité qui existe entre coûts et lois a priori, dualité déjà mentionnée au Chapitre 2. En effet, (5.3) montre qu'il est équivalent de pondérer de la même façon les deux hypothèses, $\varrho_0 = \varrho_1 = 1/2$, et de modifier les pénalités d'erreur dans $a_i' = a_i\varrho_i$ $(i = 0, 1)$ ou de pénaliser de la même façon les deux types d'erreurs $(a_1 = a_0 = 1)$, lorsque la loi a priori intègre les poids réels dans les probabilités a priori pondérées,

$$\varrho_0' = \frac{a_0\varrho_0}{a_0\varrho_0 + a_1\varrho_1}, \quad \varrho_1' = \frac{a_1\varrho_1}{a_0\varrho_0 + a_1\varrho_1}.$$

À la suite de Jeffreys (1939) et de Good (1952), le facteur de Bayes est désormais un outil à part entière (voir, par exemple, Kass et Raftery, 1995, pour une revue détaillée). En particulier, Jeffreys (1939) a développé une échelle "absolue" pour évaluer le degré de certitude en faveur ou au détriment de H_0 apporté par les données, *en l'absence d'un cadre décisionnel véritable*. L'échelle de Jeffreys est la suivante :

(i) si $\log_{10}(B_{10}^{\pi})$ varie entre 0 et 0.5, la certitude que H_0 est fausse est *faible*,

(ii) si elle est entre 0.5 et 1, cette certitude est *substantielle*,

(iii) si elle est entre 1 et 2, elle est *forte* et

(iv) si elle est au-dessus de 2, elle est *décisive*,

avec la même échelle en faveur de H_0 pour les valeurs négatives. Bien entendu, cette graduation du facteur de Bayes donne quelques indications sur le degré de certitude, mais les limites précises séparant une catégorie d'une autre sont conventionnelles et peuvent être changées de façon arbitraire, comme l'ont illustré Kass et Raftery (1995). C'est une conséquence du manque de justification décisionnelle de cette méthode et de l'absence de fonction de coût. (La critique s'applique également aux niveaux α conventionnels de 0.05 ou 0.01 utilisés pour $a_0/(a_0+a_1)$ dans (5.1).)

Le Chapitre 6 donnera des précisions sur les méthodes utilisées pour approcher les facteurs de Bayes lorsque l'intégrale dans (5.2) ne peut pas se calculer analytiquement, ce qui est souvent le cas.

Exemple 5.6. (Kass et Raftery, 1995) La "hot hand" en basket ball est une croyance répandue que les joueurs ont des bons et des mauvais jours, plutôt qu'une probabilité constante de réussir un tir. Pour un joueur donné, le modèle sous l'hypothèse nulle *(pas de hot hand)* est alors $H_0 : y_i \sim \mathscr{B}(n_i, p)$ $(i = 1, \ldots, G)$, où G est le nombre de parties et n_i (resp. y_i) le nombre de tirs (resp. de bons tirs) pendant la i-ième partie. Le modèle sous l'alternative générale est $H_1 : y_i \sim \mathscr{B}(n_i, p_i)$, la probabilité p_i variant de partie en partie. Sous une loi a priori conjuguée $p_i \sim \mathscr{B}e(\xi/\omega, (1-\xi)/\omega)$, la moyenne $\mathbb{E}[p_i|\xi,\omega] = \xi$ est distribuée selon une loi a priori uniforme $\mathscr{U}([0,1])$, comme l'est p sous H_0, et ω est fixé. Le facteur de Bayes est alors

$$\begin{aligned} B_{10} &= \int_0^1 \prod_{i=1}^{G} \int_0^1 p_i^{y_i}(1-p_i)^{n_i-y_i} p_i^{\alpha-1}(1-p_i)^{\beta-1} d\,p_i \\ &\qquad \times \frac{\{\Gamma(1/\omega)/[\Gamma(\xi/\omega)\Gamma((1-\xi)/\omega)]\}^G}{\int_0^1 p^{\sum_i y_i}(1-p)^{\sum_i (n_i-y_i)} d\,p}\, d\xi \\ &= \int_0^1 \prod_{i=1}^{G} [\Gamma(y_i+\xi/\omega)\Gamma(n_i-y_i+(1-\xi)/\omega)/\Gamma(n_i+1/\omega)] \\ &\qquad \times \frac{\{\Gamma(1/\omega)/[\Gamma(\xi/\omega)\Gamma((1-\xi)/\omega)]\}^G}{\Gamma(\sum_i y_i+1)\Gamma(\sum_i (n_i-y_i)+1)/\Gamma(\sum_i n_i+2)}\, d\xi\,, \end{aligned}$$

où $\alpha = \xi/\omega$ et $\beta = (1-\xi)/\omega$. Formellement, le numérateur peut se calculer exactement, malgré les fonctions gamma, grâce à la simplification

$$\Gamma(y_i+\xi/\omega)/\Gamma(\xi/\omega) = \prod_{j=1}^{y_i}(j-1+\xi/\omega),$$

$$\Gamma(n_i-y_i+(1-\xi)/\omega)/\Gamma((1-\xi)/\omega)] = \prod_{j=1}^{n-i-y_i}(j-1+(1-\xi)/\omega)\,,$$

mais la fonction de ξ à intégrer est alors un polynôme de degré élevé. La résolution de l'intégrale nécessite par conséquent un logiciel de calcul formel comme Maple ou Mathematica. Pour un joueur donné, la valeur de B_{10} est 0.16 pour $\omega = 0.005$ et $G = 138$, ce qui n'indique aucune preuve décisive en faveur de l'hypothèse de la *hot hand*. ‖

5.2.3 Modification de la loi a priori

La notion de facteur de Bayes permet aussi de mettre en évidence un aspect important des tests bayésiens. En fait, ce facteur n'est défini que lorsque $\varrho_0 \neq 0$ et $\varrho_1 \neq 0$. Cela implique que, si H_0 ou H_1 sont a priori impossibles, les observations ne vont pas modifier cette information absolue : des probabilités nulles a priori le restent a posteriori ! Par conséquent, une hypothèse nulle ponctuelle $H_0 : \theta = \theta_0$ ne peut pas être testée sous une loi a priori *continue*. Plus généralement, la sélection de variables (Chapitre 7) est incompatible avec des lois a priori absolument continues par rapport à la mesure de Lebesgue définies sur l'espace le plus grand.

Le test d'une hypothèse nulle ponctuelle (ou à probabilité nulle par rapport à la mesure dominante) impose par conséquent une modification radicale de la loi a priori, car il exige de construire une loi a priori pour les deux sous-ensembles Θ_0 et Θ_1, par exemple, des lois π_0 et π_1 de densités

$$g_0(\theta) \propto \pi(\theta)\mathbb{I}_{\Theta_0}(\theta), \quad g_1(\theta) \propto \pi(\theta)\mathbb{I}_{\Theta_1}(\theta),$$

(relativement aux mesures naturelles sur Θ_0 et Θ_1) bien que cette définition ne soit pas toujours dénuée d'ambiguïté (voir l'Exercice 5.5). Combinées aux probabilités a priori ϱ_0 et ϱ_1 de Θ_0 et Θ_1 données par (5.4), π_0 et π_1 définissent la loi a priori π. En d'autres termes,

$$\pi(\theta) = \varrho_0\pi_0(\theta) + \varrho_1\pi_1(\theta).$$

(Lorsque $\Theta_0 = \{\theta_0\}$, la loi a priori sur Θ_0 est juste la masse de Dirac en θ_0.)

D'un point de vue décisionnel, cette modification de la loi a priori est surprenante, puisqu'elle revient à mettre un poids a priori sur un ensemble de mesure 0. Elle souligne aussi la dichotomie imposée par l'approche habituelle des tests pour laquelle l'hypothèse nulle est soit vraie, soit fausse. Cependant, à moins que le décideur ne soit inflexible sur le choix de la loi a priori π et, dans ce cas, H_0 devrait vraiment être refusé si π ne donne aucun poids à Θ_0,

on peut considérer le problème de test comme fournissant une information supplémentaire sur θ (même si celle-ci est vague). Effectivement, tester $\theta \in \Theta_0$ signifie qu'il y a une certaine chance que θ appartienne vraiment à Θ_0 (sinon, on ne se poserait pas la question !) et par conséquent qu'une certaine information, peut-être mal définie, a été fournie sur ce fait.

Considérer les cadres de test comme sources d'information est plus convaincant encore si la décision finale n'est pas la réponse au test mais l'estimation d'une fonction de θ, c'est-à-dire lorsque le test signifie le choix d'un sous-modèle. Un test préliminaire sur l'information vague peut alors améliorer l'étape d'estimation. De plus, en gardant cette perspective du choix de modèle comme objectif réel de l'analyse, il est aussi logique de développer une loi a priori séparée pour chaque sous-espace, puisqu'un seul des deux Θ_i sera pris en compte après l'étape de test. Par exemple, pour une *hypothèse nulle ponctuelle* donnée, $H_0 : \theta = \theta_0$, la loi non informative $\pi(\theta) = 1$ ne peut pas être considérée comme une loi a priori sur Θ acceptable, car la valeur particulière θ_0 a été choisie comme une valeur possible pour θ. (Dans le Chapitre 7, nous défendrons davantage la perspective que des paramètres similaires apparaissant dans deux modèles différents doivent être considérés comme des entités séparées.) En général, considérer que les problèmes de test se produisent à cause d'observations additionnelles (indisponibles) peut aider à la construction de la loi a priori non informative, même s'il n'y a pas de consensus sur une modélisation a priori non informative des tests (voir la Section 5.3.5).

5.2.4 Hypothèses nulles ponctuelles

Une critique usuelle des hypothèses nulles ponctuelles est qu'elles ne sont pas *réalistes* (voir, pour illustration, Casella et Berger, 1987)[38]. Par exemple, comme l'a souligné Good (1980), il n'y a pas de sens à tester que la probabilité qu'il pleuve demain est de[39] $0.7163891256\ldots$ Cependant, certains problèmes statistiques nécessitent vraiment un test d'hypothèse nulle ponctuelle. Par exemple, pour l'estimation de mélanges (voir la Section 1.1 et la Section 6.4), il peut être important de savoir si une loi de mélange possède deux ou trois composantes et il est donc nécessaire de tester si le poids d'une de ces composantes est nul. De la même façon, dans le domaine de la régression linéaire, des tests de nullité des coefficients de la régression permettent l'élimination des variables exogènes inutiles, comme dans l'Exemple 5.1. D'une façon plus pertinente encore, tester si l'univers est en expansion, s'il se contracte ou s'il est stable revient à tester si la constante de Hubble est plus grande, plus petite ou égale à une valeur spécifique $\mathfrak{h}_0$.

[38] Roger Berger et non pas James Berger !

[39] En revanche, il y a un sens à tester si la prévision de 75% donnée par le météorologiste local est exacte, c'est-à-dire si la probabilité de pluie pour un jour donné est 0.75 ou une autre des probabilités annoncées par le météorologiste (voir l'Exemple 2.12).

Plus généralement, des hypothèses bilatérales telles que $H_0 : \theta \in \Theta_0 = [\theta_0 - \epsilon, \theta_0 + \epsilon]$ peuvent être approchées par $H_0 : \theta = \theta_0$, ce qui entraîne une modification des probabilités a posteriori, qui sont presque nulles lorsque ϵ est suffisamment petit. C'est le cas notamment lorsque la vraisemblance est constante autour de θ_0 (voir Berger, 1985b, et Berger et Delampady, 1987). Les hypothèses nulles ponctuelles ont aussi une grande importance pratique ; par exemple, bien qu'il y ait un sens à déterminer si un traitement médical a un effet positif ou négatif, la première question est de décider s'il a un quelconque effet.

Soit l'hypothèse nulle ponctuelle $H_0 : \theta = \theta_0$; notons ϱ_0 la probabilité a priori que $\theta = \theta_0$ et g_1 la densité a priori sous l'alternative. La loi a priori est alors $\pi_0(\theta) = \varrho_0 \mathbb{I}_{\Theta_0}(\theta) + (1 - \varrho_0) g_1(\theta)$ et la probabilité a posteriori de H_0 est donnée par

$$\pi(\Theta_0|x) = \frac{f(x|\theta_0)\varrho_0}{\int f(x|\theta)\pi(\theta)\, d\theta} = \frac{f(x|\theta_0)\varrho_0}{f(x|\theta_0)\varrho_0 + (1 - \varrho_0) m_1(x)},$$

la loi marginale sous H_1 étant

$$m_1(x) = \int_{\Theta_1} f(x|\theta) g_1(\theta)\, d\theta.$$

Cette probabilité a posteriori peut aussi s'écrire

$$\pi(\Theta_0|x) = \left[1 + \frac{1 - \varrho_0}{\varrho_0} \frac{m_1(x)}{f(x|\theta_0)}\right]^{-1}.$$

De la même façon, le facteur de Bayes est

$$B_{01}^{\pi}(x) = \frac{f(x|\theta_0)\varrho_0}{m_1(x)(1 - \varrho_0)} \Big/ \frac{\varrho_0}{1 - \varrho_0} = \frac{f(x|\theta_0)}{m_1(x)}$$

et nous obtenons la relation générale suivante entre les deux quantités :

$$\pi(\Theta_0|x) = \left[1 + \frac{1 - \varrho_0}{\varrho_0} \frac{1}{B_{01}^{\pi}(x)}\right]^{-1}.$$

Exemple 5.7. (Suite de l'Exemple 5.3) Soit le test de $H_0 : p = 1/2$ contre $p \neq 1/2$. Pour $g_1(p) = 1$, la probabilité a posteriori est alors donnée par

$$\begin{aligned} \pi(\Theta_0|x) &= \left[1 + \frac{1 - \varrho_0}{\varrho_0} 2^n B(x + 1, n - x + 1)\right]^{-1} \\ &= \left[1 + \frac{1 - \varrho_0}{\varrho_0} \frac{x!(n - x)!}{(n - 1)!} 2^n\right]^{-1}, \end{aligned}$$

puisque $m(x) = \binom{n}{x} B(x + 1, n - x + 1)$. Par exemple, si $n = 5$, $x = 3$ et $\varrho_0 = 1/2$, la probabilité a posteriori est

$$\left(1+\frac{2}{120}2^5\right)^{-1}=\frac{15}{23}$$

et le facteur de Bayes correspondant est 15/8, proche de 2. Donc, dans la plupart des cas les plus favorables, les probabilités a posteriori tendent à favoriser H_0. Lorsque la taille d'échantillon augmente, les variations des réponses possibles s'élargissent aussi. Par exemple, si $\pi(p)$ est $\mathscr{B}e(1/2,1/2)$ et $n=10$, les probabilités a posteriori sont données dans le Tableau 5.1 et soutiennent H_0 pour x proche de 5, même si la loi a priori est plutôt biaisée contre l'hypothèse nulle (car les valeurs extrêmes, 0 et 1, ont un poids important). ‖

Tab. 5.1. Probabilités a posteriori de $p=1/2$ lorsque $x\sim\mathscr{B}(10,p)$.

x	0	1	2	3	4	5
$P(p=1/2\|x)$	0.0055	0.0953	0.3737	0.6416	0.7688	0.8025

Exemple 5.8. (Suite de l'Exemple 5.4) Soit le test de $H_0:\theta=0$. Il semble raisonnable de prendre π_1 égal à $\mathscr{N}(\mu,\tau^2)$ et $\mu=0$, si aucune information additionnelle n'est disponible. Alors

$$\begin{aligned}\frac{m_1(x)}{f(x|0)}&=\frac{\sigma}{\sqrt{\sigma^2+\tau^2}}\frac{e^{-x^2/2(\sigma^2+\tau^2)}}{e^{-x^2/2\sigma^2}}\\&=\sqrt{\frac{\sigma^2}{\sigma^2+\tau^2}}\exp\left\{\frac{\tau^2x^2}{2\sigma^2(\sigma^2+\tau^2)}\right\},\end{aligned}$$

et la probabilité a posteriori se calcule comme suit :

$$\pi(\theta=0|x)=\left[1+\frac{1-\varrho_0}{\varrho_0}\sqrt{\frac{\sigma^2}{\sigma^2+\tau^2}}\exp\left(\frac{\tau^2x^2}{2\sigma^2(\sigma^2+\tau^2)}\right)\right]^{-1}.$$

Dans le cas particulier où $\varrho_0=1/2$ et $\tau=\sigma$, le Tableau 5.2 donne les probabilités a posteriori en fonction de $z=x/\sigma$. ‖

Tab. 5.2. Probabilités a posteriori de $\theta=0$ lorsque $x\sim\mathscr{N}(\theta,\sigma^2)$ pour différentes valeurs de $z=x/\sigma$ et pour $\tau=\sigma$.

z	0	0.68	1.28	1.96
$\pi(\theta=0\|z)$	0.586	0.557	0.484	0.351

Considérons maintenant l'alternative $\tau^2=10\sigma^2$, supposée indiquer une information a priori plus diffuse sur θ. Les probabilités a posteriori de H_0 sont alors modifiées comme le montre le Tableau 5.3. De manière surprenante, elles sont toutes plus favorables à H_0 : ce phénomène est lié au *paradoxe de Jeffreys-Lindley*, décrit dans la section suivante.

Tab. 5.3. Probabilités a posteriori de $\theta = 0$ lorsque $x \sim \mathcal{N}(\theta, \sigma^2)$ pour $\tau^2 = 10\sigma^2$ et $z = x/\sigma$.

z	0	0.68	1.28	1.96
$\pi(\theta = 0\|x)$	0.768	0.729	0.612	0.366

5.2.5 Lois a priori impropres

Le recours à des lois a priori non informatives pour tester des hypothèses est plutôt délicat, et DeGroot (1973) affirme que les lois a priori impropres ne devraient pas *du tout* être utilisées pour les tests. En effet, comme nous l'avons remarqué auparavant, le cadre formel des tests n'est pas cohérent avec un manque absolu d'information, car effectuer un test implique au moins une division de l'espace des paramètres en deux sous-ensembles, dont l'un peut être de mesure nulle sous une loi impropre comme la loi de Jeffreys. Cependant, l'inconvénient d'utiliser des lois a priori impropres va plus loin, car ces dernières sont incompatibles avec la plupart des tests d'hypothèses nulles ponctuelles.

Nous illustrons cette difficulté dans un cadre gaussien, $x \sim \mathcal{N}(\theta, 1)$, sous l'hypothèse nulle ponctuelle $H_0 : \theta = 0$ testée contre $H_1 : \theta \neq 0$. Si nous utilisons la loi a priori impropre $\pi(\theta) = 1$ pour $\theta \neq 0$, donc si π est la loi de densité

$$\pi(\theta) = \frac{1}{2}\mathbb{I}_0(\theta) + \frac{1}{2} \cdot 1,$$

la probabilité a posteriori de H_0 est

$$\pi(\theta = 0|x) = \frac{e^{-x^2/2}}{e^{-x^2/2} + \int_{-\infty}^{+\infty} e^{-(x-\theta)^2/2}\, d\theta} = \frac{1}{1 + \sqrt{2\pi} e^{x^2/2}}.$$

(Le choix particulier de la constante 1 dans la loi a priori est crucial pour la discussion suivante, bien qu'il soit arbitraire.) Cette probabilité a posteriori de H_0 est donc bornée supérieurement par $1/(1+\sqrt{2\pi}) = 0.285$. Ceci implique que la loi a posteriori est plutôt biaisée contre H_0, même dans le cas le plus favorable. À moins que l'échelle de comparaison, c'est-à-dire le coût, ne soit modifiée pour estimer ces valeurs faibles, l'hypothèse nulle ponctuelle sera donc assez souvent rejetée. Un phénomène similaire se produit lorsque Θ_0 est compact. Par exemple, le test de $H_0 : |\theta| \leq 1$ contre $H_1 : |\theta| > 1$ mène à la probabilité a posteriori suivante :

$$\begin{aligned}\pi(|\theta| \leq 1|x) &= \frac{\int_{-1}^{1} e^{-(x-\theta)^2/2}\, d\theta}{\int_{-\infty}^{+\infty} e^{-(x-\theta)^2/2}\, d\theta} \\ &= \Phi(1-x) - \Phi(-1-x) \\ &= \Phi(x+1) - \Phi(x-1),\end{aligned}$$

Tab. 5.4. Probabilités a posteriori de $|\theta| < 1$ pour $x \sim \mathcal{N}(\theta, 1)$.

x	0.0	0.5	1.0	1.5	2.0
$\pi(\|\theta\| \leq 1\|x)$	0.683	0.625	0.477	0.302	0.157

Tab. 5.5. Probabilités a posteriori de $\theta = 0$ pour la loi a priori de Jeffreys $\pi(\theta) = 1$ et $x \sim \mathcal{N}(\theta, 1)$.

x	0.0	1.0	1.65	1.96	2.58
$\pi(\theta = 0\|x)$	0.285	0.195	0.089	0.055	0.014

dont les valeurs numériques sont données dans le Tableau 5.4. Par conséquent, le support maximal de l'hypothèse H_0, égal à 0.683, reste modéré.

Une caractéristique intéressante de la loi a priori de Lebesgue peut être exhibée par l'hypothèse nulle ponctuelle $H_0 : \theta = 0$. La procédure résultante est en accord avec la réponse classique correspondante, comme le montre le Tableau 5.5. La probabilité a posteriori $\pi(\theta = 0|x)$ est effectivement assez proche des niveaux d'importance classiques 0.10, 0.05 et 0.01 lorsque x est 1.65, 1.96, ou 2.58 (on démontrera dans la Note 5.7.1 que cette comparaison a un sens). Cette coïncidence n'est pas vérifiée par toutes les valeurs de x mais montre que, pour les niveaux de signification habituels (et pour des objectifs de test), la réponse classique peut être considérée comme une réponse bayésienne non informative, même si elle correspond à une loi a priori difficilement justifiable.

Une autre illustration de la délicate question des lois a priori impropres dans des cadres de test est fournie par le *paradoxe de Jeffreys-Lindley*. En effet, les arguments limites ne sont pas valables pour les tests et empêchent une construction alternative des réponses non informatives. Par exemple, considérant la loi a priori conjuguée présentée dans l'Exemple 5.4, la probabilité a posteriori est

$$\pi(\theta = 0|x) = \left\{ 1 + \frac{1-\varrho_0}{\varrho_0} \sqrt{\frac{\sigma^2}{\sigma^2+\tau^2}} \exp\left[\frac{\tau^2 x^2}{2\sigma^2(\sigma^2+\tau^2)} \right] \right\}^{-1},$$

qui converge vers 1 lorsque la variance a priori τ tend vers $+\infty$, pour tout $x \neq 0$. Cette limite diffère de la réponse "non informative" construite précédemment $[1 + \sqrt{2\pi}\exp(x^2/2)]^{-1}$ et est évidemment complètement inutile. Ce phénomène peut aussi s'observer en comparant les Tableaux 5.2 et 5.3, car la probabilité est plus grande lorsque $\tau^2 = 10\sigma^2$ que lorsque $\tau = \sigma$ pour toutes les valeurs de z considérées dans les tableaux. Voir Aitkin (1991) et Robert (1993a) pour des discussions sur ce paradoxe.

Les paradoxes associés aux lois a priori impropres comme l'exemple de Jeffreys-Lindley sont en réalité dus à une indétermination des poids a priori qui n'apparaît pas dans les problèmes d'estimation ponctuelle, ni dans les tests unilatéraux.

Exemple 5.9. Soient $x \sim \mathcal{N}(\theta, 1)$ et $H_0 : \theta \leq 0$ à tester contre $H_1 : \theta > 0$. Pour l'a priori diffus $\pi(\theta) = 1$,

$$\pi(\theta \leq 0|x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{-(x-\theta)^2/2}\, d\theta = \Phi(-x)\,.$$

Dans ce cas, la réponse bayésienne généralisée est aussi une procédure classique, appelée *p-value* (voir la Section 5.3.4). ‖

Pour des problèmes bilatéraux, si g_0 et g_1 sont des mesures σ-finies correspondant à des lois a priori non informatives tronquées aux sous-espaces Θ_0 et Θ_1, le choix des constantes de normalisation influera sur l'estimateur de Bayes. En effet, si g_i est remplacé par $c_i g_i$ $(i = 0, 1)$, le facteur de Bayes est multiplié par c_0/c_1. Par exemple, si la loi a priori de Jeffreys est uniforme et si $g_0 = c_0$, $g_1 = c_1$, la probabilité a posteriori est

$$\begin{aligned}\pi(\theta \in \Theta_0|x) &= \frac{\varrho_0 c_0 \int_{\Theta_0} f(x|\theta)\, d\theta}{\varrho_0 c_0 \int_{\Theta_0} f(x|\theta)\, d\theta + (1-\varrho_0) c_1 \int_{\Theta_1} f(x|\theta)\, d\theta} \\ &= \frac{\varrho_0 \int_{\Theta_0} f(x|\theta)\, d\theta}{\varrho_0 \int_{\Theta_0} f(x|\theta)\, d\theta + (1-\varrho_0)[c_1/c_0] \int_{\Theta_1} f(x|\theta)\, d\theta}\,,\end{aligned}$$

qui dépend du rapport c_1/c_0. Par exemple, l'équivalent du Tableau 5.5 pour $\pi(\theta) = 10$ est donné dans le Tableau 5.6, avec des différences importantes pour la plupart des valeurs de x, car elles diffèrent d'une magnitude.

Tab. 5.6. Probabilités a posteriori de $\theta = 0$ pour la loi a priori de Jeffreys $\pi(\theta) = 10$.

x	0.0	1.0	1.65	1.96	2.58
$\pi(\theta = 0\|x)$	0.0384	0.0236	0.0101	0.00581	0.00143

Il est donc nécessaire d'élargir la perspective non informative de ces cadres de test en développant une technique capable de construire les poids c_i d'une façon non informative et acceptable. Bernardo (1980), Spiegelhalter et Smith (1980), Smith et Spiegelhalter (1982), Aitkin (1991), Pettit (1992), Robert (1993a) et Berger et Pericchi (1996b,a) ont fait des propositions dans ce sens, comme le détaille la Section 5.2.6. Notons que Jeffreys (1961) proposait au contraire d'utiliser des lois a priori propres dans ces cas, comme les lois $\mathcal{C}(0, \sigma^2)$ ou $\mathcal{N}(0, 10\sigma^2)$ quand $x \sim \mathcal{N}(\theta, \sigma^2)$ et $H_0 : \theta = 0$. Le problème est alors que le choix d'une loi a priori propre influera sur la réponse du test.

Avant d'introduire dans la Section 5.2.6 certains des développements récents liés à l'utilisation des lois a priori impropres, faisons la remarque suivante : utiliser des lois a priori impropres, comme celle de Jeffreys, pour des tests bilatéraux reste non satisfaisant, car elles semblent conduire à trop d'arbitraire au sens où de nombreuses solutions contradictoires abondent, reposant

sur des principes théoriques similaires mais produisant des valeurs numériques différentes, ce qui contredit le principe de vraisemblance. En d'autres termes, bien que les solutions proposées dans la section suivante soient intéressantes et convaincantes, en tant que principes constructifs, les difficultés relatives à l'utilisation de lois a priori impropres dans les tests font que celles-ci ne relèvent pas à proprement parler du paradigme bayésien. Nous considérons dans la Section 5.3 une approche alternative qui définit une réponse bayésienne la moins favorable comme une limite inférieure d'estimateurs (propres) de Bayes (mais qui présente également d'importants défauts).

Les difficultés rencontrées avec les lois a priori non informatives montrent aussi que le problème des tests ne peut pas être traité de façon cohérente s'il n'y a pas d'information a priori disponible ; en d'autres termes, l'information apportée par les observations seules n'est souvent pas suffisante pour déterminer catégoriquement si l'hypothèse est *vraie ou fausse*. Évidemment, cela renforce la motivation d'un traitement bayésien de tels problèmes, car c'est la seule approche cohérente qui profite de l'information résiduelle.

5.2.6 Pseudo-facteurs de Bayes

La plupart[40] des solutions proposées pour surmonter les difficultés liées à l'emploi de lois a priori impropres reposent sur l'utilisation d'une partie des données, afin de transformer les lois impropres en lois propres, ou le recours à des observations imaginaires pour obtenir le même résultat.

Définition 5.10. *Pour une loi a priori impropre π donnée, un échantillon $(x_1, \ldots, x_n)$ est un* échantillon d'apprentissage *si la loi a posteriori correspondante $\pi(\cdot|x_1, \ldots, x_n)$ est propre ; c'est un* échantillon d'apprentissage minimal *si aucun de ses sous-échantillons n'est un échantillon d'apprentissage.*

Exemple 5.11. Pour le modèle $\mathcal{N}(\mu, \sigma^2)$, la taille de l'échantillon d'apprentissage minimal associé à la loi a priori impropre $\pi_0(\mu, \sigma^2) = 1/\sigma^2$ est 2, car

$$\int e^{-\{(x_1-\mu)^2+(x_2-\mu)^2\}/2\sigma^2}\sigma^{-4}d\mu\, d\sigma^2$$
$$= \int_0^\infty \sigma^{-3}e^{-s^2/2\sigma^2}d\sigma^2 = \int_0^\infty \omega^{3/2-2}e^{-s^2\omega/2}d\omega\,,$$

tandis que

[40] Cette section, qui peut être omise dans une première lecture, traite de notions plus avancées, à savoir les lois a priori intrinsèques développées par Berger et Pericchi (1996b,a). Ces notions ne seront pas utilisées dans le reste du livre, sauf dans le Chapitre 7 ; voir Berger et Pericchi (2001), sur qui cette section est fondée, pour une revue beaucoup plus détaillée.

$$\int e^{-(x_1-\mu)^2/2\sigma^2}\sigma^{-3}d\mu\,d\sigma^2 = \infty\,.$$

Si nous considérons maintenant la loi a priori $\pi_1(\mu,\sigma^2) = 1/\sigma$, la taille de l'échantillon d'apprentissage est 3, car

$$\begin{aligned}\int & e^{-\{(x_1-\mu)^2+(x_2-\mu)^2\}/2\sigma^2}\sigma^{-3}d\mu\,d\sigma^2\\ &= \int_0^\infty \sigma^{-2}e^{-s^2/2\sigma^2}d\sigma^2\\ &= \int_0^\infty \omega^{-1}e^{-s^2\omega/2}d\omega = \infty\,,\end{aligned}$$

ce qui est un bon argument en faveur de l'utilisation de la loi π_0 plutôt que la loi π_1. ‖

L'idée est alors d'utiliser un échantillon d'apprentissage minimal, $x_{(\ell)}$, pour transformer la loi a priori impropre π en une loi propre $\pi(\cdot|x_{(\ell)})$ et de traiter cette loi a posteriori *comme si* c'était une loi a priori propre pour le reste de l'échantillon, $x_{(-\ell)}$, afin d'éviter une double utilisation des données, comme dans Aitkin (1991). Lorsqu'on est confronté à une hypothèse H_0 associée à une loi a priori π_0 et une hypothèse alternative H_1 plus générale de loi a priori π_1, si l'échantillon d'apprentissage minimal sous H_1 est tel que $\pi_0(\cdot|x_{(\ell)})$ soit aussi propre, le *pseudo-facteur de Bayes*

$$B_{10}^{(\ell)} = \frac{\int_{\Theta_1} f_1(x_{(-\ell)}|\theta_1)\pi_1(\theta_1|x_{(\ell)})d\theta_1}{\int_{\Theta_0} f_0(x_{(-\ell)}|\theta_0)\pi_0(\theta_0|x_{(\ell)})d\theta_0} \tag{5.5}$$

ne dépend alors pas des constantes de normalisation utilisées dans π_0 et π_1. Une décomposition utile de ce pseudo-facteur de Bayes est proposée dans Berger et Pericchi (2001).

Lemme 5.12. *Dans le cas de lois a priori indépendantes, le pseudo-facteur de Bayes peut s'écrire*

$$B_{10}^{(\ell)} = B_{10}(x) \times B_{01}(x_{(\ell)})\,, \tag{5.6}$$

avec

$$B_{10}(x) = \frac{\int_{\Theta_1} f_1(x|\theta_1)\pi_1(\theta_1)d\theta_1}{\int_{\Theta_0} f_0(x|\theta_0)\pi_0(\theta_0)d\theta_0}$$

et

$$B_{01}(x_{(\ell)}) = \frac{\int_{\Theta_0} f_0(x_{(\ell)}|\theta_0)\pi_0(\theta_0)d\theta_0}{\int_{\Theta_1} f_1(x_{(\ell)}|\theta_1)\pi_1(\theta_1)d\theta_1}\,.$$

Dans cette décomposition, $B_{10}(x)$ et $B_{01}(x_{(\ell)})$ sont les facteurs de Bayes calculés pour des lois a priori non normalisées π_1 et π_0, respectivement pour tout l'échantillon x et l'échantillon d'apprentissage $x_{(\ell)}$, comme s'il s'agissait de lois a priori régulières. Il est alors simple de voir que multiplier π_0 par c_0 et π_1 par c_1 n'a pas d'influence sur $B_{10}^{(\ell)}$, car ces constantes s'annulent. Notons l'intéressante inversion de $B_{10}(x)$ en $B_{01}(x_{(\ell)})$: l'effet de l'échantillon d'apprentissage est retiré du facteur de Bayes $B_{10}(x)$.

Bien que le problème de la constante de normalisation disparaisse, une difficulté majeure est que la solution $B_{10}^{(\ell)}$ n'est que formellement bayésienne. De plus, en dehors des modèles séquentiels, le choix de $x_{(\ell)}$ n'est pas évident, alors qu'il influe pourtant sur la valeur résultante de $B_{10}^{(\ell)}$ (ce qui viole par conséquent le principe de vraisemblance).

Exemple 5.13. (Suite de l'Exemple 5.11) Si $H_0 : \mu = 0$, avec $\pi_0(\sigma^2) = 1/\sigma^2$ et $H_1 : \mu \neq 0$, avec $\pi_1(\mu, \sigma^2) = 1/\sigma^2$, la taille de l'échantillon d'apprentissage minimal est 2 sous H_1. D'où

$$\pi_1(\mu, \sigma^2|x_1, x_2) = \frac{1}{\sigma} \exp\{-2(\mu - \bar{x}_1)^2/2\sigma^2\} s_1^5 \sigma^{-3} e^{-s_1^2/2\sigma^2}$$

et

$$\pi_0(\sigma^2|x_1, x_2) = \frac{s_0^6}{\sigma^4} e^{-s_0^2/2\sigma^2}\,,$$

avec les notations suivantes :

$$\bar{x}_1 = \frac{x_1 + x_2}{2}, \quad s_1^2 = \frac{(x_1 - x_2)^2}{2}, \quad s_0^2 = x_1^2 + x_2^2\,.$$

Alors

$$B_{10}^{(2)} = \frac{s_1^5 \int e^{-\{(n-2)(\bar{x}_2-\mu)^2 - 2(\mu-\bar{x}_1)^2 - s_2^2 - s_1^2\}/2\sigma^2} \sigma^{-n-2} d\mu d\sigma^2}{s_0^6 \int_0^\infty e^{-\{-s_3^2 - s_0^2\}/2\sigma^2} \sigma^{-n-2} d\sigma^2}$$

dépend du choix de (x_1, x_2) via $(\bar{x}_1 - \bar{x}_2)^2$, s_1^2 et s_0^2 (voir l'Exercice 5.15). ‖

Une façon de supprimer cette dépendance à l'échantillon d'apprentissage est de calculer la moyenne des différents pseudo-facteurs de Bayes (5.6) sur tous les échantillons d'apprentissage possibles $x_{(\ell)}$. La difficulté suivante est de décider quel type de moyenne devrait être utilisée. Par exemple, Berger et Pericchi (1996b, 1998, 2001) ont répertorié

– le *facteur de Bayes arithmétique intrinsèque*,

$$B_{10}^A = \frac{1}{L} \sum_{x_{(\ell)}} B_{10}^{(\ell)} = B_{10}(x) \frac{1}{L} \sum_{x_{(\ell)}} B_{01}(x_{(\ell)})\,, \tag{5.7}$$

où L est le nombre des différents échantillons d'apprentissage ;

– le *facteur de Bayes géométrique intrinsèque*,

$$B_{10}^G = \exp \frac{1}{L} \sum_{x_{(\ell)}} \log B_{10}^{(\ell)} = B_{10}(x) \exp \frac{1}{L} \sum_{x_{(\ell)}} \log B_{01}(x_{(\ell)})\,; \qquad (5.8)$$

et

– le *facteur de Bayes médian intrinsèque*,

$$B_{10}^M = \operatorname{med} B_{10}^{(\ell)} = B_{10}(x) \operatorname{med} B_{01}(x_{(\ell)})\,, \qquad (5.9)$$

où $\operatorname{med} B_{10}^{(\ell)}$ indique la médiane des $B_{10}^{(\ell)}$ sur les différents échantillons d'apprentissage.

Bien que toutes ces solutions soient proches d'une réponse bayésienne, en particulier parce qu'elles utilisent les données une seule fois (Exercice 5.16), séparant la partie utilisée pour rendre propre la loi a priori impropre de la partie utilisée pour le test lui-même, aucune d'entre elles n'est vraiment bayésienne. Nous discuterons plus loin des inconvénients plus sérieux de ces différents facteurs de Bayes intrinsèques. Il apparaît cependant que ces derniers correspondent souvent à d'authentiques facteurs de Bayes sous des lois a priori propres, appelées *lois a priori intrinsèques* dans Berger et Pericchi (1996b, 1998)[41]. (On retrouvera ce phénomène dans la Section 5.3.5 avec les bornes inférieures de Berger et Sellke, 1987.)

Exemple 5.14. (Berger et Pericchi, 1998) Dans le cas $x \sim \mathcal{N}(\theta, 1)$, lorsque $H_0 : \theta = 0$ et $\pi_1(\theta) = 1$, pour un échantillon $(x_1, \ldots, x_n)$, le facteur de Bayes arithmétique intrinsèque,

$$B_{10}^A = B_{10}(x) \frac{1}{\sqrt{2\pi}} \frac{1}{n} \sum_{i=1}^{n} e^{-x_i^2/2},$$

est presque identique au facteur de Bayes habituel associé à la loi a priori normale $\mathcal{N}(0, 2)$ sous H_1. ‖

Exemple 5.15. (Berger et Pericchi, 1998) Pour $x_1, \ldots, x_n$, observations i.i.d. d'une loi exponentielle translatée, de densité $\exp(\theta - x)\mathbb{I}_{x \geq \theta}$, si $H_0 : \theta = \theta_0$ et $H_1 : \theta > \theta_0$, avec $\pi_1(\theta) = 1$,

$$B_{10}^A = B_{10}(x) \frac{1}{n} \sum_{i=1}^{n} \left[e^{x_i - \theta_0} - 1 \right]^{-1}$$

[41] Le terme d'*intrinsèque* associé au facteur de Bayes et la loi a priori correspondante tente d'évoquer l'idée de quantités calculées uniquement à partir de la distribution des observations, mais la diversité des réponses possibles montre que ce terme est plutôt inapproprié !

correspond au facteur de Bayes standard associé à la loi a priori propre

$$\pi_2(\theta) = e^{\theta_0-\theta}\left\{1-\log\left(1-e^{\theta_0-\theta}\right)\right\}\,,$$

qui se comporte comme l'indique la Figure 5.1. ‖

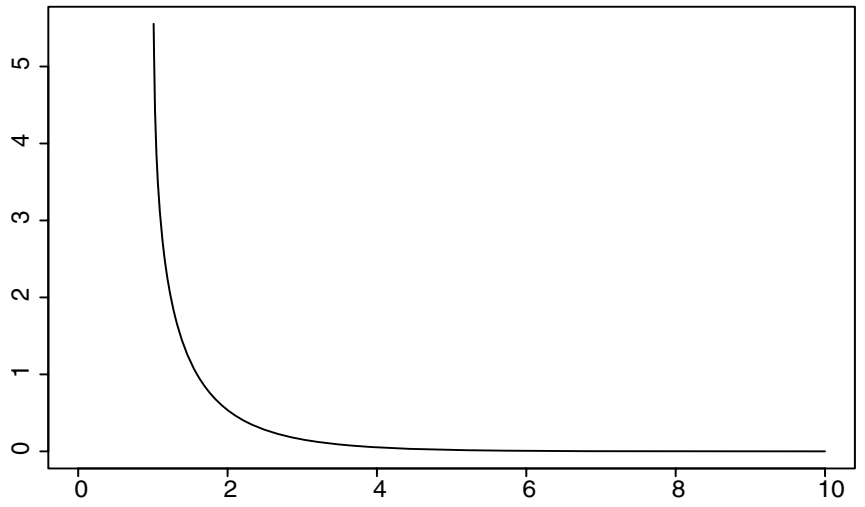

Fig. 5.1. Graphe d'une loi a priori intrinsèque associée au test exponentiel H_0 : $\theta = \theta_0$, lorsque $\theta_0 = 1$.

O'Hagan (1995) présente une alternative élégante aux facteurs de Bayes intrinsèques, alternative qui évite à la fois la sélection d'échantillons d'apprentissage et le calcul de moyenne qui en découle. Son idée est d'utiliser une *fraction* b de la vraisemblance pour rendre propre la loi a priori, c'est-à-dire prendre $0 < b < 1$ tel que

$$\int_{\Theta_0} f_0(x|\theta_0)^b \pi_0(\theta_0) d\theta_0 < \infty$$

et

$$\int_{\Theta_1} f_1(x|\theta_1)^b \pi_1(\theta_1) d\theta_1 < \infty\,.$$

La fraction restante $(1-b)$ de la vraisemblance est alors utilisée pour effectuer le test, comme dans le cas d'un facteur de Bayes intrinsèque. Le *facteur de Bayes fractionnaire* est par conséquent défini comme

$$B_{10}^F = \frac{\int_{\Theta_1} f_1(x|\theta_1)^{1-b}\pi_1^b(\theta_1|x)d\theta_1}{\int_{\Theta_0} f_0(x|\theta_0)^{1-b}\pi_0^b(\theta_0|x)d\theta_0}$$
$$= B_{10}(x)\,\frac{\int_{\Theta_0} f_0(x|\theta_0)^b\pi_0(\theta_0)d\theta_0}{\int_{\Theta_1} f_1(x|\theta_1)^b\pi_1(\theta_1)d\theta_1}\,, \tag{5.10}$$

où $\pi_0^b(\theta_0|x)$ et $\pi_1^b(\theta_1|x)$ indiquent les pseudo-lois a posteriori associées à, respectivement, $f_0(x|\theta_0)^b$ et $f_1(x|\theta_1)^b$. Pour les familles exponentielles, la quantité b correspond clairement à une fraction de taille d'échantillon, car pour n observations d'une famille exponentielle de statistique exhaustive T, on a

$$(\exp\{\theta\cdot n\,T(x) - n\Psi(\theta)\})^b = \exp\{\theta\cdot[bn]\,T(x) - [bn]\Psi(\theta)\}\,.$$

Pour les autres lois, la fraction b doit être déterminée par une approche plus empirique (voir O'Hagan, 1995, 1997).

Comme dans le cas du facteur de Bayes intrinsèque, cette solution est dans certains cas égale à un facteur de Bayes régulier, pour une certaine loi a priori "intrinsèque".

Exemple 5.16. (Suite de l'Exemple 5.14) Pour tout $0 < b < 1$,

$$B_{10}^F = \frac{\int e^{-n(1-b)(\bar{x}-\theta)^2/2}\sqrt{b}e^{-nb((\bar{x}-\theta)^2/2}d\theta}{\sqrt{2\pi}e^{-n(1-b)\bar{x}^2/2}}$$
$$= \sqrt{b}e^{n(1-b)\bar{x}^2/2}\,, \tag{5.11}$$

qui est égal au facteur de Bayes associé à la loi propre $\theta \sim \mathcal{N}(0,(1-b)/nb)$ sous H_1. ‖

Ces pseudo-facteurs de Bayes présentent cependant suffisamment de difficultés pour que nous remettions en cause leur utilisation dans les problèmes de test et de choix de modèle :

(i) Lorsque les facteurs de Bayes sont associés à des lois a priori, ils satisfont certaines propriétés de *cohérence* telles que

$$B_{12} = B_{10}B_{02} \qquad \text{et} \qquad B_{01} = 1/B_{10}\,.$$

La plupart des pseudo-facteurs de Bayes n'y satisfont pas, même si le facteur de Bayes fractionnaire satisfait $B_{01}^F = 1/B_{10}^F$.

(ii) Lorsque les pseudo-facteurs de Bayes peuvent s'exprimer comme de vrais facteurs de Bayes, les lois a priori intrinsèques correspondantes ne sont pas nécessairement satisfaisantes, comme le montrent l'Exemple 5.15 pour le facteur de Bayes arithmétique et l'Exemple 5.16 pour le facteur de Bayes fractionnaire. Ces lois a priori dépendent du choix des lois a priori de référence impropres π_0 et π_1, donc elles ne sont pas véritablement intrinsèques.

(iii) En relation avec le point précédent, les pseudo-facteurs de Bayes peuvent aussi être biaisés vers l'une des hypothèses, au sens où ils peuvent s'exprimer comme un vrai facteur de Bayes multiplié par un certain facteur.

Exemple 5.17. (Suite de l'Exemple 5.15) Pour le facteur de Bayes intrinsèque médian,

$$\begin{aligned} B_{10}^M &= B_{10}(x) \left[e^{\text{med}(x_i)} - \theta_0 \right]^{-1} \\ &= 0.69 \tilde{B}_{10}(x) \end{aligned} \tag{5.12}$$

où $\tilde{B}_{10}(x)$ est le facteur de Bayes associé à la loi a priori $\pi_3(\theta) \propto (2\exp\{\theta - \theta_0\} - 1)^{-1}$, qui, bien qu'elle soit similaire à π_2, ne fournit pas exactement la même couverture des régions proches de 1. ‖

Dans de tels cas, les pseudo-facteurs de Bayes peuvent être perçus comme attribuant aux deux hypothèses des probabilités différentes de la valeur de référence 1/2, une caractéristique que nous rencontrerons aussi pour les bornes les moins favorables dans la Section 5.3.5.

(iv) Le plus souvent cependant, les pseudo-facteurs de Bayes ne correspondent pas du tout à un vrai facteur de Bayes et donnent des solutions fortement biaisées. Par exemple, Berger et Pericchi (2001) confirment que les facteurs de Bayes arithmétiques intrinsèques ne sont pas associés à des lois a priori intrinsèques pour la plupart des problèmes de test unilatéraux.

Exemple 5.18. (Suite de l'Exemple 5.15) Le facteur de Bayes fractionnaire

$$B_{10}^F = B_{10}(x) bn \left\{ e^{-bn(x_{(1)} - \theta_0)} - 1 \right\}^{-1}, \tag{5.13}$$

est toujours plus grand que 1, par conséquent, il favorise toujours l'hypothèse alternative, selon l'échelle de Jeffreys. Ce comportement paradoxal peut être attribué au fait que la fraction b ne modifie pas la fonction indicatrice. ‖

(v) Les pseudo-facteurs de Bayes peuvent simplement ne pas exister pour toute une catégorie de modèles.

Exemple 5.19. Les mélanges de lois normales

$$p\mathcal{N}(\mu_1, \sigma_1^2) + (1-p)\mathcal{N}(\mu_2, \sigma_2^2)$$

ont été présentés dans l'Exemple 1.6. Comme on le voit dans l'Exercice 1.56, les lois a priori impropres de la forme $\pi_1(\mu_1, \sigma_1)\pi_2(\mu_2, \sigma_2)\pi_3(p)$ ne peuvent pas être utilisées dans ce cadre, quelle que soit la taille de l'échantillon n. (La raison fondamentale de cette interdiction est qu'il

existe une probabilité $(1-p)^n > 0$ qu'aucune observation soit associée à la première composante $\mathcal{N}(\mu_1, \sigma_1^2)$.) Par conséquent, il n'existe jamais d'échantillon d'apprentissage pour les lois a priori non informatives standard et on ne peut pas calculer de facteur de Bayes. La même règle s'applique aux facteurs de Bayes fractionnaires (voir l'Exercice 5.22). ‖

(vi) Comme le montre cette section, il existe plusieurs approches pour définir les pseudo-facteurs de Bayes et, bien que la plupart soient sans doute logiques, il n'y a pas de méthode cohérente de les classer par ordre de préférence. Les pseudo-facteurs de Bayes, tels qu'ils sont définis ici, sont en accord avec le principe de vraisemblance, mais la multiplication des réponses possibles, même si celles-ci sont proches, n'est pas un bon signal pour les utilisateurs[42]. De la même façon, il n'existe pas une procédure précise pour le choix de b dans les facteurs de Bayes fractionnaires, car la taille minimale de l'échantillon d'apprentissage n'est pas toujours clairement définie.

(vii) Jusqu'ici, le problème du calcul des pseudo-facteurs de Bayes n'a pas été évoqué, faute d'outils appropriés, qui seront introduits dans les Chapitres 6 et 7. Mais notons que chaque facteur de Bayes $B_{10}^{(\ell)}$ peut être une intégrale complexe et le calcul d'une moyenne de facteurs de Bayes intrinsèques peut impliquer $\binom{m}{n}$ intégrales de ce type, si m est la taille minimale de l'échantillon d'apprentissage. Les facteurs de Bayes fractionnaires sont plus faciles à calculer dans des cadres exponentiels, mais les autres lois sont plus difficiles à manipuler (Exercice 5.23).

5.3 Comparaisons avec l'approche classique

5.3.1 Tests UPP et UPPS

L'approche classique de la théorie des tests est la théorie de Neyman-Pearson, présentée, par exemple, dans Lehmann (1986). Sous le coût $0-1$, noté L ci-dessous, la notion fréquentiste d'optimalité est fondée sur la *puissance* d'un test, définie comme suit :

Définition 5.20. *La* puissance *d'une procédure de test φ est la probabilité de rejeter H_0 sous l'hypothèse alternative, c'est-à-dire $\beta(\theta) = 1 - \mathbb{E}_\theta[\varphi(x)]$ lorsque $\theta \in \Theta_1$. La quantité $1-\beta(\theta)$ est appelée* erreur de deuxième espèce*, tandis que l'*erreur de première espèce *est $\mathbb{E}_\theta[\varphi(x)]$ lorsque $\theta \in \Theta_0$.*

[42] Berger et Pericchi (2001) soutiennent que la multiplicité des facteurs de Bayes intrinsèques possibles n'est pas plus inquiétante que la multiplicité des lois a priori possibles par défaut. La comparaison est cependant légèrement déficiente, puisque *chaque* loi a priori choisie induit de multiples facteurs de Bayes intrinsèques !

Les tests fréquentistes optimaux sont alors ceux qui minimisent le risque $\mathbb{E}_\theta[\mathrm{L}(\theta, \varphi(x))]$ *sous* H_1 *seulement* :

Définition 5.21. *Si* $\alpha \in]0,1[$ *et* $\mathscr{C}_\alpha$ *est la classe des procédures* φ *satisfaisant la contrainte suivante sur l'erreur de première espèce :*

$$\sup_{\theta \in \Theta_0} \mathbb{E}_\theta[\mathrm{L}(\theta, \varphi(x))] = \sup_{\theta \in \Theta_0} P_\theta(\varphi(x) = 0) \leq \alpha, \tag{5.14}$$

une procédure de test φ *est dite* uniformément plus puissante au niveau α *ou UPP si elle minimise dans* $\mathscr{C}_\alpha$ *le risque* $\mathbb{E}_\theta[\mathrm{L}(\theta, \varphi(x))]$ *uniformément sur* Θ_1.

Cette optimalité est beaucoup plus faible que la notion d'admissibilité développée dans la Section 2.4. En effet, le coût est *bidimensionnel*, du fait de la restriction sur l'*erreur de première espèce* (5.14). Cette restriction est généralement nécessaire pour obtenir une procédure de test optimale, car les fonctions de risque des procédures admissibiles se croisent, mais :

(i) Elle entraîne une asymétrie entre les hypothèses nulle et alternative, ce qui implique un comportement anormal des procédures de test. En effet, puisque l'erreur de première espèce est fixée, un équilibre entre les deux erreurs (acceptation sous H_1 et rejet sous H_0) est impossible, d'où une erreur de seconde espèce beaucoup plus grande. Cette asymétrie explique aussi le fait que la théorie ne fasse pas intervenir de considérations de minimaxité. C'est ce qui se passe notamment lorsque deux hypothèses H_0 et H_1 sont *contiguës*, c'est-à-dire lorsqu'il est possible de passer de Θ_0 à Θ_1 par une transformation connexe.

(ii) Elle implique la sélection d'un *niveau de confiance* α par le décideur, en plus du choix de la fonction de coût L, ce qui entraîne généralement le recours à des niveaux "standard", comme 0.05 ou 0.01, et les inconvénients qui sont liés à de tels niveaux "universels" (voir ci-dessous).

(iii) Elle ne suggère pas nécessairement une réduction suffisante de la classe des procédures de test et ne permet pas toujours la sélection d'une procédure unique optimale. Il est parfois nécessaire d'imposer plus de contraintes sur ces classes.

Dans le cas le plus simple, c'est-à-dire si les hypothèses nulle et alternative sont ponctuelles, $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$, le lemme de *Neyman-Pearson* établit l'existence de procédures de test UPP, de la forme[43]

$$\varphi(x) = \begin{cases} 1 & \text{si } f(x|\theta_1) < k f(x|\theta_0), \\ 0 & \text{sinon,} \end{cases}$$

[43] Conservant l'interprétation d'une procédure de test comme estimateur de $\mathbb{I}_{\Theta_0}(\theta)$, les procédures de test sont dans ce livre les compléments à 1 des procédures de Neyman-Pearson classiques, pour lesquelles la valeur de 1 correspond au *rejet* de H_0.

k étant donné par le niveau de confiance choisi α. Évidemment, le fait que Θ_1 se réduise à $\{\theta_1\}$ est assez utile, car ceci permet un ordre total sur les procédures de $\mathscr{C}_\alpha$. Pour les familles à *rapport de vraisemblance monotone*, c'est-à-dire les familles paramétriques pour lesquelles il existe une statistique $T(x)$ telle que

$$\frac{f(x|\theta')}{f(x|\theta)}$$

soit croissant en $T(x)$ pour $\theta' > \theta$, Karlin et Rubin (1956) ont établi l'extension suivante du lemme de Neyman-Pearson (voir Lehmann, 1986, p. 79, pour une démonstration).

Proposition 5.22. *Soit $f(x|\theta)$ à rapport de vraisemblance monotone dans $T(x)$. Pour $H_0 : \theta \leq \theta_0$ et $H_1 : \theta > \theta_0$ il existe un test UPP tel que*

$$\varphi(x) = \begin{cases} 1 & \text{si } T(x) < c, \\ \gamma & \text{si } T(x) = c, \\ 0 & \text{sinon,} \end{cases}$$

γ et c étant déterminés par la contrainte

$$\mathbb{E}_{\theta_0}[\varphi(x)] = \alpha.$$

Karlin et Rubin (1956) ont aussi montré que, pour les fonctions de coût de type (5.1), les procédures de test fournies dans le Théorème 5.22 forment *une classe essentiellement complète*, c'est-à-dire une classe de procédures suffisamment grande pour être au moins aussi bonne que n'importe quelle autre procédure (voir le Chapitre 8). De plus, si le support de la loi $f(x|\theta)$ ne dépend pas de θ, la classe obtenue dans la Proposition 5.22 est *essentiellement complète minimale* : elle ne peut être réduite plus avant (voir Lehmann, 1986, p. 82-83) et, par conséquent elle ne contient *que* les procédures optimales.

Notons qu'une classe importante de familles à rapport de vraisemblance monotone est celle des familles exponentielles, car

$$\frac{f(x|\theta')}{f(x|\theta)} = \frac{e^{\theta' x - \psi(\theta')}}{e^{\theta x - \psi(\theta)}} = \frac{e^{(\theta'-\theta)x}}{e^{\psi(\theta')-\psi(\theta)}}$$

est croissant en x. Pfangzagl (1968) a aussi établi la réciproque de la Proposition 5.22 dans l'esprit du lemme de Pitman-Koopman (Section 3.3.3), à savoir que l'existence d'un test UPP pour toute taille d'échantillon et un niveau donné α implique que la loi appartienne à une famille exponentielle.

Exemple 5.23. Soient $x \sim \mathscr{P}(\lambda)$ et $H_0 : \lambda \leq \lambda_0$, $H_1 : \lambda > \lambda_0$. Pour m observations indépendantes de cette loi, une statistique exhaustive est $s = \sum_i x_i \sim \mathscr{P}(m\lambda)$ et, selon la Proposition 5.22, un test UPP est donné par

$$\varphi(x) = \begin{cases} 1 & \text{si } s < k, \\ \gamma & \text{si } s = k, \\ 0 & \text{sinon,} \end{cases}$$

pour $\mathbb{E}_{\lambda_0}[\varphi(x)] = P_{m\lambda_0}(s > k) + \gamma P_{m\lambda_0}(s = k) = \alpha$. ‖

La Proposition 5.22 et l'exemple ci-dessus mettent en avant une difficulté majeure de l'approche de Neyman-Pearson, à savoir que des niveaux de confiance arbitraires ne sont pas nécessairement accessibles, à moins de faire appel à une *randomisation*. En effet, comme l'espace de décision est $\mathscr{D} = \{0, 1\}$, $\varphi(x) = \gamma$ signifie que $\varphi(x) = 1$ avec probabilité γ (et 0 autrement). De telles procédures sont évidemment incompatibles avec le principe de vraisemblance, même si elles n'apparaissent que pour des cas discrets. Lehmann (1986) indique que le niveau de confiance α devrait être modifié jusqu'à ce que la randomisation soit évitée, mais cette modification provoque un autre inconvénient : le choix du niveau de confiance dépend des observations et non pas d'une fonction d'utilité.

De plus, la Proposition 5.22 s'applique uniquement aux hypothèses unilatérales. Dans un cas particulier d'hypothèses bilatérales, nous pouvons exposer un résultat d'optimalité (voir Lehmann, 1986, p. 101-103).

Proposition 5.24. *Soient une famille exponentielle*

$$f(x|\theta) = e^{\theta T(x) - \psi(\theta)} h(x)$$

et $H_0 : \theta \leq \theta_1$ *ou* $\theta \geq \theta_2$, $H_1 : \theta_1 < \theta < \theta_2$. *Il existe un test UPP de la forme*

$$\varphi(x) = \begin{cases} 0 & \textit{si } c_1 < T(x) < c_2, \\ \gamma_i & \textit{si } T(x) = c_i \quad (i = 1, 2), \\ 1 & \textit{sinon,} \end{cases}$$

avec $(i = 1, 2)$

$$\mathbb{E}_{\theta_i}[\varphi(x)] = \alpha.$$

Cependant, il n'existe pas de test UPP correspondant au cas opposé, à savoir $H_0 : \theta_1 \leq \theta \leq \theta_2$. Ce paradoxe montre avec force l'absence de symétrie–et donc de cohérence–du critère UPP et jette un doute sur la validité de l'analyse de Neyman-Pearson ou sur la pertinence d'un coût asymétrique comme le coût $0 - 1$. Dans ces cas, la solution de Neyman-Pearson est de proposer une réduction additionnelle de la classe des procédures en considérant des *tests sans biais*, c'est-à-dire satisfaisant de plus

$$\sup_{\Theta_0} P_\theta(\varphi(x) = 0) \leq \inf_{\Theta_1} P_\theta(\varphi(x) = 0).$$

En d'autres termes, φ doit aussi satisfaire

$$\inf_{\Theta_0} \mathbb{E}_\theta[\varphi(x)] \geq \sup_{\Theta_1} \mathbb{E}_\theta[\varphi(x)].$$

La notion de tests *uniformément plus puissants sans biais* (UPPS) en découle. Néanmoins, cette restriction provoque encore plus d'asymétrie entre H_0 et H_1. Bien qu'intuitivement acceptable, cette notion de test sans biais est un autre exemple des restrictions imposées à la notion d'optimalité par l'approche fréquentiste, qui dénaturent le vrai objectif de la Théorie de la Décision.

Exemple 5.25. Si, pour $x \sim \mathcal{N}(\theta, 1)$, on teste $H_0 : \theta = 0$ contre $H_1 : \theta \neq 0$, il n'existe pas de test UPP. Un test UPPS au niveau $\alpha = 0.05$ est

$$\varphi(x) = \begin{cases} 1 & \text{si } |x| \leq 1.96, \\ 0 & \text{sinon.} \end{cases}$$

‖

5.3.2 Lois a priori les moins favorables

Lorsque aucun test UPPS n'existe, il devient assez difficile de défendre, ou même de construire, une procédure de test spécifique dans un cadre fréquentiste. À moins de restreindre plus encore la classe des procédures acceptables, une approche habituelle est de considérer le rapport de vraisemblance

$$\frac{\sup_{\theta \in \Theta_0} f(x|\theta)}{\sup_{\theta \in \Theta_1} f(x|\theta)} \tag{5.15}$$

et sa distribution, ou de fonder le test sur la loi asymptotique de (5.15). Le rapport ci-dessus illustre un lien avec l'approche bayésienne, car, comme on l'a déjà dit précédemment, il s'agit formellement d'un facteur de Bayes pour une loi a priori π de support réduit aux points $\hat{\theta}_0$ et $\hat{\theta}_1$, estimateurs du maximum de vraisemblance de θ sur Θ_0 et Θ_1. Cette analogie est en effet formelle, puisque les masses de Dirac sont des lois a priori artificielles et, de plus, les $\hat{\theta}_i$ dépendent des observations. Cependant, elle indique aussi que le rapport de vraisemblance a une motivation bayésienne.

Des relations entre procédures de test bayésiennes et procédures optimales de Neyman-Pearson sont présentées dans Lehmann (1986), via la notion de *lois les moins favorables*, décrite ci-dessous[44]. Soient $H_0 : \theta \in \Theta_0$, $H_1 : \theta = \theta_1$ avec π une loi a priori sur Θ_0. D'un point de vue bayésien, ce problème de test

[44] Le reste de cette section n'est pas utilisé dans la suite. La passerelle signalée ici est d'importance moindre que la relation correspondante obtenue dans la théorie de la minimaxité (voir la Section 2.4.3). De plus, elle ne peut s'appliquer qu'à des cas spécifiques et ne valide pas plus avant les réponses classiques, qui ne peuvent pas être obtenues comme limites de procédures bayésiennes (voir la Section 5.4).

peut être représenté comme le test de $H_\pi : x \sim m_\pi$ contre $H_1 : x \sim f(x|\theta_1)$, où m est la loi marginale sous H_0

$$m_\pi(x) = \int_{\Theta_0} f(x|\theta)\pi(\theta)\, d\theta.$$

Puisque les deux hypothèses (H_π et H_1) sont des hypothèses ponctuelles, le lemme de Neyman-Pearson assure l'existence d'un test UPP φ_π, à un niveau de signification α et de puissance $\beta_\pi = P_{\theta_1}(\varphi_\pi(x) = 0)$. Ce test est de la forme

$$\varphi_\pi(x) = \begin{cases} 1 & \text{si } m_\pi(x) > kf(x|\theta_1), \\ 0 & \text{sinon.} \end{cases}$$

Définition 5.26. *Une* loi la moins favorable *est une loi a priori π qui maximise la puissance β_π.*

Cette définition est utilisée dans le résultat suivant (Lehmann, 1986, p. 105).

Théorème 5.27. *Soit $H_0 : \theta \in \Theta_0$ à tester contre l'alternative $H_1 : \theta = \theta_1$. Si le test UPP φ_π au niveau α pour H_π contre H_1 satisfait*

$$\sup_{\theta \in \Theta_0} \mathbb{E}_\theta[\mathrm{L}(\theta, \varphi_\pi)] \leq \alpha,$$

alors

(i) *φ_π est UPP au niveau α ;*

(ii) *si φ_π est le seul test de niveau α de H_π contre H_1, φ_π est le seul test UPP au niveau α pour tester H_0 contre H_1 ; et*

(iii) *π est une loi la moins favorable.*

La condition dans le théorème ci-dessus peut sembler superflue, mais notons que φ_π est défini par

$$\int_{\{m_\pi(x) > kf(x|\theta_1)\}} m_\pi(x)\, dx = \alpha.$$

Ce rapport ne garantit pas que $\mathbb{E}_\theta[\mathrm{L}(\theta, \varphi_\pi)] \leq \alpha$ pour tout $\theta \in \Theta_0$.

5.3.3 Critiques

Le Théorème 5.27 exhibe une connexion entre les tests bayésien et UPP, de la même façon que les lois les moins favorables mènent aux estimateurs minimax dans les problèmes d'estimation ponctuelle avec une valeur (voir la Section 2.4), bien qu'une procédure de Bayes corresponde à un test modifié impliquant π. Nous ne poursuivrons pas l'analogie au-delà de cette connexion, car, comme d'autres, nous nous opposons à l'approche de Neyman-Pearson

dans son ensemble. En effet, en plus des problèmes de randomisation évoqués ci-dessus, un inconvénient majeur de cette perspective est de restreindre l'espace de décision au couple $\{0,1\}$, ce qui oblige par conséquent à prendre une décision catégorique. Il nous semble qu'une réponse plus adaptative est préférable. De plus, les tests UPP (et UPPS), lorsqu'ils existent, dépendent d'une mesure d'évaluation (le niveau de signification α) non révisée après observation. Dans l'Exemple 5.25 notamment, si le niveau est fixé à 0.05, la réponse classique est identique pour $x = 1.96$ et $x = 100$. D'un point de vue purement décisionnel, il semble aussi paradoxal de restreindre les procédures inférentielles à un cadre limité, puisque ce dernier peut (et doit) mener à des procédures sous-optimales. En particulier, la notion de "*sans biais*", qui a été déconsidérée en estimation ponctuelle grâce à l'effet Stein (Note 2.8.2), devrait aussi disparaître des procédures de test.

Une critique plus fondamentale de l'approche de Neyman-Pearson (et, au fond, de toute approche fréquentiste) est qu'elle fonde le rejet de H_0 sur des *événements improbables qui ne se sont pas produits*, pour reprendre les termes de Jeffreys (1939, 1961). En effet, une région de rejet UPP est de la forme

$$\mathscr{R} = \{T(X) \geq T(x)\}$$

si la loi a un rapport de vraisemblance monotone en T, car, sous l'hypothèse nulle,

$$P(T(X) \geq T(x)) < \alpha. \tag{5.16}$$

Cependant, l'événement qui se produit en réalité est $\{T(X) = T(x)\}$. Il y a donc perte d'information dans le processus (classique) de décision, qui se trouve en général être biaisé contre l'hypothèse nulle. En effet, la région $\{T(X) \geq T(x)\}$ est relativement plus improbable qu'un voisinage de $T(x)$, ce qui explique le fait que les réponses bayésiennes soient plus optimistes (voir la Section 5.3.5). Bien entendu, la seule approche cohérente qui permette de conditionner sur $\{T(X) = T(x)\}$, c'est-à-dire sur les observations elles-mêmes, est l'approche bayésienne. En revanche, choisir une procédure sur la base de (5.16) fait intervenir la loi complète de x et, par conséquent, contredit potentiellement le principe de vraisemblance, comme le montrent les Exemples 1.16 et 1.18. En effet, le principe des règles d'arrêt n'est pas compatible avec une théorie des tests fréquentiste, car la loi de la taille de l'échantillon ne devrait pas avoir d'impact sur la sélection de la procédure de test. Le principe de vraisemblance présente effectivement la propriété paradoxale qu'une procédure fondée sur un rapport de vraisemblance reste acceptable tant qu'elle ne dépend pas de la loi de ce rapport.

Exemple 5.28. Le *test du khi deux* est une procédure simple (mais approximative) pour tester l'adéquation d'un échantillon à une loi (ou une famille de lois). Si l'échantillon de taille n est divisé en k classes, de tailles théoriques $N_i = np_i$ et de tailles observées n_i, on déduit du Théorème Central Limit que

$$D^2 = \sum_{i=1}^{k} \frac{(n_i - N_i)^2}{N_i}$$

est approximativement distribué comme une loi du χ^2_ℓ, de degrés de liberté ℓ dépendant du problème (et valant généralement $k-1$ moins le nombre de paramètres estimés). Comme l'a souligné Jeffreys (1961), l'approche classique rejette l'hypothèse nulle (adéquation à la famille des lois proposées) si D^2 est trop grand, par exemple, si

$$P(z > D^2) < 0.05$$

pour $z \sim \chi^2_\ell$. Cependant, il n'y a pas de raison d'accepter l'hypothèse nulle (qui est que D^2 est approximativement distribué comme χ^2_ℓ) si

$$P(z < D^2) \leq 0.05,$$

puisque de telles valeurs de D^2 ne sont pas plus compatibles avec la loi que lorsque $P(z > D^2) \leq 0.05$. De ce point de vue, il serait aussi justifié de rejeter l'hypothèse nulle, ce que ne fait pas l'approche classique. ‖

Exemple 5.29. Une critique bayésienne bien connue de la théorie de Neyman-Pearson est le contre-exemple suivant présenté par Lindley (1957, 1961). Soient $\bar{x}_n \sim \mathcal{N}(0, 1/n)$ la moyenne d'un échantillon normal et $\theta \sim \mathcal{N}(0,1)$. Pour tester $H_0 : \theta = 0$ contre $H_1 : \theta \neq 0$, les tests UPPS correspondants ne dépendent que de $z_n = |x_n|\sqrt{n}$. Supposons $z_n = 1.97$. Au niveau de significativité 5%, la procédure de test rejette H_0 pour tout n. Au contraire, la probabilité a posteriori de H_0 est (voir l'Exemple 5.4)

$$\pi(\theta = 0|z_n) = \left(1 + \frac{1-\varrho_0}{\varrho_0} \frac{1}{\sqrt{n+1}} \exp\{z_n^2 n/2(n+1)\}\right)^{-1},$$

et par conséquent tend vers 1 quand n tend vers l'infini. En fait, ce résultat se vérifie pour la plupart des lois a priori, de par la normalité asymptotique des lois a posteriori (voir Hartigan, 1983). Ce paradoxe peut être relié au *problème de Kepler* (voir Jeffreys, 1961 ou Berger, 1985b), qui est que, en astronomie, une hypothèse nulle–par exemple, la nature elliptique de la trajectoire des planètes–est toujours rejetée d'un point de vue fréquentiste pour une taille d'échantillon suffisamment grande, c'est-à-dire lorsque suffisamment d'observations ont été accumulées. ‖

Une autre difficulté majeure de l'approche de Neyman-Pearson est que la sélection du niveau α devrait être équivalente à la sélection des poids a_0 et a_1 dans la fonction de coût et que, par conséquent, elle devrait être fondée sur des considérations d'utilité. Au lieu de cela, la pratique courante d'omettre

complètement cette étape de sélection et, suivant une suggestion faite par Fisher (1956), de choisir un niveau α classique de 5% ou 1%, est à présent devenue une règle formelle, quels que soient le problème, la taille de l'échantillon, ou l'erreur de seconde espèce. Puisque l'approche de Neyman-Pearson est plutôt prédominante de nos jours, cette attitude dogmatique a entraîné un biais de publication. En effet, les résultats des expériences qui ne sont pas "*significatifs au niveau* 5%" sont le plus souvent rejetés par les éditeurs ou même censurés par les auteurs eux-mêmes dans plusieurs domaines, incluant la biologie, la médecine et les sciences sociales.

5.3.4 Les *p*-values

Les fréquentistes (et praticiens) ont tenté de compenser les inconvénients de l'approche de Neyman-Pearson en supprimant le niveau de signification α et en proposant une réponse prenant ses valeurs dans $[0,1]$ et, de façon plus importante, dépendant des observations de manière plus adaptative qu'une acceptation ou un rejet établis en comparant $T(x)$ à un seuil donné. La notion suivante a été introduite pour la première fois par Fisher (1956).

Définition 5.30. *La* p-value *associée à un test est le niveau de signification* α *le plus petit pour lequel l'hypothèse nulle est rejetée.*

Une définition générale pour les hypothèses nulles ponctuelles (voir Thompson, 1989) est qu'une p-value est une statistique admettant une loi uniforme sous l'hypothèse nulle ; se pose alors le difficile problème du choix de l'une de ces statistiques, comme d'ailleurs pour le *test* introduit dans la définition ci-dessus. En réalité, si un test de région critique R_α est disponible pour tout niveau de signification α et si ces régions sont imbriquées (c'est à dire si $R_\alpha \subset R_\beta$ pour $\beta > \alpha$), la procédure

$$p(x) = \inf\{\alpha;\, x \in R_\alpha\}$$

est distribuée selon une loi uniforme si $\mathbb{E}_{\theta_0}[\mathbb{I}_{R_\alpha}(x)] = \alpha$ (voir Goutis *et al.*, 1996). Dans l'éventualité de plusieurs tests donnant des réponses opposées, nous suggérons d'utiliser la loi du rapport de vraisemblance sous l'hypothèse nulle, si cette dernière est ponctuelle.

Exemple 5.31. (Suite de l'Exemple 5.25) Puisque la région critique (qui est la région de rejet pour H_0) du test UPPS est $\{|x| > k\}$, une p-value usuelle est

$$\begin{aligned} p(x) &= \inf\{\alpha;\, |x| > k_\alpha\} \\ &= P^X(|X| > |x|), \qquad\qquad X \sim \mathcal{N}(0,1) \\ &= 1 - \Phi(|x|) + \Phi(|x|) = 2[1 - \Phi(|x|)]. \end{aligned}$$

Par conséquent, si $x = 1.68$, $p(x) = 0.10$ et, si $x = 1.96$, $p(x) = 0.05$. ‖

Exemple 5.32. Soit $x \sim \mathscr{B}(n,p)$, lorsque l'hypothèse à tester est $H_0 : p = 1/2$ contre $H_1 : p \neq 1/2$. La p-value associée au rapport de vraisemblance

$$\frac{f(x|1/2)}{\sup_p f(x|p)} = \frac{(1/2)^n}{\left(\frac{x}{n}\right)^x \left(1-\frac{x}{n}\right)^{n-x}} \propto x^{-x}(n-x)^{-(n-x)}$$

est la fonction

$$\tilde{p}(x) = P_{1/2}\left(X^X(n-X)^{(n-X)} \leq x^x(n-x)^{(n-x)}\right),$$

où $X \sim \mathscr{B}(n,1/2)$. ‖

Les p-values sont donc des procédures adaptatives qui peuvent être acceptables d'un point de vue fréquentiste et qui, en outre, répondent aux exigences de Kiefer (1977) et Robinson (1979) d'une *approche fréquentiste conditionnelle*. Cependant, elles restent critiquées, car

(i) Les p-values évaluent aussi la mauvaise quantité, à savoir, la probabilité de *dépasser* la valeur observée de la statistique de test. Elles contredisent donc le principe de vraisemblance, car elles dépendent de toute la loi des observations.

(ii) Même si elles sont calculées à partir de procédures de test optimales, les p-values ne sont pas intrinsèquement optimales, car elles ne sont pas évaluées sous une fonction de coût. En effet, comme le montre la Section 5.4, elles peuvent être sous-optimales.

(iii) Le nouvel espace de décision, $\mathscr{D} = [0,1]$, n'est pas motivé par des considérations de Théorie de la Décision et donc l'utilisation des p-values n'est pas rendue explicite. En particulier, les p-values sont souvent perçues comme fournissant une approximation fréquentiste de $P(\theta \in \Theta_0|x)$, même si cette expression n'a pas de sens dans un cadre non bayésien.

(iv) Dans une perspective classique, les p-values ne résument pas toute l'information disponible pour un problème de test ; elles devraient être comparées aux *erreurs de seconde espèce*, qui sont habituellement omises dans l'analyse. Berger et Wolpert (1988) illustrent le danger de n'utiliser que des p-values dans l'exemple suivant. Si $x \sim \mathscr{N}(\theta,1/2)$, tester $\theta = -1$ contre $\theta = 1$ lorsque $x = 0$ mène à une p-value de 0.072 (pour un test UPP), indiquant apparemment un fort rejet de l'hypothèse nulle, alors que la p-value correspondante pour le test inverse de H_1 contre H_0 prend exactement la même valeur. En fait, un rejet de H_0 ne devrait pas toujours impliquer l'acceptation de H_1, cependant les praticiens considèrent souvent la p-value comme étant *la* procédure de test et supposent qu'elle englobe toute l'information sur le problème de test en jeu et concluent néanmoins à l'acceptation. (Voir la Note 5.7.4.)

5.3.5 Réponses bayésiennes moins favorables

Le problème d'évaluation des p-values sous un coût adapté est considéré dans la Section 5.4. Nous terminons cette section par une comparaison entre les p-values et leurs contreparties bayésiennes, les probabilités a posteriori. Considérer la probabilité a posteriori la plus petite pour une classe de lois a priori fournit la *réponse bayésienne la moins favorable* par rapport à l'hypothèse nulle. Cette limite inférieure ne peut pas être utilisée comme une procédure non informative, car elle sélectionne la loi a priori la plus opposée à l'hypothèse nulle et elle est à la fois biaisée contre H_0 et dépendante des observations. Elle devrait être interprétée comme un indicateur des *variations* des probabilités a posteriori, la réponse la plus favorable étant 1. Une littérature étendue est désormais disponible sur cette approche et les lecteurs pourront consulter Berger et Sellke (1987), Berger et Delampady (1987) et Berger et Mortera (1991) pour des références supplémentaires. La Note 5.7.4 présente une perspective différente due à Berger *et al.* (1997) qui réconcilient les tests fréquentistes et bayésiens en modifiant le cadre décisionnel.

Berger et Sellke (1987) et Berger et Delampady (1987) considèrent le cas d'une hypothèse nulle ponctuelle, $H_0 : \theta = \theta_0$, contre l'hypothèse alternative $H_1 : \theta \neq \theta_0$. Pour une famille G de lois a priori sous l'hypothèse alternative, les mesures d'évaluation de la vraisemblance de H_0 sont données par les limites inférieures

$$\underline{B}(x, G) = \inf_{g \in G} \frac{f(x|\theta_0)}{\int_\Theta f(x|\theta) g(\theta)\, d\theta},$$
$$\underline{P}(x, G) = \inf_{g \in G} \frac{f(x|\theta_0)}{f(x|\theta_0) + \int_\Theta f(x|\theta) g(\theta)\, d\theta}$$

sur les facteurs de Bayes et les probabilités a posteriori (pour $\varrho_0 = 1/2$, de façon à donner des poids égaux aux deux hypothèses). Ces limites peuvent aussi s'écrire

$$\underline{B}(x, G) = \frac{f(x|\theta_0)}{\sup_{g \in G} \int_\Theta f(x|\theta) g(\theta) d\theta}, \quad \underline{P}(x, G) = \left[1 + \frac{1}{\underline{B}(x, G)}\right]^{-1}.$$

Elles varient bien évidemment en fonction de la classe G considérée. Dans un cas plus général, lorsque G est égal à G_A, l'ensemble de toutes les lois a priori, le résultat suivant se démontre aisément.

Lemme 5.33. *S'il existe un estimateur du maximum de vraisemblance de θ, $\hat{\theta}(x)$, les limites inférieures des facteurs de Bayes et des probabilités a posteriori de H_0 sont, respectivement,*

$$\underline{B}(x, G_A) = \frac{f(x|\theta_0)}{f(x|\hat{\theta}(x))}, \quad \underline{P}(x, G_A) = \left[1 + \frac{f(x|\hat{\theta}(x))}{f(x|\theta_0)}\right]^{-1}.$$

Une conséquence du Lemme 5.33 est que la réponse bayésienne ne sera jamais *fortement* en faveur de l'hypothèse nulle, car

$$\underline{B}(x, G_A) \leq 1, \qquad \underline{P}(x, G_A) \leq \frac{1}{2}.$$

Ce comportement n'est pas particulièrement surprenant, car les limites inférieures correspondent au pire choix possible de g par rapport à H_0. Un phénomène plus inattendu est que la décroissance de ces limites lorsque $|x|$ augmente est plus lente que pour les p-values, comme le montre l'exemple suivant.

Exemple 5.34. (Suite de l'Exemple 5.31) Dans le cas gaussien, les limites inférieures associées à $H_0 : \theta_0 = 0$ sont

$$\underline{B}(x, G_A) = e^{-x^2/2} \quad \text{et} \quad \underline{P}(x, G_A) = \left(1 + e^{x^2/2}\right)^{-1},$$

ce qui donne le Tableau 5.7, qui compare les p-values aux réponses bayésiennes les moins favorables.

La différence avec les réponses fréquentistes est donc assez importante. Les p-values sont plus petites pour des niveaux de signification usuels et rejettent donc l'hypothèse nulle H_0 "trop souvent". Bien entendu, pour des valeurs plus petites de x, les p-values sont plus grandes que les limites inférieures, mais le point le plus important est que, pour les valeurs de x où la décision est le plus difficile à prendre, soit donc pour des niveaux de signification entre 0.01 et 0.1, une telle divergence apparaisse entre les réponses fréquentistes et bayésiennes. ‖

Tab. 5.7. Comparaison entre les p-values et les réponses bayésiennes dans un cas gaussien. (*Source* : Berger et Sellke, 1987.)

p-value	0.10	0.05	0.01	0.001
$\underline{P}$	0.205	0.128	0.035	0.004
$\underline{B}$	0.256	0.146	0.036	0.004

Des résultats de ce type sont assez surprenants, car les procédures classiques appartiennent habituellement à la gamme des réponses bayésiennes. De plus, la classe G_A est plutôt déraisonnable, car elle inclut des masses de Dirac menant à la limite inférieure. La seule justification pour ce type de lois a priori se rapporte au principe minimax et à la notion correspondante de loi la moins favorable. L'exemple ci-dessus montre que les p-values ne sont pas minimax en ce sens. Bien entendu, la divergence est plus importante pour des classes de lois plus petites. Par exemple, si G est égal à G_S, l'ensemble des lois qui sont symétriques en θ_0, l'équivalent du Lemme 5.33 est :

Lemme 5.35. *Le facteur de Bayes le plus petit lorsque $g \in G_S$ est*

$$\underline{B}(x, G_S) = \frac{f(x|\theta_0)}{\sup_\xi \frac{1}{2}[f(x|\theta_0 - \xi) + f(x|\theta_0 + \xi)]},$$

qui mène à la limite inférieure correspondante pour les probabilités a posteriori.

Ce résultat se déduit du fait que toute loi symétrique est un mélange de lois dont le support se réduit à deux points, de la forme $\{\theta_0 - \xi, \theta_0 + \xi\}$. Pour des extensions multidimensionnelles, le suprémum doit être pris sur les lois uniformes pour des sphères centrées sur θ_0 (voir Berger et Delampady, 1987). Les problèmes discrets nécessitent quelques raffinements, notamment la définition d'une notion de loi symétrique. Par exemple, dans le cas binomial, la classe correspondante est G_S, l'ensemble des lois qui sont symétriques en

$$\frac{p - p_0}{\sqrt{p(1-p)}}.$$

Exemple 5.36. (Suite de l'Exemple 5.32) Pour $H_0 : p = 1/2$, le Tableau 5.8 fournit les p-values et les limites inférieures bayésiennes associées à G_S ($p_0 = 1/2$). ‖

Tab. 5.8. Comparaison entre p-values et réponses bayésiennes dans un cas binomial. (*Source :* Berger et Delampady, 1987.)

p-value	0.0093	0.0507	0.1011
$\underline{P}$	0.0794	0.2210	0.2969

Notons que dans ce cas les p-values ne sont pas des niveaux standard, de par la nature discrète de la loi binomiale.

Une autre classe intéressante de lois a priori est celle des lois unimodales symétriques en θ_0, G_{SU}. Ces lois peuvent s'écrire comme des mélanges de lois symétriques uniformes en dimension 1 (Berger et Sellke, 1987). Cependant, le calcul des limites inférieures reste faisable. De telles classes sont nécessaires dans des cadres multidimensionnels, car les limites inférieures associées à des classes plus générales comme G_A sont proches de 0 pour la plupart des valeurs des observations.

Exemple 5.37. (Suite de l'Exemple 5.25) Dans le cas gaussien, si $|x| \leq 1$, $\underline{B}(x, G_{SU}) = 1$ et $\underline{P}(x, G_{SU}) = 1/2$. Cependant, si $|x| > 1$ et si on définit $g(\theta) = (1/2K)\mathbb{I}\{|\theta| < K\}$, on a

$$\int f(x|\theta)g(\theta)\,d\theta = \frac{1}{2K}[\Phi(K - x) - \Phi(-K - x)]$$

et la limite inférieure est associée au K maximisant cette expression. Le Tableau 5.9 donne les valeurs de $\underline{B}$ et $\underline{P}$ correspondant aux p-values de 0.1 et 0.01, qui diffèrent significativement de la réponse fréquentiste. ‖

Tab. 5.9. Réponses bayésiennes pour les p-values de 0.01 (haut) et 0.1 (bas) dans le cas normal. (*Source :* Berger et Delampady, 1987.)

dim.	1	3	5
$\underline{P}$	0.109	0.083	0.076
	0.392	0.350	0.339
$\underline{B}$	0.123	0.090	0.082
	0.644	0.540	0.531

Une première conséquence de cette comparaison est que, d'un point de vue bayésien, les p-values ne sont pas un outil valable pour mettre en œuvre des expériences de test d'hypothèses nulles. Contrairement aux problèmes réguliers d'estimation ponctuelle comme ceux développés dans le Chapitre 4, les réponses fréquentistes ne semblent pas s'exprimer comme limites de réponses bayésiennes ; nous donnons dans la Section 5.4 une preuve formelle de ce fait. Puisque les p-values sont strictement plus petites que les réponses bayésiennes (pour des niveaux qui comptent vraiment dans un processus de test décisionnel), l'hypothèse nulle H_0 est rejetée plus souvent sous une approche fréquentiste, tandis que l'approche bayésienne montre que le rapport des vraisemblances a posteriori de H_0 et H_1 est assez modéré pour des niveaux de signification usuels (0.05 ou 0.01). Cette différence importante entre les deux approches justifie clairement une modélisation bayésienne, car cette approche inclut plus naturellement la notion de probabilité d'une hypothèse. Elle montre aussi que l'argument de *validité fréquentiste*, c'est-à-dire la justification de long terme fournie par un niveau de signification de 5% ou de 1%, est plutôt illusoire et que la division introduite par la théorie de Neyman-Pearson dans le traitement de H_0 et H_1 (entre les erreurs de première et de seconde espèces) mène à un biais en faveur de l'hypothèse alternative pour des valeurs plus grandes de x ou $T(x)$.

5.3.6 Le cas unilatéral

Les hypothèses unilatérales (c'est-à-dire $H_0 : \theta \leq \theta_0$ contre $H_1 : \theta > \theta_0$) n'exhibent pas de tels contrastes entre solutions fréquentistes et solutions bayésiennes. En effet, comme le montre l'Exemple 5.9, la p-value peut alors s'écrire comme un estimateur de Bayes généralisé et donc comme une limite de solutions bayésiennes (puisque la renormalisation n'a pas d'impact). Par conséquent, il n'est pas possible d'exhiber une dichotomie entre les deux approches comme dans le cas bilatéral. Casella et Berger (1987) considèrent ce cadre et généralisent le phénomène de "réconciliation" décrit plus haut.

Théorème 5.38. *Soit $x \sim f(x-\theta)$, avec f symétrique en 0. L'hypothèse nulle à tester est $H_0 : \theta \leq 0$. Si f est une loi à rapport de vraisemblance monotone, la p-value $p(x)$ est égale à la limite inférieure des probabilités a posteriori, $\underline{P}(x, G_{SU})$, lorsque cette limite est calculée sur la classe G_{SU} des lois a priori symétriques unimodales et lorsque $x > 0$.*

Preuve. Dans ce cas la p-value est

$$p(x) = P_{\theta=0}(X > x) = \int_x^{+\infty} f(t)\,dt$$

et

$$\begin{aligned}
\underline{B}(x, G_{SU}) &= \inf_{\pi \in G_{SU}} P^{\pi}(\theta \leq 0|x) \\
&= \inf_{\pi \in G_{SU}} \frac{\int_{-\infty}^{0} f(x-\theta)\pi(\theta)\,d\theta}{\int_{-\infty}^{+\infty} f(x-\theta)\pi(\theta)\,d\theta} \\
&= \inf_{K} \frac{\int_{-K}^{0} f(x-\theta)\,d\theta}{\int_{-K}^{K} f(x-\theta)\,d\theta}, \qquad (5.17)
\end{aligned}$$

de par la représentation des lois a priori unimodales symétriques comme mélanges de lois uniformes sur $[-K, K]$. La propriété de rapport de vraisemblance monotone implique que (5.17) est atteint en $K = +\infty$. □

Une conséquence du Théorème 5.18 est que la limite inférieure des réponses bayésiennes sur toutes les lois a priori est plus petite que la p-value.

Exemple 5.39. Soit $X \sim \mathscr{C}(\theta, 1)$, la loi de Cauchy, et l'hypothèse à tester est $H_0 : \theta \leq 0$ contre $H_1 : \theta > 0$. Si la loi a priori de θ est supposée appartenir à la classe des lois symétriques par rapport à 0, la limite inférieure des réponses bayésiennes et les p-values correspondantes sont données dans le Tableau 5.10. Les différences entre les valeurs numériques ne sont pas aussi frappantes que dans les exemples précédents. ‖

Tab. 5.10. Comparaison entre les p-values et les probabilités a posteriori bayésiennes dans le cas d'une loi de Cauchy. (*Source :* Casella et Berger, 1987.)

p-value	0.437	0.102	0.063	0.013	0.004
$\underline{P}$	0.429	0.077	0.044	0.007	0.002

Cette différence entre les cas unilatéral et bilatéral appelle les commentaires suivants :

(i) Comme il a déjà été dit plusieurs fois, une modélisation bayésienne est généralement assez délicate dans les cas bilatéraux, en particulier pour des hypothèses nulles ponctuelles, car cela implique une *modification de la loi a priori imposée par le problème inférentiel*. Ceci ne contredit pas les principes bayésiens si nous considérons que cette modification est le résultat d'une information (vague) additionnelle ; mais la façon d'utiliser cette information reste incertaine. Une illustration de cette difficulté est donnée par le cas des lois non informatives, où plusieurs approches bayésiennes (et pas entièrement compatibles) donnent des résultats contradictoires, comme le détaille la Section 5.2.6.

(ii) Que la p-value soit proche de la limite inférieure dans le cas unilatéral montre le comportement conservateur (ou minimax) de la procédure. Puisque cette dernière peut s'écrire comme une réponse bayésienne généralisée, cela nous incite à penser que la p-value devrait aussi s'exprimer comme une réponse non informative dans les cas bilatéraux. Bien entendu, cela n'implique pas forcément que cette réponse devrait être utilisée, car une utilisation efficace de l'information contenue dans le problème de test lui-même est généralement possible.

(iii) Les p-values sont construites à partir de tests UPP ou UPPS par une construction empirique sur mesure. Les comparaisons dans Berger et Sellke (1987) et Casella et Berger (1987) montrent qu'elles diffèrent (ou non) de leurs contreparties bayésiennes. Bien que ces études signalent l'existence d'un problème théorique, elles ne sont pas suffisantes d'un point de vue fréquentiste pour rejeter l'utilisation des p-values. Il est donc nécessaire d'utiliser une perspective décisionnelle adaptée à l'évaluation des p-values. La section suivante traite de cette comparaison. Elle fournit aussi des explications théoriques à la dichotomie bilatérale/unilatérale présentée ci-dessus.

(iv) Une perspective différente, qui permet d'agrandir l'espace de décision en incluant l'option "pas de décision", donne des réponses fréquentistes et bayésiennes beaucoup plus proches, conceptuellement et numériquement. Elle est détaillée dans la Note 5.7.4.

5.4 Une deuxième approche décisionnelle

Comme on vient de le souligner[45], les p-values n'ont pas de justification intrinsèque, car leur prétendue "optimalité" découle de celle des procédures de test, dont elles sont dérivées. En un sens, la même remarque s'applique aux probabilités a posteriori, car, bien qu'elles soient intuitivement justifiables, celles-ci ne sont pas validées par un processus de décision. Dans cette section, nous construisons une alternative à l'approche de Neyman-Pearson pour justifier les probabilités a posteriori et évaluer les p-values.

[45] Cette section, de niveau plus avancé, peut être omise lors d'une première lecture.

Comme le montre la Section 5.2, le problème de test formalisé par Neyman et Pearson peut s'exprimer comme l'estimation de la fonction indicatrice $\mathbb{I}_{\Theta_0}(\theta)$ sous le coût $0-1$ ou, de façon équivalente, *le coût en erreur absolue*

$$\mathrm{L}_1(\theta,\varphi) = |\varphi - \mathbb{I}_{\Theta_0}(\theta)|\,. \tag{5.18}$$

En effet, si les estimateurs φ ne prennent que les valeurs 0 et 1, il existe de nombreuses manières d'écrire le coût $0-1$, (5.18) étant l'une d'elles. Mais, comme il est indiqué ci-dessus, la théorie de Neyman-Pearson est essentiellement une théorie "pré-données" que ne fournit pas de solution "post-données" (ou plus adaptative). Nous nous tournons alors vers une théorie moins restrictive, pour laquelle les estimateurs prennent leurs valeurs dans $\mathscr{D} = [0,1]$ et peuvent être considérés comme des indicateurs du degré de certitude contre ou en faveur de H_0.

Parallèlement à Schaafsma *et al.* (1989), Hwang *et al.* (1992) examinent cette approche des problèmes de test, pour laquelle les estimateurs de $\mathbb{I}_{\Theta_0}(\theta)$ appartiennent à $[0,1]$. Lorsque la restriction à $\{0,1\}$ est levée, le choix du coût devient plus important. Par exemple, (5.18) est trop semblable à la fonction de coût $0-1$, car elle fournit les mêmes procédures de Bayes

$$\varphi^\pi(x) = \begin{cases} 1 & \text{si } P^\pi(\theta \in \Theta_0|x) > P^\pi(\theta \notin \Theta_0|x), \\ 0 & \text{sinon.} \end{cases}$$

En revanche, les coûts strictement convexes, comme les coûts quadratiques

$$\mathrm{L}_2(\theta,\varphi) = \left(\varphi - \mathbb{I}_{\Theta_0}(\theta)\right)^2, \tag{5.19}$$

mènent à des estimateurs plus adaptatifs.

Proposition 5.40. *Sous le coût* (5.19)*, l'estimateur de Bayes associé à* π *est la probabilité a posteriori*

$$\varphi^\pi(x) = P^\pi(\theta \in \Theta_0|x).$$

En effet, l'espérance a posteriori de $\mathbb{I}_{\Theta_0}(\theta)$ n'est autre que la probabilité a posteriori de Θ_0. Le coût quadratique (5.19) fournit alors une base décisionnelle pour l'utilisation de probabilités a posteriori comme réponses bayésiennes. De tels coûts sont dits *réguliers* (voir Lindley, 1985 et Schervish, 1989 ; l'Exercice 2.15 caractérise ces coûts). Il existe d'autres coûts réguliers que les coûts quadratiques, mais Hwang et Pemantle (1994) ont montré qu'il suffit de considérer le coût quadratique en termes d'*admissibilité* et de *classes complètes* (voir aussi le Chapitre 8).

Nous examinons dans cette section le cas particulier des familles exponentielles naturelles,

$$f(x|\theta) = e^{\theta x - \psi(\theta)}, \qquad \theta \in \Theta \subset \mathbb{R},$$

et nous introduisons la définition suivante, due à Farrell (1968b), qui nous permet d'évaluer les procédures dans un intervalle lorsqu'elles sont constantes en dehors de cet intervalle.

Définition 5.41. *Pour un test unilatéral, c'est-à-dire pour une hypothèse de la forme* $H_0 : \theta \leq \theta_0$ *contre* $H_1 : \theta > \theta_0$*, un intervalle* $[t_1, t_2]$ *est appelé* ensemble de troncature *pour l'estimateur* φ *si* $\varphi(t) = 1$ *lorsque* $t < t_1$ *et* $\varphi(t) = 0$ *lorsque* $t > t_2$*. Pour un test bilatéral de* $H_0 : \theta \in [\theta_1, \theta_2]$*, l'intervalle* $[t_1, t_2]$ *est appelé* ensemble de troncature *pour l'estimateur* φ *si* $\varphi(t) = 0$ *lorsque* $t \notin [t_1, t_2]$.

Les résultats suivants ont été obtenus par Hwang *et al.* (1992), à partir des travaux de Brown (1986b) ; celui-ci montre que tout estimateur admissible est une limite ponctuelle d'estimateurs de Bayes pour une suite de mesures de support fini (voir la Section 8.3.4).

Théorème 5.42. *Pour le problème bilatéral*

$$H_0 : \theta \in [\theta_1, \theta_2] \quad \textit{contre} \quad H_1 : \theta \notin [\theta_1, \theta_2], \tag{5.20}$$

un estimateur φ *d'ensemble de troncature* $[t_1, t_2]$ *est admissible s'il existe une mesure de probabilité* π_0 *sur* $[\theta_1, \theta_2]$ *et une mesure* σ*-finie* π_1 *sur* $[\theta_1, \theta_2]^c$ *telles que*

$$\varphi(x) = \frac{\int f(x|\theta)\pi_0(\theta)\, d\theta}{\int f(x|\theta)\pi_0(\theta) d\theta + \int f(x|\theta)\pi_1(\theta)\, d\theta}, \tag{5.21}$$

pour $x \in [t_1, t_2]$*. Réciproquement, si* φ *est admissible, il existe* $[t_1, t_2]$*,* π_0 *et* π_1 *tels que* (5.21) *soit satisfait.*

Dans le cas unilatéral, nous ne pouvons proposer qu'une condition nécessaire d'admissibilité, mais celle-ci implique que les estimateurs de Bayes généralisés forment une classe complète.

Théorème 5.43. *Pour le problème unilatéral*

$$H_0 : \theta \leq \theta_0 \quad \textit{contre} \quad H_1 : \theta > \theta_0, \tag{5.22}$$

si φ *est admissible, il existe une procédure croissante* φ' *telle que* φ' *est équivalente à* φ *(en termes de risque). Si* φ *est une procédure admissible croissante et* $[t_1, t_2]$ *est un ensemble de troncature tel que* $0 < \varphi(x) < 1$ *sur* $[t_1, t_2]$*, il existe deux mesures* σ*-finies sur* $(-\infty, \theta_0]$ *et* $[\theta_0, +\infty)$*,* π_0 *et* π_1*, telles que*

$$1 = \int e^{t_0\theta - \psi(\theta)}(\pi_0(\theta) + \pi_1(\theta))\, d\theta$$

pour $t_1 < t_0 < t_2$ *et* φ *est donné par* (5.21) sur $[t_1, t_2]$.

Ces deux théorèmes de *classes complètes* montrent qu'il suffit de considérer des estimateurs de Bayes généralisés pour obtenir des estimateurs admissibles sous un coût quadratique. Le Théorème 5.43 montre de plus que les estimateurs monotones forment une *classe essentiellement complète*. Ces résultats peuvent être utilisés pour évaluer les p-values. Rappelons de nouveau que les estimateurs de Bayes sous-tendent les estimateurs optimaux (classiques). (Le Chapitre 8 expose plus en détail les bases bayésiennes de l'admissibilité.)

Rappelons aussi que Casella et Berger (1987) ont montré que les p-values prenaient des valeurs sensiblement similaires à celles des probabilités a posteriori bayésiennes dans des cadres unilatéraux. Il est donc naturel d'examiner l'admissibilité des p-values. Les exemples ci-dessous montrent qu'elles sont admissibles pour la plupart des tests unilatéraux.

Exemple 5.44. Soient de nouveau $x \sim \mathcal{N}(\theta, 1)$ et H_0 de la forme (5.22). Nous avons montré dans l'Exemple 5.9 que

$$p(x) = P_{\theta_0}(X > x) = 1 - \Phi(x - \theta_0)$$

est un estimateur de Bayes généralisé par rapport à la mesure de Lebesgue. De plus, le risque de la p-value est

$$\begin{aligned} r(\pi, p) &= \int_{-\infty}^{+\infty} R(p, \theta)\, d\theta \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (p(x) - \mathbb{I}_{\Theta_0}(\theta))^2 f(x|\theta)\, dx\, d\theta \\ &= \int_{-\infty}^{\theta_0} \int_{-\infty}^{+\infty} (1 - \Phi(x - \theta_0))^2 f(x|\theta)\, dx\, d\theta \\ &\qquad + \int_{\theta_0}^{+\infty} \int_{-\infty}^{+\infty} \Phi(x - \theta_0)^2 f(x|\theta)\, dx\, d\theta \\ &= 2 \int_{-\infty}^{+\infty} (1 - \Phi(x - \theta_0))^2 \Phi(x - \theta_0)\, dx \end{aligned}$$

par le théorème de Fubini. Cette intégrale est finie. Par conséquent, $r(\pi) < +\infty$ et p est admissible sous (5.19) (voir Section 2.4). ‖

Exemple 5.45. Soit $x \sim \mathcal{B}(n, \theta)$. La p-value pour le test de (5.21) est alors

$$p(x) = P_{\theta_0}(X \geq x) = \sum_{k=x}^{n} \binom{n}{k} \theta_0^k (1 - \theta_0)^{n-x},$$

qui est aussi un estimateur de Bayes généralisé sous la loi a priori $\pi(\theta) = 1/\theta$. Il est de nouveau possible de montrer que p a un risque de Bayes fini et est par conséquent admissible. Un résultat similaire peut être établi pour une loi de Poisson, $\mathcal{P}(\theta)$ (voir Hwang *et al.*, 1992). ‖

En revanche, les p-values ne sont pas admissibles dans les cas bilatéraux, comme le suggèrent les comparaisons de la Section 5.3.5.

Théorème 5.46. *Pour le test de* (5.20)*, lorsque la distribution d'échantillonnage est absolument continue par rapport à la mesure de Lebesgue, la p-value est inadmissible pour le coût* (5.19)*.*

Preuve. Ce résultat repose sur le fait que la p-value vaut 1 avec une probabilité strictement positive (voir Hwang *et al.*, 1992, Section 4.1.2). En effet, si p est admissible, elle peut s'écrire sous la forme (5.21). Puisqu'elle est positive,

$$\int f(x|\theta)\pi_1(\theta)\,d\theta < +\infty.$$

Par conséquent, l'égalité (5.21) est par continuité vraie partout et $p(x_0) = 1$ implique $\pi = \pi_0$, soit $p(x) = 1$ pour tout x, ce qui ne peut pas être vrai. □

Ce résultat s'accorde avec les observations de Berger et Sellke (1987), qui ont montré que les p-values n'appartiennent pas à la catégorie des réponses bayésiennes. Cela justifie donc le rejet des p-values pour les hypothèses bilatérales. En outre, Hwang et Pemantle (1994) ont montré que l'inadmissibilité des p-values peut s'étendre à la plupart des coûts réguliers bornés. Comme remarque finale, notons qu'il semble désormais nécessaire de construire des estimateurs qui dominent les p-values. Dans le cas normal, Hwang *et al.* (1992) montrent que cela ne peut pas être fait avec un estimateur de Bayes régulier, tandis que Hwang et Pemantle (1994) donnent des arguments numériques en faveur d'un estimateur dominant explicite.

5.5 Régions de confiance

En plus de fournir au décideur des approximations de la "vraie" valeur du paramètre θ, à savoir des estimateurs ponctuels et des réponses aux questions sur l'inclusion de θ dans un domaine spécifique, c'est-à-dire des procédures de test, il est souvent nécessaire de construire également des *régions de confiance* pour θ, sous-ensembles C_x de l'espace des paramètres Θ où θ devrait se trouver avec une forte probabilité (dans un sens fréquentiste ou bayésien). Cette notion s'étend aussi aux transformations non bijectives de θ. Elle est par ailleurs d'un intérêt considérable dans les problèmes de prévision.

Exemple 5.47. Reprenons le prix des actions IBM de l'Exemple 4.23, représenté dans la Figure 4.2. Si les séries (x_t) ont été observées jusqu'au temps T, la valeur au temps $T+1$, x_{T+1}, est évidemment cruciale et il est important de ne pas communiquer à l'investisseur uniquement la valeur la plus probable de x_{T+1}, sachant les observations précédentes, mais aussi l'éventail des valeurs vraisemblables de x_{T+1}, afin qu'il puisse prendre une décision par rapport aux profits possibles correspondants. ‖

Une fois de plus, le fait que, dans la formulation bayésienne, θ ait une probabilité *donnée* d'appartenir à une région *fixée* C_x est plus attrayant que l'interprétation fréquentiste d'une région *aléatoire* C_x ayant une probabilité *donnée* de contenir le paramètre inconnu θ.

5.5.1 Intervalles de crédibilité

Comme dans le cadre des tests, le paradigme bayésien propose une notion de région de confiance qui est plus naturelle que son équivalent fréquentiste, car la notation $P(\theta \in C_x)$ a un sens même conditionnellement à x.

Définition 5.48. *Pour une loi a priori π, un ensemble C_x est un* ensemble α-crédible *si*

$$P^\pi(\theta \in C_x|x) \geq 1-\alpha.$$

Cet ensemble est appelé région α-crédible HPD *(HPD pour* Highest Posterior Density*, soit densité a posteriori la plus forte) s'il peut s'écrire sous la forme*[46]

$$\{\theta;\ \pi(\theta|x) > k_\alpha\} \subset C_x^\pi \subset \{\theta;\ \pi(\theta|x) \geq k_\alpha\},$$

où k_α est la plus grande borne telle que

$$P^\pi(\theta \in C_x^\alpha|x) \geq 1-\alpha.$$

Considérer uniquement les régions HPD est motivé par le fait qu'elles *sont de volume minimal parmi les régions α-crédibles* et, par conséquent, peuvent être perçues comme des solutions optimales dans un cadre de décision.

Exemple 5.49. Si $\theta \sim \mathcal{N}(0,\tau^2)$, la loi a posteriori de θ est $\mathcal{N}(\mu(x),\omega^{-2})$ avec $\omega^2 = \tau^{-2}+\sigma^{-2}$ et $\mu(x) = \tau^2 x/(\tau^2+\sigma^2)$. Alors

$$C_\alpha^\pi = \left[\mu(x) - k_\alpha\omega^{-1}, \mu(x) + k_\alpha\omega^{-1}\right],$$

où k_α est le quantile $\alpha/2$ de $\mathcal{N}(0,1)$. En particulier, si τ tend vers $+\infty$, $\pi(\theta)$ converge vers la mesure de Lebesgue sur $\mathbb{R}$ et donne

$$C_\alpha = [x - k_\alpha\sigma, x + k_\alpha\sigma],$$

c'est-à-dire l'intervalle de confiance habituel, en tant qu'estimateur de Bayes généralisé. ‖

Exemple 5.50. Soient $x \sim B(n,p)$ et la loi non informative $p \sim \mathcal{B}e(1/2, 1/2)$. Alors $p|x \sim \mathcal{B}e(x+1/2, n-x+1/2)$ et les intervalles de confiance pour p peuvent être calculés à partir de la fonction de répartition de la loi bêta. Le Tableau 5.11 donne ces intervalles pour $n=5$ et $\alpha = 5\%$, 10%. ‖

Tab. 5.11. Intervalles α-crédibles pour la loi binomiale $\mathscr{B}(n,p)$.

x	0	1	2
$\alpha = 5\%$	$[0.000, 0.38]$	$[0.022, 0.621]$	$[0.094, 0.791]$
$\alpha = 10\%$	$[0.000, 0.308]$	$[0.036, 0.523]$	$[0.128, 0.74]$

Notons l'avantage significatif de l'approche bayésienne par rapport à l'approche classique pour traiter des *lois discrètes*. En effet, les intervalles de confiance classiques requièrent une étape de *randomisation* pour atteindre les niveaux de confiance standard (voir Blyth, 1961, pour une illustration dans un cas binomial). Une modélisation a priori évite cette adjonction d'un bruit aléatoire et, au contraire, tire profit de l'information a priori disponible. Notons aussi que *les lois a priori impropres* peuvent être utilisées dans ce cadre, sans présenter les mêmes difficultés que pour des hypothèses nulles ponctuelles. En effet, les régions crédibles a posteriori peuvent être obtenues dès que la loi a posteriori est définie. Certaines régions de confiance peuvent s'exprimer comme des régions crédibles associées à des lois généralisées.

Exemple 5.51. Soient $x_1, \ldots, x_n$ i.i.d. $\mathscr{N}(\theta, \sigma^2)$ et la loi a priori non informative

$$\pi(\theta, \sigma^2) = \frac{1}{\sigma^2}.$$

Nous avons montré dans la Section 4.4.2 que la loi a posteriori marginale pour $1/\sigma^2$ est une loi gamma $\mathscr{G}\left((n-1)/2, s^2/2\right)$ avec $s^2 = \sum(x_i - \bar{x})^2$. Par conséquent,

$$\frac{s^2}{\sigma^2}|\bar{x}, s^2 \sim \chi^2_{n-1}$$

et nous obtenons le même intervalle de confiance que dans l'approche classique, mais sa justification est ici conditionnelle à s^2. ‖

Exemple 5.52. Soient $x \sim \mathscr{B}(n,p)$ et $p \sim \mathscr{B}e(\alpha, \beta)$. Dans ce cas, $\pi(p|x)$ est la loi $\mathscr{B}e(\alpha + x, \beta + n - x)$. Selon les valeurs de α, β, n et x, les régions de confiance sont de quatre types :

(i) $0 \le p \le K(x)$;

(ii) $K(x) \le p \le 1$;

(iii) $K_1(x) \le p \le K_2(x)$; et

(iv) $0 \le p \le K_1(x)$ ou $K_2(x) \le p \le 1$.

La dernière région est assez artificielle et plutôt inutile. Notons qu'elle correspond au cas

$$\alpha + x < 1 \quad \text{et} \quad \beta + n - x < 1,$$

[46] Cette formulation permet de couvrir le cas particulier où $\{\theta; \pi(\theta|x) = k_\alpha\}$ n'est pas vide.

ce qui implique par conséquent que α et β doivent être suffisamment négatifs, car $\alpha+\beta < 2-n$. Cette possibilité disparaît donc pour n assez grand, à moins que α et β ne dépendent de n, ce qui n'est pas désirable d'un point de vue bayésien. De plus, le cas limite $\alpha = \beta = 0$, qui correspond à la loi de Haldane (1931)

$$\pi(p) = [p(1-p)]^{-1},$$

conduit déjà aux régions de types (i)-(iii), bien que la loi a posteriori ne soit pas définie pour tous les x (Exemple 1.27). ‖

Lorsque des phénomènes comme ceux du cas (iv) de l'Exemple 5.52 se produisent, c'est-à-dire lorsque la région de confiance n'est pas connexe (voir aussi l'Exemple 5.5), la solution habituelle est de remplacer la région α-crédible HPD par un intervalle à queues égales, soit $[C_1(x), C_2(x)]$ tel que

$$P^\pi(\theta < C_1(x)|x) = P^\pi(\theta > C_2(x)|x) = \alpha/2.$$

Berger (1985b) fait remarquer que l'occurrence de régions HPD non connexes met aussi en lumière une divergence entre la loi a priori et les observations, et que ce phénomène devrait conduire à une remise en question du choix de la loi a priori ou de la distribution de l'échantillon. Il peut aussi permettre d'exhiber une structure de non-identifiabilité responsable de la multimodalité de la loi a posteriori.

Si la construction d'ensembles crédibles est plutôt simple conceptuellement, la détermination pratique de ces régions peut être assez complexe, en particulier lorsque la dimension de Θ est grande ou lorsque la loi a posteriori n'est pas disponible explicitement. Une première solution est d'utiliser des méthodes numériques similaires à celles développées dans le Chapitre 6, le problème étant d'évaluer l'erreur correspondante (qui peut être beaucoup plus grande que les erreurs d'approximation dans les problèmes d'estimation ponctuelle). (Notons que les régions crédibles à queues égales sont généralement plus faciles à approcher que les régions HPD ; voir Eberly et Casella, 1999.) Une deuxième solution, suggérée par Berger (1980b, 1985b), est d'utiliser une approximation normale, donc de considérer que la loi a posteriori de θ est approximativement $\mathcal{N}_p(\mathbb{E}^\pi(\theta|x), \mathrm{Var}^\pi(\theta|x))$ et de construire les régions de confiance à partir de cette approximation

$$C_\alpha = \left\{\theta; (\theta - \mathbb{E}^\pi(\theta \mid x))^t \,\mathrm{Var}^\pi(\theta|x)^{-1}(\theta - \mathbb{E}^\pi(\theta|x)) \leq k_\alpha^2\right\},$$

où k_α^2 est le quantile de niveau α de χ_p^2. Cette approximation n'est justifiée que pour une grande taille d'échantillon (voir Hartigan, 1983), mais elle permet des calculs rapides et plutôt efficaces.

5.5.2 Intervalles de confiance classiques

Dans la théorie de Neyman-Pearson, les régions de confiance peuvent se déduire des tests UPPS par un argument de *dualité* : Si

$$C_\theta = \{x; \varphi_\theta(x) = 1\}$$

est la région d'acceptation de l'hypothèse nulle $H_0 : \theta = \theta_0$, φ_{θ_0} étant un test UPPS au niveau α, la région de confiance correspondante est

$$\begin{aligned} C_x &= \{\theta; x \in C_\theta\} \\ &= \{\theta; \varphi_\theta(x) = 1\} \end{aligned}$$

et $P(\theta \in C_x) = 1 - \alpha$. De façon plus générale, une région C_x est dite *région de confiance au niveau* α (dans un sens fréquentiste) si, pour tout $\theta \in \Theta$, $P(\theta \in C_x) \geq 1 - \alpha$.

Exemple 5.53. (Suite de l'Exemple 5.49) Si $x \sim \mathcal{N}(\theta, \sigma^2)$, le test UPPS à 95% est $\varphi_\theta(x) = \mathbb{I}_{[0,1.96]}(|x - \theta|/\sigma)$ et la région de confiance correspondante, lorsque σ est connu, est

$$C_x = [x - 1.96\sigma, x + 1.96\sigma].$$

‖

Exemple 5.54. Soit $x \sim \mathcal{T}_p(N, \theta, I_p)$, loi de Student à N degrés de liberté de densité

$$f(x \mid \theta) \propto \left(1 + \frac{1}{N} \parallel x - \theta \parallel^2\right)^{-(N+p)/2}.$$

Puisque $||x - \theta||^2/p \sim \mathcal{F}(p, N)$, nous pouvons construire une *boule de confiance* au niveau $1 - \alpha\%$

$$C_x = \left\{\theta; \parallel x - \theta \parallel^2 \leq p f_\alpha(p, N)\right\},$$

où $f_\alpha(p, N)$ est le quantile de niveau α de $\mathcal{F}(p, N)$. ‖

Ces régions de confiance, bien qu'elles soient utilisées de façon assez extensive dans la pratique (par exemple, dans le cas des régressions linéaires), ont été critiquées en termes fréquentistes, conditionnels et bayésiens. Tout d'abord, comme on l'a vu dans les sections précédentes, l'approche de Neyman-Pearson elle-même n'est pas sans inconvénient et l'optimalité des tests UPPS peut être contestée. Par conséquent, les régions de confiance construites à partir de ces tests (appelées *régions uniformément plus précises* par Lehmann, 1986) n'ont pas nécessairement un comportement adéquat. De plus, même dans une perspective fréquentiste, la transformation de procédures de test optimales en régions de confiance n'accorde pas automatiquement à ces régions une forme d'optimalité, malgré la dénomination ci-dessus.

En plus des critiques conditionnelles des régions de confiance (voir la Note 5.7.3), il existe aussi des critiques fréquentistes. À la suite de Stein (1962a) et Lindley (1962), Brown (1966) et Joshi (1967a) ont en effet établi que ces

régions C_x^0 ne sont pas toujours optimales, car il peut exister un autre ensemble C_x' tel que

$$P_\theta(\theta \in C_x') \geq P_\theta(\theta \in C_x^0) \quad \text{et} \quad \text{vol}(C_x') \leq \text{vol}(C_x^0).$$

Par conséquent, l'ensemble C_x' est préférable à C_x^0, car, pour un volume plus petit, il a une probabilité plus grande de contenir la vraie valeur du paramètre. Par exemple, dans le cas normal, Joshi (1967a) a établi que, si $x \sim \mathcal{N}_p(\theta, I_p)$, la région de confiance

$$C_x^0 = \left\{\theta;\ ||\theta - x||^2 \leq c_\alpha\right\}$$

est admissible (au sens ci-dessus) si et seulement si $p \leq 2$ (voir aussi Cohen et Strawderman, 1973). Pour des dimensions plus grandes, il est possible d'exhiber des régions de confiance plus efficaces.

Ce phénomène se rapporte à l'effet Stein, qui établit la non-admissibilité de l'estimateur du maximum de vraisemblance pour $p \geq 3$ (voir la Note 2.8.2). Hwang et Casella (1982) ont tiré profit de cette analogie pour montrer que, si

$$\delta^{\text{JS}}(x) = \left(1 - \frac{a}{||x||^2}\right)^+ x$$

est un estimateur de James-Stein tronqué, la *région de confiance recentrée*

$$C_x^{\text{JS}} = \left\{\theta;\ ||\theta - \delta^{\text{JS}}(x)||^2 \leq c_\alpha\right\},$$

a le même volume que la boule usuelle C_x^0 et satisfait

$$P_\theta(\theta \in C_x^{\text{JS}}) > P_\theta(\theta \in C_x^0) = 1 - \alpha \tag{5.23}$$

pour a suffisamment petit. Par conséquent, C_x^{JS} domine C_x^0 dans le sens ci-dessus.

Une part importante de la littérature sur les régions de confiance recentrées a été initiée par Hwang et Casella (1982, 1984), à l'instar des développements sur l'estimation ponctuelle associée à l'effet Stein (voir la Section 2.8.2). De nouvelles régions recentrées ont été proposées par Hwang et Casella (1984) et Casella et Hwang (1983, 1987). Hwang et Chen (1986) et Robert et Casella (1990) ont élargi les résultats de domination aux lois à symétrie sphérique, bien que le cas gaussien avec variance inconnue soit toujours sans solution (voir Hwang et Ullah, 1994). Shinozaki (1990) a aussi imaginé une région de confiance avec exactement la même probabilité de couverture, mais avec un volume plus petit, tirant profit de la non-admissibilité de la région usuelle d'une façon opposée à (5.23). Lu et Berger (1989a), Robert et Casella (1993) et George et Casella (1994) se sont aussi inspirés de (5.23) pour proposer des estimateurs de confiance améliorés pour les ensembles standard et recentrés. Pour le problème d'estimation de la variance d'une loi normale, des améliorations similaires sont données par Cohen (1972), Shorrock (1990) et Goutis et Casella (1991).

5.5.3 Évaluation décisionnelle des ensembles de confiance

Comme les lecteurs ont pu le constater, la construction des régions de confiance ci-dessus a été conduite de manière plutôt empirique, pour des justifications décisionnelles limitées. Le choix des régions HPD est généralement lié à la nécessité de minimiser le volume de cette région, sous une contrainte de couverture

$$P(\theta \in C_\alpha | x) \geq 1 - \alpha.$$

Plusieurs auteurs ont proposé des constructions différentes des régions de confiance selon des critères purement décisionnels. Ces auteurs considèrent des fonctions de coût intégrant simultanément les exigences de volume et de couverture. (Dans un sens, l'approche ci-dessus correspond à un coût bidimensionnel, dont les composantes sont $\text{vol}(C)$ et $1 - \mathbb{I}_C(\theta)$.) Par exemple, une version simple de cette perspective décisionnelle est de considérer une combinaison linéaire

$$\text{L}(C, \theta) = \text{vol}(C) + c\mathbb{I}_{\theta \notin C}, \tag{5.24}$$

ce qui donne le risque

$$R(C, \theta) = \mathbb{E}[\text{vol}(C_x)] + cP(\theta \notin C_x).$$

(La constante c peut être reliée à un niveau de confiance particulier.) De plus, Cohen et Sackrowitz (1984) ont montré que le coût bidimensionnel ci-dessus est lié au coût linéaire (5.24) lorsque c est traité comme un paramètre supplémentaire du modèle.

Un défaut important des coûts (5.24) a été souligné par James Berger (voir Casella *et al.*, 1993b,a) : Le problème provient d'une pénalisation inégale entre volume et couverture. En effet, la fonction indicatrice varie entre 0 et 1 tandis que le volume peut augmenter jusqu'à l'infini ; cette asymétrie mène à un biais en faveur des ensembles de confiance petits.

Exemple 5.55. Soient $x_1, \ldots, x_n$ i.i.d. $\mathcal{N}(\theta, \sigma^2)$. L'intervalle classique de Student en θ,

$$C_k(\bar{x}, s) = \left[\bar{x} - k\frac{s}{\sqrt{n}}, \bar{x} + k\frac{s}{\sqrt{n}} \right],$$

est une région HPD lorsque

$$\bar{x} = \sum_{i=1}^{n} x_i/n, \quad s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2/(n-1), \quad \text{et} \quad \pi(\theta, \sigma^2) = \frac{1}{\sigma^2},$$

loi non informative de Jeffreys. Dans ce cas, en effet,

$$\sqrt{n}\,\frac{\theta - \bar{x}}{s} \mid \bar{x}, s \sim \mathcal{T}_{n-1}\,.$$

Sous (5.24), le coût a posteriori est

$$\begin{aligned}\varrho(\pi, C_k(\bar{x}, s)|\bar{x}, s) &= 2k\frac{s}{\sqrt{n}} - cP^{\pi}(\theta \in C_k(\bar{x}, s)|\bar{x}, s) \\ &= 2k\frac{s}{\sqrt{n}} - cP(|T_{n-1}| \leq k).\end{aligned}$$

Il est alors facile de voir que la région HPD est dominée par une région tronquée

$$C_t'(\bar{x}, s) = \begin{cases} C_t(\bar{x}, s) & \text{si } s < \sqrt{n}c/(2k), \\ \{\bar{x}\} & \text{sinon.} \end{cases}$$

Cette domination est contraire à l'intuition : C_t' ne contient que le point $\{\bar{x}\}$, ce qui semble indiquer une forte certitude, alors que la variance empirique augmente, ce qui signifie que l'incertitude grandit. Un phénomène similaire se produit lorsque k dépend de s : la taille de la région de crédibilité décroît vers 0 quand s augmente (voir Casella *et al.*, 1993b,a). ‖

Le paradoxe ci-dessus montre les limitations du coût linéaire (5.24). Casella et al. (1993a) proposent une classe alternative de fonctions de coût qui évite ce paradoxe. Le plus simple de ces coûts est le coût dit *rationnel*

$$\mathrm{L}(C, \theta) = \frac{\mathrm{vol}(C)}{\mathrm{vol}(C) + k} + \mathbb{I}_{\theta \notin C} \qquad (k > 0),$$

où les deux termes sont inférieurs à un. Les estimateurs de Bayes associés à ces coûts restent des régions HPD mais sont non vides pour toutes les lois a priori conjuguées dans le cas normal. Le paramètre k peut s'obtenir par des techniques similaires à celles développées pour des coûts réguliers, à savoir en comparant les pénalisations associées au volume pour des régions différentes et en approchant la fonction d'utilité.

Nous n'irons pas plus loin dans l'étude décisionnelle des régions de confiance bayésiennes. En effet, un aspect important souvent négligé dans la construction de régions de confiance est la façon dont elles seront utilisées, bien que cette façon soit essentielle dans la construction de la fonction de coût. En effet, l'objectif du décideur peut être de

(1) considérer l'estimation d'ensemble comme une étape préliminaire à une phase *d'estimation ponctuelle* (et, par exemple, construire une loi a priori empirique de support égal à la région de confiance estimée) ;
(2) se fonder sur la région de confiance obtenue pour résoudre un problème de *test* (et rejeter l'hypothèse nulle si la région de confiance ne contient pas une certaine valeur) ;
(3) déduire de la taille (volume) de la région de confiance un indicateur de *performance* d'un estimateur associé, par exemple, le centre de la région. Une *courbe de performance* pour cet estimateur peut être obtenue en faisant correspondre la taille et les niveaux de confiance.

Ces trois perspectives de l'estimation par régions de confiance mènent à des fonctions de coût fondamentalement différentes et il peut paraître irréaliste

d'essayer de construire une fonction de coût globale unifiant des objectifs si opposés. En effet, des fonctions de coût distinctes sont préférables, car, en accord avec les bases de la Théorie de la Décision, le décideur devrait choisir une fonction de coût selon ses besoins. Notons aussi que les trois objectifs considérés ci-dessus correspondent à des problèmes inférentiels déjà étudiés auparavant et donc qu'une approche spécifique aux régions de confiance peut être partiellement inutile. Par conséquent, il nous semble que, pour le moins, une approche *conditionnelle* devrait être utilisée pour la construction de régions de confiance (voir la Note 5.7.3). À la suite de Kiefer (1977), nous suggérons d'associer à l'ensemble donné C_x un *indicateur de confiance* $\gamma(x)$, évalué sous le coût

$$\mathrm{L}(C, \gamma, \theta) = (\mathbb{I}_C(\theta) - \gamma)^2. \tag{5.25}$$

La région de confiance est alors remplacée par une procédure de confiance, liée à la perspective conditionnelle de Robinson (1979). De ce point de vue, la procédure $\{\Theta, 1\}$ est malheureusement parfaite, un inconvénient qui indique qu'une évaluation additionnelle de C_x devrait être incluse dans la fonction de coût, comme dans Rukhin (1988a,b). De la même façon, la procédure bayésienne associée à une région HPD C_α est $[C_\alpha, 1-\alpha]$, comme on peut le vérifier en minimisant le coût a posteriori. Pour une région arbitraire, C_x, la procédure correspondante est $[C_x, \gamma^\pi(x)]$, où

$$\gamma^\pi(x) = P^\pi(\theta \in C_x | x).$$

L'introduction d'une fonction de coût globale combinant volume, couverture et rapport de confiance comme dans (5.25) donnerait pour procédures optimales les procédures minimisant l'erreur a posteriori (ou fréquentiste) maximale. Cette approche n'a cependant pas encore été traitée dans la littérature.

5.6 Exercices

Section 5.2.1

5.1 Dans le cadre de l'Exemple 5.4, étudier la modification de la probabilité a posteriori de H_0 lorsque $x = 0$ et τ/σ tend vers $+\infty$. Comparer à la réponse non informative associée à $\pi(\theta) = 1$.

Section 5.2.2

5.2 Soit $x \sim \mathcal{N}(\theta, 1)$. L'hypothèse à tester est $H_0 : |\theta| \leq c$ contre $H_1 : |\theta| > c$, avec $\pi(\theta) = 1$.

a. Tracer la courbe de la probabilité maximale de H_0 en fonction de c.

b. Déterminer les valeurs de c pour lesquelles ce maximum est 0.95 et le facteur de Bayes est 1. Ces valeurs sont-elles satisfaisantes ?

5.3 Un professeur doit donner un examen sur deux jours différents. Puisque les étudiants s'assoient les uns à côté des autres, il distribue deux sujets différents,

en alternance, afin de réduire les possibilités de tricherie. Il utilise la même technique et les mêmes sujets avec une autre classe le jour suivant. Les résultats sont : $n_{1A} = 17$ étudiants ont planché sur l'examen A le premier jour, $n_{2A} = 19$ le second jour, $n_{1B} = 15$ sur le sujet B le premier jour et $n_{2B} = 19$ le second jour. Les notes moyennes (sur 20) sont $\hat{\mu}_{1A} = 10.3$, $\hat{\mu}_{2A} = 10.9$, $\hat{\mu}_{1B} = 7.9$ et $\hat{\mu}_{2B} = 8.7$ et les écarts types sont $\hat{\sigma}_{1A} = 2.67$, $\hat{\sigma}_{2A} = 2.09$, $\hat{\sigma}_{1B} = 2.98$ et $\hat{\sigma}_{2B} = 2.91$.

a. Tester la présence d'un effet de classe, de sujet, ou d'un effet croisé classe-sujet en modélisant les résultats par une approche d'*analyse de la variance*, c'est-à-dire en supposant que chaque note d'étudiant x est distribuée selon une loi normale de moyenne $\mu_0+\mu_e+\mu_c$ et de variance σ^2_{ec} ($e = A, B$, $c = 1, 2$) avec $\mu_A + \mu_B = 0$, $\mu_1 + \mu_2 = 0$.

b. Un étudiant planchant sur le sujet A a oublié de rendre sa copie le premier jour. Est-il possible de détecter une tricherie le second jour ?

Section 5.2.3

5.4 (Pearl, 1988) Après que vous ayez fait part d'une rumeur à un voisin, celui-ci vous la répète quelques jours plus tard. Construire un modèle pour tester la possibilité que ce voisin ait entendu cette rumeur d'une autre personne.

5.5 *Soient deux observations indépendantes normales standard x et y. Les coordonnées polaires de (x, y) sont (r, θ), avec $x = r\cos\theta$ et $y = r\sin\theta$.

a. Pour $2r^2 = (x - y)^2 + (x + y)^2$ et étant donné que les variables $x - y$ et $x + y$ sont indépendantes, montrer que la distribution de r^2 sachant $x = y$ est $\mathscr{G}(1/2, 1)$.

b. Montrer que les variables r et θ sont indépendantes et en déduire que la distribution de r^2 sachant $\theta = \pi/4, 5\pi/4$ est $\mathscr{G}(1/2, 1/2)$.

c. Puisque $\{x = y\} = \{\theta = \pi/4, 5\pi/4\}$, expliquer ce paradoxe apparent, dit *paradoxe de Borel*, de deux distributions conditionnelles différentes pour un même événement. (*Indication :* Replacer le conditionnement dans une perspective de σ-algèbres et comparer les σ-algèbres engendrées par $x - y$ et par θ.)

Section 5.2.4

5.6 Pour $x \sim \mathscr{N}(\theta, 1)$ et $\theta \sim \mathscr{N}(0, \sigma^2)$, comparer les réponses bayésiennes pour les deux problèmes de test

$$H_0^1 : \ \theta = 0 \text{ contre } H_1^1 : \ \theta \neq 0,$$
$$H_0^2 : \ |\theta| \leq \epsilon \text{ contre } H_1^2 : \ |\theta| > \epsilon,$$

lorsque ϵ et σ varient.

5.7 Dans le cadre de l'Exemple 5.3, pour $x \sim \mathscr{B}(n, p)$ et le test de $H_0 : \ p = 1/2$, étudier comment varient les réponses bayésiennes en fonction de n pour $x = 0$, $x = n/2$ et la loi a priori de Jeffreys.

5.8 * (Berger et Delampady, 1987) Soit $x \sim \mathscr{N}(\theta, 1)$. Le but de cet exercice est de comparer $H_0 : \ |\theta - \theta_0| \leq \epsilon$ avec l'approximation $H_0^* : \ \theta = \theta_0$. On note g_0 et g_1 les densités a priori sur $\{|\theta - \theta_0| \leq \epsilon\}$ et $\{|\theta - \theta_0| > \epsilon\}$. Soit g une densité sur $\mathbb{R}$ telle que

$$g(\theta) \propto g_1(\theta) \quad \text{si} \quad |\theta - \theta_0| > \epsilon,$$

et telle que

$$\lambda = \int_{|\theta-\theta_0|\le\epsilon} g(\theta)\,d\theta,$$

soit suffisamment petit. On note

$$B = \frac{\int_{|\theta-\theta_0|\le\epsilon} f(x|\theta)g_0(\theta)\,d\theta}{\int_{|\theta-\theta_0|>\epsilon} f(x|\theta)g_1(\theta)\,d\theta} \quad \text{et} \quad \hat{B} = \frac{f(x|\theta_0)}{m_g(x)} = \frac{f(x|\theta)}{\int f(x|\theta)g(\theta)\,d\theta},$$

$t = (x - \theta_0)$ et

$$\gamma = \frac{1}{2\epsilon\varphi(t)}[\Phi(t+\epsilon) - \Phi(t-\epsilon)] - 1.$$

Montrer que, si $|t| \ge 1$, $\epsilon < |t| - 1$ et $\hat{B} \le (1+\gamma)^{-1}$, alors

$$B = \hat{B}(1+\varrho)$$

avec

$$-\lambda \le \frac{\lambda(\hat{B}-1)}{1-\lambda\hat{B}} \le \varrho \le \frac{\gamma + \lambda(1+\gamma)(\hat{B}-1)}{1-\lambda\hat{B}(1+\gamma)} \le \gamma.$$

Section 5.2.5

5.9 Soit $x \sim \mathscr{P}(\lambda)$. L'hypothèse à tester est $H_0 :\ \lambda \le 1$ contre $H_1 : \lambda > 1$. Donner la probabilité a posteriori de H_0 pour $x = 1$ et $\lambda \sim \mathscr{G}(\alpha, \beta)$.

a. Comment varie cette probabilité lorsque α et β tendent vers 0 ? Est-ce que la réponse à cette question dépend des taux de convergence de α et β vers 0 ?

b. Comparer aux probabilités associées à la loi non informative $\pi(\lambda) = 1/\lambda$. Est-il toujours possible d'utiliser cet a priori impropre ?

5.10 Soient $x \sim \mathscr{B}(n, p)$, $H_0 :\ p = 1/2$ et $H_1 :\ p \ne 1/2$. L'a priori $\pi(p)$ est une loi $\mathscr{B}e(\alpha, \alpha)$. Déterminer la limite de la probabilité a posteriori de H_0 lorsque $n = 10$, $x = 5$ et $n = 15$, $x = 7$ pour α tendant vers $+\infty$. Ces valeurs sont-elles intuitives ? Donner les probabilités a posteriori pour les lois a priori non informatives de Laplace, Jeffreys et Haldane.

5.11 Résoudre les Exercices 5.9 et 5.10 pour les facteurs de Bayes plutôt que les probabilités a posteriori.

5.12 Dans un cadre gaussien, déterminer s'il existe un problème de normalisation associé à des lois a priori non informatives pour des tests d'hypothèses unilatérales telles que

$$H_0 :\ \theta \in [0,1] \qquad \text{contre} \qquad H_1 :\ \theta > 1.$$

Remplacer $[0, 1]$ par $[0, \epsilon]$ et étudier les variations de la solution optimale lorsque ϵ tend vers 0.

5.13 Dans le test de

$$H_0 :\ |\theta| < \epsilon \qquad \text{contre} \qquad H_1 :\ |\theta| > \epsilon,$$

montrer que le facteur de Bayes tend vers le facteur de Bayes associé au test de

$$H_0 :\ \theta = 0 \qquad \text{contre} \qquad H_1 :\ \theta \ne 0\,,$$

quand ϵ tend vers 0. (*Indication :* On supposera que la règle de L'Hospital s'applique.)

Section 5.2.6

5.14 Établir la décomposition (5.6) à partir de la définition originale (5.5) de $B_{10}^{(\ell)}$. (*Indication :* Utiliser la formule de Bayes pour obtenir $\pi_1(\theta_1|x_{(\ell)})$ et $\pi_0(\theta_1|x_{(\ell)})$.)

5.15 Dans le cadre de l'Exemple 5.13, montrer comment $B_{10}^{(2)}$ dépend du choix de (x_1, x_2) en calculant les constantes de normalisation de $\pi_0(\sigma^2|x_1, x_2)$ et $\pi_1(\mu, \sigma^2|x_1, x_2)$ et en concluant le calcul d'intégrales dans $B_{10}^{(2)}$.

5.16 Aitkin (1991) suggère de contourner la difficulté liée aux lois a priori impropres en utilisant les données *deux fois* : pour $x \sim f(x|\theta)$, un a priori impropre π et une hypothèse à tester $H_0 : \theta = \theta_0$, prendre $\tilde{\pi}(\theta) = \pi(\theta|x)$ et utiliser $\tilde{\pi}$ comme a priori dans le facteur de Bayes.

a. Si $f(\cdot|\theta)$ est la densité de la loi $\mathcal{N}(\theta, 1)$ et $\pi(\theta) = 1$, calculer les pseudo-facteurs de Bayes correspondants.

b. Même question que a. lorsque $f(\cdot|\theta)$ est la fonction de probabilité de la loi $\mathcal{P}(\lambda)$ et $\pi(\lambda) = 1/\lambda$.

c. Analyser le comportement limite de ce pseudo-facteur de Bayes lorsque cette procédure est répétée, c'est-à-dire lorsque π est remplacé itérativement par $\tilde{\pi}$. [*Note :* D'un point de vue numérique, cette technique peut être utile pour le calcul d'estimateurs du maximum de vraisemblance et d'estimateurs MAP ; voir Robert et Casella, 1999, Section 5.2.4.]

5.17 Dans le cadre de l'Exemple 5.14, calculer le facteur de Bayes lorsque $\pi_1(\theta)$ est la densité de la loi $\mathcal{N}(0, 2)$ et comparer avec le facteur de Bayes intrinsèque arithmétique. (*Indication :* Calculer $\mathbb{E}[\exp(-x^2/2)]$.)

5.18 (Suite de l'Exercice 5.17) Pour le facteur de Bayes fractionnaire (5.11),

a. Montrer que la valeur minimale de b est $1/n$.

b. Montrer que (5.11) correspond à la loi a priori intrinsèque $\mathcal{N}(0, (1-b)/nb)$.

c. Montrer qu'une valeur fixée de b mène à une réduction de la variance vers 0 dans l'a priori intrinsèque.

d. Comparer les valeurs numériques des facteurs de Bayes intrinsèques arithmétique et fractionnaire.

e. Déterminer s'il existe une valeur de b telle que ces pseudo-facteurs de Bayes soient équivalents.

5.19 Dans le cadre de l'Exemple 5.15,

a. Montrer que π_2 s'intègre bien à 1.

b. Montrer que B_{10}^A correspond bien à un facteur de Bayes sous π_2.

5.20 Les conditions de cohérence pour les facteurs de Bayes sont données par

$$B_{12} = B_{10}B_{02} \qquad \text{et} \qquad B_{01} = 1/B_{10}\,,$$

lorsque trois hypothèses, H_0, H_1 et H_2, sont considérées avec, pour lois a priori respectives, π_0, π_1 et π_2.

a. Montrer que ces conditions sont satisfaites lorsque les π_i sont des lois a priori propres.

b. Montrer que les facteurs de Bayes fractionnaires satisfont $B_{01} = 1/B_{10}$ mais pas $B_{12} = B_{10}B_{02}$.

c. Montrer que ni les facteurs de Bayes arithmétiques ni les facteurs de Bayes géométriques intrinsèques ne satisfont ces conditions.

5.21 Pour la loi a priori intrinsèque considérée dans l'Exemple 5.17,

a. Montrer que

$$\int_{\theta_0}^{\infty} \left(2e^{\theta-\theta_0} - 1\right)^{-1} d\theta = \log(2).$$

(*Indication :* Utiliser un changement de variable de θ à $\omega = \exp(\theta - \theta_0)$ et une décomposition fractionnelle de $1/\omega(2\omega - 1)$.)

b. En déduire l'expression (5.12).

5.22 Dans le cadre de l'Exemple 5.19,

a. Montrer que

$$\int \left(\prod_{t=1}^{n} \left\{ \frac{p}{\sigma_1} e^{-(x_t-\mu_1)^2/2\sigma_1^2} + \frac{1-p}{\sigma_2} e^{-(x_t-\mu_2)^2/2\sigma_2^2} \right\}\right)^b d\pi(\mu,\sigma)$$
$$\geq \int \left(\prod_{t=1}^{n} \frac{p}{\sigma_1} e^{-(x_t-\mu_1)^2/2\sigma_1^2}\right)^b d\pi(\mu,\sigma)\,.$$

b. En déduire que le facteur de Bayes fractionnaire n'existe pas pour ce modèle.

5.23 Soient n observations $x_1, \ldots, x_n$ d'une loi de Student $\mathcal{T}(\nu, \mu, \sigma)$ et l'hypothèse nulle $H_0 : \mu = 0$.

a. Déterminer la taille d'échantillon d'apprentissage minimale pour les lois a priori $\pi_0(\sigma) = 1/\sigma$ et $\pi_1(\mu, \sigma) = 1/\sigma$.

b. Montrer que les facteurs de Bayes fractionnnaires ne peuvent pas être obtenus explicitement dans ce cas.

Section 5.3.1

5.24 Soient f et g deux fonctions réelles croissantes.

a. Montrer que

$$\mathbb{E}_\theta[f(X)g(X)] \geq \mathbb{E}_\theta[f(X)]\mathbb{E}_\theta[g(X)]$$

pour toute loi P_θ de x.

b. Utiliser a. pour montrer que, si $f(x|\theta)$ est une densité de rapport de vraisemblance monotone en $T(x)$, l'espérance $\mathbb{E}_\theta[g(T(x))]$ est une fonction croissante de θ. (*Indication :* Utiliser $g(x) = 1 - f(x|\theta')/f(x|\theta)$ et montrer que $\mathbb{E}_\theta[g(X)] = 0$.)

5.25 Montrer que les lois de Student et du χ^2 décentré sont à rapport de vraisemblance monotone.

Section 5.3.4

5.26 Pour la p-value $\tilde{p}$ définie dans l'Exemple 5.32, déterminer les valeurs de $\tilde{p}(x)$ pour $n = 15$ et comparer avec

$$p(x) = P_{1/2}\left[f(X|1/2) > f(x|1/2)\right].$$

5.27 (Johnson et Lindley, 1995) Soit une hypothèse nulle ponctuelle $H_0 : \theta = \theta_0$ telle que la p-value φ soit bien définie. La seule information disponible est que les données sont significatives au niveau α, donc que $\varphi(x) < \alpha$.

a. Donner le facteur de Bayes R_α de H_0 contre $H_1 : \theta \neq \theta_0$ lorsque les données sont significatives au niveau α, pour une loi a priori π arbitraire.
b. Étant donné un second niveau de significativité β tel que $\beta < \alpha$, on suppose $R_\alpha < R_\beta$. Établir une condition suffisante sur π pour que cette condition soit vérifiée.
c. Si $R_{\alpha|\beta}$ est le facteur de Bayes fondé sur l'information $\beta < \varphi(x) < \alpha$, montrer que $R_\alpha = \omega R_\beta + (1-\omega)R_{\alpha|\beta}$ et en déduire que $R_\beta > R_\alpha > R_{\alpha|\beta}$.
d. Dans le cas particulier où $\pi(\theta)$ est $\varrho_0 \mathbb{I}_{\theta_0}(\theta) + (1-\varrho_0)\mathcal{N}(\theta_0, \tau^2)$ et $x_1, \ldots, x_n \sim \mathcal{N}(\theta, \sigma^2)$, montrer que R_α converge vers $(1-\varrho_0)/\varrho_0\alpha$ lorsque n tend vers l'infini et $R_{\alpha|\beta}$ vers 0.

Section 5.3.5

5.28 Pour $x \sim \mathcal{N}(\theta, 1)$ et $H_0 : \theta = 0$, déterminer si les p-values prennent des valeurs inférieures à $\underline{P}(x, G_A)$ et $\underline{P}(x, G_S)$.

5.29 (Berger et Delampady, 1987) Soient $x \sim \mathcal{B}(n, p)$ et $H_0 : p = 1/2$. Pour la classe de lois a priori G_C formée des lois conjuguées de moyenne $1/2$, montrer que

$$\begin{aligned}\underline{P}(x, G_C) &= \inf_{g \in G_C} P(H_0|x) \\ &= \left[1 + \frac{1-\pi_0}{\pi_0} \sup_{c>0} \frac{\Gamma(c)\Gamma(x+c/2)\Gamma(n-x+c/2)}{\Gamma(c/2)^2\Gamma(n+c)}\right]^{-1}\end{aligned}$$

et établir la table de ces bornes inférieures et les p-values correspondantes pour $n = 10, 20, 30$ et x variant de 0 à $n/2$.

5.30 *(Casella et Berger, 1987) Établir le lemme suivant, utilisé dans le Lemme 5.35 et le Théorème 5.38 : *dans le cas où G est la famille des lois de mélange*

$$g(\theta) = \int_\Xi g_\xi(\theta) h(\xi)\, d\xi,$$

pour toute densité h sur Ξ, avec $g_\xi \in G_0$ et

$$G_0 = \{g_\xi;\ \xi \in \Xi\},$$

alors, pour tout f,

$$\sup_{g \in G} \int f(x|\theta) g(\theta)\, d\theta = \sup_{\xi \in \Xi} \int f(x|\theta) g_\xi(\theta)\, d\theta.$$

5.31 Dans le cas où $x \sim \mathcal{N}(\theta, 1)$ et $H_0 : \theta \leq 0$, déterminer la borne inférieure

$$\begin{aligned}\underline{P}(x, G_{SU}) &= \inf_{g \in G_{SU}} P^g(\theta \leq 0|x) \\ &= \inf_{g \in G_{SU}} \frac{\int_{-\infty}^0 f(x-\theta) g(\theta)\, d\theta}{\int_{-\infty}^{+\infty} f(x-\theta) g(\theta)\, d\theta}\end{aligned}$$

pour $x < 0$. Est-ce que la conclusion de Casella et Berger (1987) tient toujours ? Pouvez-vous expliquer pourquoi ?

5.32 *(Casella et Berger, 1987) Soit une fonction symétrique unimodale bornée g. La famille des *mélanges d'échelle de g* est définie par

$$G_g = \{\pi_\sigma;\ \pi_\sigma(\theta) = (1/\sigma)g(\theta/\sigma),\ \sigma > 0\}.$$

Si la densité des observations est $f(x-\theta)$, avec f symétrique en 0, et si elle vérifie la propriété de rapport de vraisemblance monotone, montrer que, pour $x > 0$,

$$\underline{P}(x, G_g) = p(x)$$

pour le test de $H_0 :\ \theta \leq 0$.

5.33 *(Casella et Berger, 1987) Soit le test de $H_0 :\ \theta \leq 0$ contre $H_1 :\ \theta > 0$ avec $x \sim f(x-\theta)$. Soient h et g des densités sur $]-\infty, 0]$ et $]0, +\infty[$.

a. Montrer que, si $\pi(\theta) = \varrho_0 h(\theta) + (1-\varrho_0)g(\theta)$,

$$\sup_h P^\pi(\theta \leq 0|x) = \frac{\varrho_0 f(x)}{\varrho_0 f(x) + (1-\varrho_0)\int_0^{+\infty} f(x-\theta)g(\theta)\,d\theta}$$

et en déduire que le suprémum favorise en fait H_0 en concentrant toute la masse à la frontière $\theta = 0$.

b. Si

$$\pi(\theta) = \varrho_0 h(\theta/\sigma_1)\frac{1}{\sigma_1} + (1-\varrho_0)g(\theta/\sigma_2)\frac{1}{\sigma_2}\,,$$

montrer que, lorsque σ_1 est fixé,

$$\lim_{\sigma_2 \to \infty} P^\pi(\theta \leq 0|x) = 1$$

et que, lorsque σ_2 est fixé,

$$\lim_{\sigma_1 \to \infty} P^\pi(\theta \leq 0|x) = 0.$$

5.34 *(Caron, 1994) Afin de répondre aux critiques à l'égard des hypothèses nulles ponctuelles, $H_0 : \theta = \theta_0$, la formulation de l'hypothèse nulle peut être modifiée pour tenir compte de la loi a priori. Par exemple, pour une loi a priori donnée π sur Θ admettant un mode en θ_0 mais n'attribuant pas de poids a priori à θ_0, on peut proposer l'hypothèse transformée $H_0^\pi : \pi(\theta) > k^\pi$, de façon telle que la taille de la région HPD soit déterminée par la condition "objective" $\pi(\pi(\theta) > k^\pi) = 0.5$. Considérons le cas $x \sim \mathcal{N}(\theta, 1)$ et $\theta_0 = 0$.

a. Lorsque π appartient à la famille des lois $\mathcal{N}(0, \sigma^2)$, déterminer k^π et calculer la borne inférieure des réponses bayésiennes pour cette famille. Comparer avec les probabilités a posteriori de Berger et Sellke (1987) pour les valeurs d'intérêt.

b. Déterminer si le paradoxe de Jeffreys-Lindley a lieu dans ce cas.

c. Pour les familles alternatives $\mathcal{U}_{[-c,c]}$ $(c > 0)$ et $\pi(\theta|\lambda) \propto \exp(-\lambda|\theta|)$ $(\lambda > 0)$, calculer les bornes inférieures correspondantes.

5.35 ***(Suite de l'Exercice 5.34)** Considérons le cas $x \sim \mathcal{C}(\theta, 1)$ pour $H_0 : \theta = 0$.

a. Pour l'approche de Berger et Sellke (1987), montrer que la probabilité a posteriori de H_0 lorsque π_c est $\mathcal{U}_{[-c,c]}$ vaut

$$\pi_c(H_0|x) = \left[1 + (1+x^2)(\arctan(c-x) + \arctan(c+x))/2c\right]^{-1}.$$

b. Pour l'approche développée dans l'exercice précédent, montrer que la probabilité correspondante est

$$\pi_c(H_0^\pi|x) = \frac{\arctan(c/2-x)+\arctan(c/2+x))}{\arctan(c-x)+\arctan(c+x))}.$$

c. Calculer et comparer les bornes inférieures pour les deux approches.
d. Montrer que

$$\lim_{x\to\infty} \frac{\inf_c \pi_c(H_0^\pi|x)}{\inf_c \pi_c(H_0|x)} = \frac{2}{3}.$$

Section 5.4

5.36 (Hwang *et al.*, 1992) Montrer que, pour la fonction de coût (5.19), les p-values définies dans l'Exemple 5.45 sont effectivement admissibles. (*Indication :* Montrer que les risques de Bayes sont finis.)

5.37 (Hwang *et al.*, 1992) Le but de cet exercice est de montrer que, pour le test bilatéral (5.20), la p-value $p(x)$ peut prendre la valeur 1. (*Indication :* On rappelle que le test UPPS est de la forme

$$\varphi(x) = \begin{cases} 0 & \text{si } T(x) < c_0 \text{ ou } T(x) > c_1, \\ 1 & \text{sinon,} \end{cases}$$

dans ce cadre, avec $c_0 = c_0(\alpha)$ et $c_1 = c_1(\alpha)$.)
a. Soient $\theta_1 \neq \theta_2$ et

$$c^* = \inf\{T(x);\ f(x|\theta_2) > f(x|\theta_1)\}.$$

Montrer que $c^* \in [c_0(\alpha), c_1(\alpha)]$ pour tout $0 < \alpha < 1$.
b. On suppose $\theta_1 = \theta_2$. Appliquer le résultat précédent à

$$f(x|\theta^*) = \mathbb{E}_{\theta_1}[T(x)]f(x|\theta_1), \qquad f(x|\theta^{**}) = T(x)f(x|\theta_1),$$

et conclure.

5.38 (Hwang *et al.*, 1992) Dans un cadre gaussien, considérer l'hypothèse nulle ponctuelle $H_0 :\ \theta = 0$. Montrer que, sous la fonction de coût (5.19), la p-value ne peut pas être dominée par une probabilité a posteriori propre. (*Indication :* Démontrer d'abord que, pour tout a et ϵ,

$$\frac{P_\theta(a < |x| < a+\epsilon)}{P_\theta(|x| < a)} \to +\infty$$

lorsque θ tend vers l'infini.)

5.39 (Hwang *et al.*, 1992) Pour la fonction de coût (5.19), montrer que $\varphi(x) = 1/2$ *est l'unique estimateur minimax*. Étendre ce résultat à toutes les fonctions de coût convexes. Dans ce cadre, existe-t-il des lois les moins favorables ?

5.40 (Robert et Casella, 1994) Une modification possible de la fonction de coût (5.18) est d'introduire une pondération fondée sur une distance, afin de pénaliser d'une façon différente les erreurs proches de la frontière entre H_0 et H_1 de celles qui en sont loin.

a. Si l'hypothèse nulle est $H_0 : \theta \leq \theta_0$ pour $x \sim \mathcal{N}(\theta, 1)$ et la fonction de coût est

$$\mathrm{L}(\theta, \varphi) = (\theta - \theta_0)^2 (\mathbb{I}_{H_0}(\theta) - \varphi)^2,$$

donner l'expression générale des estimateurs de Bayes.

b. Si $\pi(\theta) = 1$, montrer que l'estimateur de Bayes est plus petit que la p-value si $x > \theta_0$ et plus grand si $x < \theta_0$.

5.41 (Robert et Casella, 1994) D'un point de vue de choix de modèle, la fonction de perte incorpore les conséquences d'une acceptation ou d'un rejet de l'hypothèse nulle $H_0 : \theta = \theta_0$ en termes d'estimation.

a. Pour la fonction de coût

$$\mathrm{L}_1(\theta, (\varphi, \delta)) = d(\theta - \delta)|1 - \varphi| + d(\theta_0 - \theta)|\varphi|,$$

montrer que les estimateurs de Bayes sont $(0, \delta^\pi(x))$ où $\delta^\pi(x)$ est l'estimateur de Bayes régulier de θ sous $d(\theta - \delta)$ pour tout d et π.

b. Pour la fonction de coût

$$\mathrm{L}_2(\theta, (\varphi, \delta)) = d(\theta - \delta)|1 - \varphi| + d(\theta_0 - \delta)|\varphi|,$$

montrer que la règle de Bayes est $(1, \theta_0)$ pour tout π et d.

c. Pour la fonction de coût

$$\mathrm{L}_3(\theta, (\varphi, \delta)) = (\delta - \theta)^2 (\mathbb{I}_{H_0}(\theta) - \varphi)^2,$$

montrer que la règle de Bayes associée est $(0, \theta_0)$, c'est-à-dire que cette règle rejette systématiquement l'hypothèse nulle $H_0 : \theta = \theta_0$, mais utilise toujours θ_0 comme estimateur de θ.

d. Étudier les procédures bayésiennes sous le coût modifié

$$\mathrm{L}_4(\theta, (\varphi, \delta)) = \left[1 + (\delta - \theta)^2\right] \left[1 + (\mathbb{I}_{H_0}(\theta) - \varphi)^2\right],$$

afin d'établir si elles sont moins paradoxales.

e. Montrer que la fonction de perte

$$\mathrm{L}_5(\theta, (\varphi, \delta)) = \xi(\delta - \theta)^2|1 - \varphi| + \{(\delta - \theta_0)^2 + (\theta - \theta_0)^2\}|\varphi|,$$

fournit une procédure de pré-test bayésien raisonnable qui évite les paradoxes de L_1, L_2 et L_3 si et seulement si $\xi > 1$.

Section 5.5.1

5.42 Soient deux observations indépendantes x_1, x_2 tirées d'une loi de Cauchy $\mathscr{C}(\theta, 1)$. Pour $\pi(\theta) = 1$, donner la forme de la région HPD α-crédible. Quelle autre région de niveau α plus convaincante pourriez-vous proposer ?

5.43 Donner la région α-crédible pour $x \sim \mathscr{P}(\lambda)$ et $\lambda \sim \mathscr{G}(\delta, \beta)$. Etudier l'évolution de cette région en fonction de δ et β. Examiner le cas particulier de la loi non informative.

5.44 *Cet exercice traite d'une alternative aux régions α-crédibles. *Le meilleur centre bayésien au niveau* α est l'estimateur $\delta^\pi_\alpha(x)$, qui est le centre de la boule de plus petit rayon et de couverture $1 - \alpha$, c'est-à-dire

$$P^\pi(||\theta - \delta^\pi_\alpha(x)|| < k|x) = \sup_\delta P^\pi(||\theta - \delta(x)|| < k|x) = 1 - \alpha.$$

a. Montrer que, si la loi a posteriori est à symétrie sphérique et unimodale, la région correspondante est HPD.
b. Soient $x \sim \mathcal{N}(\theta, 1)$, $\theta \sim \mathcal{N}(0, \tau^2)$ et $\pi(\tau^2) = 1/\tau^{3/2}$. Déterminer la loi a posteriori. Montrer que la densité correspondante est unimodale lorsque $0 < x^2 < 2$ et bimodale sinon, de second mode

$$\delta(x) = \left(1 - \frac{1 - \sqrt{1 - (2/x^2)}}{2}\right) x.$$

Calculer le meilleur centre de Bayes et montrer que, si α est suffisamment grand, δ_α^π n'est pas continu et proche de

$$\phi(x) = \left(1 - \frac{1}{2x^2}\right)^+ x,$$

c'est-à-dire que cet estimateur de Bayes reproduit l'estimateur de James-Stein.
c. Généraliser b. pour $\pi(\tau^2) = \tau^{-v}$.
d. Montrer que le meilleur centre de Bayes associé à un a priori propre π est admissible sous le coût

$$\mathrm{L}(\theta, \delta) = \mathbb{I}_{(k,+\infty)}(||\theta - \delta||^2).$$

5.45 *(Thatcher, 1964) Soit $x \sim \mathscr{B}(n, \theta)$. Pour $0 < \alpha < 1$ et l'a priori π sur θ, on définit θ_x^π par $P^\pi(\theta \leq \theta_x^\pi | x) = \alpha$.
a. Si $\pi(\theta) = (1-\theta)^{-1}$, montrer que $P_\theta(\theta \leq \theta_x^\pi) \leq \alpha$ pour $\theta > 0$.
b. Si $\pi(\theta) = \theta^{-1}$, montrer que $P_\theta(\theta \leq \theta_x^\pi) \geq \alpha$ pour $\theta < 1$.
c. Définir θ_x^λ associé à $\pi(\theta) = \theta^{\lambda-1}(1-\theta)^{-\lambda}$, $0 \leq \lambda \leq 1$. Montrer que θ_x^λ croît en λ et en déduire que

$$\lim_{\theta \uparrow \theta_x^\lambda} P_\theta(\theta \leq \theta_x^\lambda) \geq \alpha \geq \lim_{\theta \downarrow \theta_x^\lambda} P_\theta(\theta \leq \theta_x^\lambda).$$

5.46 *(Hartigan, 1983) Soit $x \sim \mathscr{P}(\lambda)$. Pour $0 < \alpha < 1$ et l'a priori π sur λ, on définit λ_x^π par

$$P^\pi(0 \leq \lambda \leq \lambda_x^\pi | x) = \alpha.$$

a. Montrer que, si $\pi(\lambda) = 1/\lambda$, $P_\lambda(\lambda \leq \lambda_x^\pi) \leq \alpha$ pour tout λ.
b. Montrer que, si $\pi(\lambda) = 1$, $P_\lambda(\lambda \leq \lambda_x^\pi) \geq \alpha$ pour tout λ. (*Indication :* Utiliser la relation suivante :

$$\sum_{x=x_0}^{\infty} e^{-\lambda} \frac{\lambda^x}{x!} = \int_0^\infty \frac{u^{x_0-1}}{(x_0-1)!} e^{-u} du.\Big)$$

5.47 Un problème célèbre en Statistique classique est celui de *Behrens-Fisher*. Il découle de la simple situation de deux populations normales de moyennes et variances inconnues ; il n'existe pas en effet de test UPP ou UPPS pour comparer les moyennes dans ce cas. Soient $x_1, \ldots, x_n$ un échantillon tiré de $\mathcal{N}(\theta, \sigma^2)$ et $y_1, \ldots, y_m$ un échantillon de $\mathcal{N}(\mu, \tau^2)$, où $\theta, \mu, \tau, \sigma$ sont inconnus.
a. *Montrer qu'il n'existe pas de test UPPS pour l'hypothèse $H_0 : \theta = \mu$. (*Indication :* Conditionner en s_x^2 et s_y^2, définis ci-dessous, afin de montrer que les procédures UPPS varient avec s_x^2 et s_y^2.)

b. Expliquer pourquoi un test raisonnable devrait dépendre de la quantité pivotale

$$T = \frac{(\theta - \mu) - (\bar{x} - \bar{y})}{\sqrt{s_x^2/n + s_y^2/m}}$$

avec $\bar{x} = \sum_i x_i/n$, $\bar{y} = \sum_j y_j/m$, $s_x^2 = \sum_i (x_i - \bar{x})^2/n - 1$ et $s_y^2 = \sum_j (y_j - \bar{y})^2/m - 1$.

c. Montrer que la distribution de T dépend de σ/τ même quand $\theta = \mu$ et qu'il *ne s'agit pas* d'une loi de Student.

d. Donner la loi a posteriori de T pour $\pi(\theta, \mu, \sigma, \tau) = 1/\sigma^2\tau^2$ et montrer qu'elle ne dépend que de $(s_x/\sqrt{n})(s_y/\sqrt{m})$. [*Note :* Voir Robinson, 1982, pour une revue détaillée des différents points reliés à ce problème.]

Section 5.5.2

5.48 (Casella et Berger, 2001) Soient $x \sim \mathcal{N}(\mu, 1)$ et

$$C_a(x) = \{\mu;\ \min(0, x - a) \leq \mu \leq \max(0, x + a)\}.$$

a. On pose $a = 1.645$. Montrer que C_a est un intervalle de confiance à 95% tel que

$$P_0(0 \in C_a(x)) = 1.$$

b. Pour $\pi(\mu) = 1$ et $a = 1.645$, montrer que C_a est aussi une région 0.1-crédible et que

$$P^\pi(\mu \in C_a(x)|x) = 0.90$$

si $|x| \leq 1.645$ et

$$\lim_{|x| \to +\infty} P^\pi(\mu \in C_a(x)|x) = 1.$$

5.49 Soient $x \sim f(x|\theta)$ avec $\theta \in \mathbb{R}$ et π loi a priori sur θ. Si on définit l'ensemble α-crédible $(-\infty, \theta_x)$ par $P^\pi(\theta \geq \theta_x|x) = \alpha$, montrer que cet intervalle unilatéral ne peut pas être de niveau α au sens fréquentiste. (*Indication :* Montrer que $P(\theta \geq \theta_x|\theta \leq \theta_0) > \alpha$ pour une certaine valeur θ_0.)

5.50 *(Fieller, 1954) Dans un cadre de *calibration* (voir l'Exercice 4.48), les intervalles de confiance doivent avoir une longueur infinie pour maintenir un niveau de confiance fixé, comme le montrent Gleser et Hwang (1987). Soit $(x_1, y_1), \ldots, (x_n, y_n)$ un échantillon tiré de $\mathcal{N}_2(\mu, \Sigma)$. Le paramètre d'intérêt est θ, le rapport des espérances μ_x/μ_y.

a. Définir $\bar{z}_\theta = \bar{y} - \theta\bar{x}$. Montrer que

$$\bar{z}_\theta \sim \mathcal{N}\left(0, \frac{1}{n}(\sigma_y^2 - 2\theta\sigma_{xy} + \theta^2\sigma_x^2)\right)$$

et que

$$\hat{v}_\theta = \frac{1}{n-1}(s_y^2 - 2\theta s_{xy} + \theta^2 s_x)$$

est un estimateur sans biais de v_θ, la variance de $\bar{z}_\theta$, où $\bar{x}$, $\bar{y}$, s_x^2, s_{xy} et s_y^2 sont les moments empiriques usuels et

$$\Sigma = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}.$$

b. Montrer que $\bar{z}_\theta$ et $\hat{v}_\theta$ sont indépendants et que $(n-1)\hat{v}_\theta/v_\theta \sim \chi^2_{n-1}$. En déduire que $\{\theta;\ \bar{z}_\theta/\hat{v}_\theta \leq t^2_{n-1,\alpha/2}\}$ définit un ensemble de confiance à $(1-\alpha)$.
c. Montrer que cet ensemble de confiance dépend d'une parabole en θ et peut être un intervalle, le complément d'un intervalle ou l'ensemble des nombres réels.

Section 5.5.3

5.51 *La domination de l'estimateur usuel en tant que centre d'une région de confiance ne découle pas forcément de la domination correspondante pour le coût quadratique. Montrer que, dans le cas gaussien, si

$$\delta_a^{\mathrm{JS}}(x) = \left(1 - \frac{a}{||x||^2}\right) x,$$

la région de confiance recentrée

$$C_a^{\mathrm{JS}}(x) = \{\theta;\ ||\theta - \delta_a^{\mathrm{JS}}(x)||^2 \leq c_\alpha\},$$

ne domine pas la région de confiance usuelle, même si δ_a^{JS} domine δ_0 lorsque $a \leq 2(p-2)$. (*Indication :* Considérer $\theta = 0$.)

5.52 (Casella *et al.*, 1993a) Montrer que la fonction de coût rationnel donnée en Section 5.5,

$$\mathrm{L}(\theta, C) = \frac{\mathrm{vol}(C)}{k + \mathrm{vol}(C)} - \mathbb{I}_C(\theta),$$

ne mène pas au paradoxe de Berger dans le cas gaussien.

5.53 *(Casella *et al.*, 1993a) Soit une fonction de coût générale de la forme

$$\mathrm{L}(\theta, C) = S(\mathrm{vol}(C)) - \mathbb{I}_C(\theta),$$

où S est croissante et $0 \leq S(t) \leq 1$.
a. Montrer que les estimateurs de Bayes sont des régions HPD.
b. Montrer que, si $x \sim \mathscr{N}_p(\theta, I_p)$ et $\theta \sim \mathscr{N}_p(\mu, \tau^2 I_p)$, les ensembles crédibles bayésiens C^π ne sont pas vides si $S(t) = t/(a+t)$.
c. Déterminer le rayon minimal de C^π lorsque τ varie.
d. Soient $\bar{x} \sim \mathscr{N}(\theta, \sigma^2/n)$ et $s^2 \sim \sigma^2\chi^2_q$. Sous le coût rationnel, montrer que

$$C^\pi(\bar{x}, s^2) = \left\{\theta;\ |\theta - \bar{x}| \leq \frac{t^* s}{\sqrt{n}}\right\},$$

où t^* est la solution de

$$\min_t \left(\frac{2ts/\sqrt{n}}{a + 2ts/\sqrt{n}} - P(|T_{n-1}| < t)\right).$$

En déduire que $P(|T_{n-1}| < t^*(s)|s) \geq 1/2$.

5.54 (Walley, 1991) Soit la loi double-exponentielle, $f(x|\theta) = (1/2)\exp(-|x-\theta|)$.
a. Montrer que $C_x =]-\infty, x]$ est un intervalle de confiance à 50%.
b. Montrer que $P_\theta(\theta \in C_x | x < 0) < 0.5$ pour tout θ.
c. Soit $\varphi(x) = (e^{2x}/2)\mathbb{I}_{x<0}$. Montrer que

$$\mathbb{E}_\theta[\mathbb{I}_{x<0}(\mathbb{I}_{C_x}(\theta) - 1/2) + \varphi(x)] \geq 0$$

et en déduire que $\gamma(x) = 1/2$ n'est pas un estimateur de confiance admissible sous le coût quadratique pour C_x.

Note 5.7.3

5.55 *(Brown, 1967) Dans le cadre de l'Exemple 5.55, montrer que

$$P\left(\sqrt{n}|\bar{x}-\theta| \le ks | s \le 1\right) \le \alpha > P\left(\sqrt{n}|\bar{x}-\theta| \le ks | s > 1\right)$$

et calculer un sous-ensemble positivement pertinent. (*Indication :* Montrer que

$$P\left(\sqrt{n}|\bar{x}-\theta| \le ks | s\right)$$

est croissant en s.)

5.56 (Walley, 1991) Soit un échantillon $x_1, \ldots, x_n$ tiré de $\mathscr{U}_{[\theta,\theta+1]}$.

a. Montrer que les intervalles de confiance unilatéraux uniformément plus précis sont de la forme $C_x = [(x_{(1)} + 1 - K) \wedge (x_{(n)} - 1), x_{(1)} + 1]$ et vérifier que le niveau de confiance est $\gamma = 1 - (1 - K/2)^n$.

b. Pour $n = 1$ et $\gamma = 1/2$, montrer que $C_x = [x, x+1]$. Considérer une fonction bornée strictement décroissante f et poser $\varphi(x) = (f(x) - f(x+1)) \wedge (f(x-1) - f(x))$. Vérifier que

$$\mathbb{E}_\theta[f(\mathbb{I}_{C_x}(\theta) - 0.5)] = 0.25 \int_\theta^{\theta+1} (f(x-1) - f(x))\, dx$$

et

$$\mathbb{E}_\theta[\varphi(x)] \le \frac{1}{8} \int_\theta^{\theta+1} (f(x-1) - f(x))\, dx.$$

c. En déduire que

$$\mathbb{E}_\theta[f(\mathbb{I}_{C_x}(\theta) - 0.5) - \varphi(x)] \ge 0$$

pour tout θ et que $\gamma = 1/2$ n'est pas un estimateur admissible.

d. On définit, pour $n \ge 2$,

$$B = \{(x_1, \ldots, x_n);\ x_{(n)} - x_{(1)} \ge 2 - K\}.$$

Montrer que

$$P_\theta(\theta \in C(x_1, \ldots, x_n) | (x_1, \ldots, x_n) \in B) = 1$$

et conclure que B est un sous-ensemble pertinent.

Note 5.7.4

5.57 (Berger *et al.*, 1998) Pour l'estimateur de confiance $\gamma(x)$ donné en (5.26), montrer que

$$\gamma(x) = \frac{s}{1+s} \quad \text{si} \quad s < r, \qquad \gamma(x) = \frac{1}{1+s} \quad \text{si} \quad s > a\,.$$

5.58 Montrer que, dans le cadre de l'Exemple 5.59, $\Psi(1) > 1$ et donner le facteur de Bayes en faveur de H_0.

5.59 (Lindley, 1990) Considérant une troisième décision -1 dans un problème de test, soit l'extension suivante de la fonction de coût $0-1$:

$$\mathrm{L}(\theta, \varphi) = \begin{cases} \ell_i & \text{si } \varphi = 1 - i \text{ et } H_i \text{ est vraie,} \\ m_i & \text{si } \varphi = -1 \text{ et } H_i \text{ est vraie.} \end{cases}$$

Calculer les coûts a posteriori et montrer que $\varphi = -1$ si

$$\frac{m_1 \varrho}{\ell_0 - m_0} < B_{10}(x) < \frac{(\ell_1 - m_1)\varrho}{m_0}\,,$$

où ϱ est le rapport des chances a priori, soit π_1/π_0.

5.60 (Lindley, 1990) Montrer que la statistique $S(x)$ donnée en (5.27) n'est pas libre, sauf lorsque

$$\tau(t) + \varrho = 1 + \frac{\varrho}{t}, \qquad t > c\,,$$

où c est défini par $F_0(c) = 1 - \varrho F_1(c)$ et $\tau(t)$ est donné par $F_0(t) = 1 - \varrho F_1(\tau(t))$. Montrer que cette propriété est vérifiée lorsque $B_{10}(x)$ a la même loi sous m_1 que $B_{01}(x)$ sous m_0. [*Note :* Voir Berger *et al.*, 1994, p. 1798.]

5.7 Notes

5.7.1 P-values et décisions bayésiennes

Une critique radicale de la comparaison de la Section 5.3.5 est qu'elle n'a en fait aucun sens : ces deux types de réponses sont différentes conceptuellement et des p-values ne sont pas des probabilités. La réponse à cette critique est que, au-delà du fait qu'elles sont utilisées *comme des probabilités* en pratique, les p-values, d'un point de vue décisionnel, tentent de répondre au même problème inférentiel que les probabilités a posteriori. Il est donc sensé de les comparer.

Considérons la fonction de coût $a_0 - a_1$, comme dans (5.1). Le *test minimax* UPPS est alors

$$\varphi(x) = \begin{cases} 1 & \text{si } p(x) > \frac{a_1}{a_0+a_1}, \\ 0 & \text{sinon.} \end{cases}$$

En fait, lorsque les fonctions de puissance sont continues et les hypothèses sont contiguës (voir Lehmann, 1986, Chapitre 4), un test UPPS vérifie

$$\sup_{\Theta_0} P_\theta(\varphi(x) = 0) = \alpha = \inf_{\Theta_1} P_\theta(\varphi(x) = 0) = 1 - \sup_{\Theta_1} P_\theta(\varphi(x) = 1).$$

De plus, lorsque φ est minimax sous cette fonction de coût, il satisfait

$$\begin{aligned} \sup_{\Theta_0} R(\theta, \varphi) &= a_0 \sup_{\Theta_0} P_\theta(\varphi(x) = 0) \\ &= \sup_{\Theta_1} R(\theta, \varphi) = a_1 \sup_{\Theta_1} P_\theta(\varphi(x) = 1). \end{aligned}$$

Donc, sous certaines conditions de régularité, satisfaites, par exemple, par des familles exponentielles, φ est tel que

$$\sup_{\Theta_0} P_\theta(\varphi = 0) = \frac{a_1}{a_1 + a_0}.$$

Il découle alors de la Proposition 5.2 qu'il est légitime de comparer la p-value $p(x)$ à des probabilités a posteriori, puisque la procédure de décision bayésienne est donnée par

$$\gamma^\pi(x) = \begin{cases} 1 & \text{si } P^\pi(\theta \in \Theta_0 | x) > \frac{a_0}{a_0+a_1}, \\ 0 & \text{sinon} \end{cases}$$

et les deux approches comparent une évaluation continue (p-value ou probabilité a posteriori) à la même borne.

5.7.2 Probabilités a priori inégales

Une autre critique à l'égard de l'évaluation des bornes de la Section 5.3.5, avancée, par exemple, par Casella et Berger (1987), est que cette borne inférieure n'est pas calculée sur l'ensemble de *toutes* les lois a priori, puisque n'est considérée que la probabilité a priori $\varrho_0 = 1/2$. Bien entendu, si ϱ_0 peut aussi être modifié, il est toujours possible de trouver une réponse bayésienne plus petite que la p-value, puisque la borne inférieure sur toutes les réponses bayésiennes est alors 0 pour tout x (ce qui correspond au cas $\varrho_0 = 0$). À l'inverse, pour une valeur fixée de $\varrho_0 \neq 0$, il y a toujours des valeurs de x pour lesquelles la borne inférieure sur les probabilités a posteriori est plus grande que la p-value.

Une version plus sophistiquée de cette critique est de considérer que le poids $\varrho_0 = 1/2$ n'est pas nécessairement la probabilité la plus objective et qu'elle devrait être déterminée en fonction de l'a priori π choisi. En fait, comme nous l'avons mentionné ci-dessus, les lois a priori de la forme $\pi(\theta) = \varrho_0 \mathbb{I}_{\theta_0}(\theta) + (1 - \varrho_0)\pi_1(\theta)$ sont assez artificielles. Même si de telles lois a priori sont nécessaires à la résolution du problème de test, il est plus naturel de penser π comme une modification de l'a priori original π_1, à la lumière de ce problème. Le problème inférentiel, c'est-à-dire le fait qu'on s'intéresse particulièrement à θ_0, contient une certaine information résiduelle suffisante pour justifier une modification de la loi a priori (sinon, la question du test devrait elle-même être modifiée pour devenir compatible avec l'information a priori). Il est donc sensé d'imposer que le poids ϱ_0 dépende de π_1. (Ce point sera repris dans le Chapitre 7 sur le choix de modèles, pour le cas des modèles imbriqués : le modèle le plus général, c'est-à-dire celui qui contient tous les autres, devrait être plus probable que les autres.)

Exemple 5.56. (Suite de l'Exemple 5.34) Puisqu'il s'agit de tester H_0 : $\theta = 0$, la probabilité a priori de H_0 est nulle pour toute densité a priori continue π_1. Cependant, il est raisonnable d'imposer que H_0 ait une probabilité a priori plus élevée si π_1 est $\mathcal{N}(0, 1)$ que si π_1 est $\mathcal{N}(0, 10)$, puisque tout voisinage de 0 est moins probable sous la deuxième loi a priori. Voilà pourquoi le paradoxe de Jeffreys-Lindley est bien un "paradoxe" : l'accroissement des probabilités du Tableau 5.2 au Tableau 5.3 semble contre-intuitive. ‖

Malheureusement, une détermination du poids ϱ_0 comme fonction de π_1 prête à controverse et nous nous contentons de mentionner brièvement une solution proposée dans Robert et Caron (1996) (voir Spiegelhalter et Smith, 1980, pour une autre approche fondée sur des observations virtuelles les plus favorables). L'idée sous-jacente est que le poids ϱ_0 devrait satisfaire

$$(1 - \varrho_0)\pi_1(\theta_0) = \varrho_0,$$

afin que θ_0 soit pondéré de la même façon sous les deux hypothèses. Bien entendu, cela revient à comparer un poids sous une masse de Dirac en 0, ϱ_0, à un poids instantané relativement à la mesure de Lebesgue, $(1 - \varrho_0)\pi_1(\theta_0)$, et

la comparaison n'est pas justifiée mathématiquement parlant (puisque la valeur que prend la densité π_1 en un point tel que θ_0 est arbitraire). De plus, l'équation ci-dessus n'admet pas toujours de solution.

Exemple 5.57. (Suite de l'Exemple 5.25) Lorsque $\pi_1(\theta)$ est une loi a priori gaussienne $\mathcal{N}(0, n)$, l'égalité ci-dessus donne comme expression du poids

$$\varrho_0 = \frac{\pi_1(0)}{1 + \pi_1(0)} = \frac{1}{1 + \sqrt{2\pi n}},$$

et la probabilité a posteriori de H_0 est alors

$$\begin{aligned}\left(1 + \frac{1-\varrho_0}{\varrho_0}\frac{m_1(x)}{\varphi(x)}\right)^{-1} &= \left(1 + \sqrt{2\pi\frac{n}{n+1}}e^{x^2/2 - x^2/2(n+1)}\right)^{-1} \\ &= \left(1 + \sqrt{\frac{2\pi n}{n+1}}e^{\frac{n}{2(n+1)}x^2}\right)^{-1}.\end{aligned}$$

Notons que cette approche évite le paradoxe de Jeffreys-Lindley, puisque la probabilité limite (pour n tendant vers $+\infty$) est

$$\left(1 + \sqrt{2\pi}e^{x^2/2}\right)^{-1}.$$

Cette valeur se trouve aussi être la probabilité a posteriori associée à la densité a priori de Lebesgue, $\pi(\theta) = 1$. ‖

5.7.3 Évaluation conditionnelle des régions de confiance

Une évaluation critique des régions de confiance de Neyman-Pearson (et plus généralement des procédures fréquentistes) dérive de l'analyse *conditionnelle* de Kiefer (1977) et Robinson (1979). Lehmann (1986, Chapitre 10) donne une description de cette approche (voir aussi Buehler, 1959, Pierce, 1973, Casella, 1987, 1992, Maatta et Casella, 1990, et Goutis et Casella, 1991, 1992). Ces travaux démontrent que des procédures classiques de construction de régions de confiance sont souvent sous-optimales lorsqu'elles sont considérées d'un point de vue conditionnel.

Définition 5.58. *Soit C_x, une région de confiance de niveau α. Un ensemble $A \subset \mathcal{X}$ est dit* sous-ensemble pertinent biaisé négativement *pour la région de confiance C_x s'il existe $\epsilon > 0$ tel que*

$$P_\theta(\theta \in C_x | x \in A) \leq 1 - \alpha - \epsilon$$

pour tout $\theta \in \Theta$.

On peut définir de même des *sous-ensembles pertinents biaisés positivement*. Cette notion est généralisée par Robinson (1979) à celle de *procédures de paris pertinentes*. L'existence de tels ensembles remet en cause le concept même de niveau de confiance α, puisque, selon l'ensemble de conditionnement, la probabilité de couverture varie et peut même tomber sous le niveau de confiance

nominal minimal. Bien entendu, cette critique peut s'étendre aux procédures de test par un argument de dualité.

Dans le cadre de l'Exemple 5.54 et pour des tests de Student, Brown (1967) établit l'existence d'ensembles pertinents biaisés positivement de la forme $\{|x| < k\}$; ce qui implique

$$P_\theta(\theta \in C_x | |x| > k) \leq 1 - \alpha$$

(voir aussi l'Exercice 5.55). De tels phénomènes ont mené Kiefer (1977) à suggérer de partitionner l'espace d'échantillonnage $\mathscr{X}$ et d'allouer à chaque sous-ensemble de la partition un niveau de confiance différent (voir aussi Brown, 1978). Suivant l'analyse de Fisher, il a proposé que ces sous-ensembles soient indexés par une statistique libre. Par exemple, la statistique libre adéquate pour l'Exemple 2.9 est $x_1 - x_2$.

Malheureusement, le choix d'une statistique libre modifie dans la plupart des cas la région de confiance obtenue ; Berger et Wolpert (1988) donnent un exemple où des statistiques libres différentes produisent des résultats différents, ce qui est incompatible avec le principe de vraisemblance. Nous considérons que, fondamentalement, le problème de l'existence d'ensembles biaisés pertinents n'est pas lié à la région de confiance C_x même, mais plutôt au niveau de confiance α, qu'il faudrait remplacer par un niveau plus adaptatif (ou plus conditionnel) $\alpha(x)$ (voir la Section 4.2). En fait, l'existence de procédures de paris pertinentes est équivalente à la domination de l'estimateur de confiance constant sous le coût quadratique (Robinson, 1979).

5.7.4 Perspective de réconciliation

Alors que la Section 5.3 a montré que les réponses fréquentistes, c'est-à-dire les p-values, sont intrinséquement *et* numériquement différentes de leurs équivalents bayésiens (voir aussi la Note 5.7.1), une modification du cadre décisionnel, proposée par Berger *et al.* (1994), permet une réconciliation partielle des deux approches. Bien qu'une telle réconciliation ne soit pas une caractéristique importante d'un point de vue bayésien–une procédure se doit avant tout d'être optimale pour le problème décisionnel considéré, plutôt que d'être stable sur le long terme–, elle a différents avantages en pratique : premièrement, les statisticiens sont plus enclins à utiliser une procédure bayésienne lorsque celle-ci jouit aussi de propriétés fréquentistes. Deuxièmement, ceci élimine le problème de l'interprétation d'une p-value comme une probabilité a posteriori.

Cette modification revient à ajouter l'option "pas de décision" aux réponses "acceptation" et "rejet" utilisées dans les tests classiques. Même si cette possibilité peut sembler absurde d'un point de vue décisionnel, elle est certainement défendable d'un point de vue statistique : il existe bien des cas où les données ne permettent pas une réponse concluante à l'égard de H_0 et nous font demander au client plus d'observations ou une information a priori plus précise. En fait, une telle approche existait déjà pour les tests séquentiels, comme les tests du rapport de vraisemblance séquentiels de Wald (voir Lehmann, 1986). (Notons cependant que cette procédure de Berger *et al.*, 1994, ne prend pas en compte les tests répétés, ce qui a un impact sur les niveaux de confiance ; voir aussi l'Exemple 1.18.)

Dans le cas de deux hypothèses simples,

$$H_0 : \ x \sim m_0(x) \quad \text{contre} \quad H_1 : \ x \sim m_1(x)\,,$$

où m_0 et m_1 sont des densités connues, le facteur de Bayes B_{10} est égal au rapport de vraisemblance $m_1(x)/m_0(x)$. Si l'option "pas de décision" est représentée par -1 , le test bayésien modifié de Berger *et al.* (1994) s'écrit

$$\varphi(x) = \begin{cases} 1 & \text{si } B_{10}(x) \le r, \\ 0 & \text{si } B_{10}(x) \ge a, \\ -1 & \text{si } r < B_{10}(x) < a, \end{cases} \tag{5.26}$$

avec pour estimateur associé

$$\gamma(x) = \begin{cases} 1/(1+B_{10}(x)) & \text{si } B_{10}(x) \ge a, \\ B_{10}(x)/(1+B_{10}(x)) & \text{si } B_{10}(x) \le r. \end{cases}$$

Notons que $\gamma(x)$ est la probabilité a posteriori de l'hypothèse rejetée et est donc optimale sous le coût quadratique. (Mais φ ne semble pas être une procédure décisionnelle ; voir l'Exercice 5.59.)

Si on note F_0 et F_1 les fonctions de répartition de $B_{10}(x)$ associées respectivement à m_0 et m_1 et si on définit $\Psi(b) = F_0^{-1}(1 - F_1(b))$, alors $\Psi^{-1}(b) = F_1^{-1}(1 - F_0(b))$ et Berger et al. (1994) prennent

$$(r, a) = \begin{cases} (1, \Psi(1)) & \text{si } \Psi(1) > 1\,, \\ (\Psi^{-1}(1), 1) & \text{si } \Psi(1) < 1\,. \end{cases}$$

Ces auteurs démontrent que l'estimateur $\gamma(x)$ est valide dans une perspective fréquentiste conditionnelle : conditionnellement en

$$S(x) = \min\{B_{10}(x), \Psi^{-1}(B_{10}(x))\}\,, \tag{5.27}$$

la procédure (φ, γ) vérifie

$$P_0(B_{10}(x) \ge a | S(x) = s) = \gamma(s)\,, \qquad P_1(B_{10}(x) \le r | S(x) = s) = \gamma(s)\,,$$

où $\gamma(x)$ ne dépend que de s (Exercice 5.57). Notons cependant que $S(x)$ n'est une statistique libre que dans quelques cas particuliers (Exercice 5.60).

La généralisation de ce résultat aux hypothèses composites,

$$H_0 : \ \theta = \theta_0 \qquad \text{contre} \qquad H_1 : \ \theta \in \Theta_1$$

s'obtient en réécrivant H_1 comme dans la Section 5.3.5, soit,

$$H_1 : \ x \sim m_1(x) = \int_{\Theta_1} f(x|\theta)\pi_1(\theta)d\theta\,.$$

Berger *et al.* (1997) montrent alors que l'évaluation fréquentiste conditionnelle sous H_0 coïncide de nouveau avec l'estimateur bayésien, mais dans un sens plus faible, car, si la procédure obtenue a de bonnes propriétés bayésiennes, sa validité fréquentiste est plus contestable (Hinkley, 1997, Louis, 1997).

Exemple 5.59. (Berger *et al.*, 1997) Pour $x_1, \ldots, x_n$ i.i.d. $\mathcal{N}(\theta, \sigma^2)$, avec σ connu, considérons le test de $H_0 : \ \theta = \theta_0$ sous l'a priori conjugué $\theta \sim \mathcal{N}(\mu, k\sigma^2)$. Si $z = \sqrt{n}(\bar{x}_n - \theta_0)/\sigma$, on obtient

$$m_0(z) = \frac{1}{\sqrt{2\pi}} \exp\{-z^2/2\}$$

et

$$m_1(z) = \frac{1}{\sqrt{2\pi}\sqrt{1+kn}} \exp\left\{ \frac{-(z+\sqrt{kn}\Delta)^2}{2(1+kn)} \right\},$$

avec $\Delta = (\theta_0 - \mu)/\sqrt{k}\sigma$. Le facteur de Bayes est alors

$$B_{10}(x) = \sqrt{1+kn} \exp\left\{ -\frac{kn}{2(1+kn)} \left[z - \frac{\Delta}{\sqrt{kn}} \right]^2 + \frac{\Delta^2}{2} \right\},$$

$\Psi(1) > 1$, $r = 1$ et $a = F_0^{-1}(1 - F_1(1))$. ‖

6

Méthodes de calcul bayésien

"The contraption began to quiver, steam hissing out from two or three places. The hiss grew to a shriek, and the thing began trembling."

Robert Jordan, *Lord of Chaos.*

6.1 Difficultés de mise en œuvre

À ce stade du livre, nous devons discuter de l'aspect pratique du paradigme bayésien, à savoir le calcul des estimateurs de Bayes. La simplicité ultime de l'approche bayésienne est que, pour une fonction de coût L et une loi a priori π données, l'estimation bayésienne associée à une observation x est la décision (habituellement unique) d minimisant le coût a posteriori

$$\mathrm{L}(\pi, d|x) = \int_{\Theta} \mathrm{L}(\theta, d)\pi(\theta|x)\, d\theta. \tag{6.1}$$

Dans la pratique cependant, minimiser (6.1) peut être rendu difficile pour deux raisons :

(i) le calcul explicite de la loi a posteriori, $\pi(\theta|x)$, peut être impossible ; et

(ii) même si $\pi(\theta|x)$ est connu, cela n'implique pas nécessairement que minimiser (6.1) soit facile ; en effet, lorsque l'intégration analytique est impossible, la minimisation numérique nécessite parfois un temps de calcul considérable, en particulier lorsque Θ et $\mathscr{D}$ sont de grandes dimensions.

Le point (i) peut sembler être une difficulté mineure et formelle, puisque minimiser (6.1) revient en réalité à minimiser

$$\int_{\Theta} \mathrm{L}(\theta, d)\pi(\theta) f(x|\theta)\, d\theta,$$

qui ne requiert pas une évaluation de $\pi(\theta|x)$. Cependant, nous avons vu dans les Chapitres 2 et 4 que les coûts classiques, comme le coût quadratique, mènent directement à des estimateurs s'exprimant en fonction de la loi a posteriori, notamment la moyenne a posteriori

$$\begin{aligned}\delta^{\pi}(x) &= \int_{\Theta} \theta\, \pi(\theta|x)\, d\theta \\ &= \frac{\int_{\Theta} \theta\, \pi(\theta) f(x|\theta)\, d\theta}{\int_{\Theta} \pi(\theta) f(x|\theta)\, d\theta},\end{aligned}$$

pour le coût quadratique ; ils nécessitent donc un calcul direct des moments. Une remarque similaire s'applique à l'obtention d'autres quantités a posteriori d'intérêt, comme les quantiles a posteriori, les facteurs de Bayes ou les régions de confiance.

Une réponse simpliste à ces difficultés de calcul est de n'utiliser que des modèles d'échantillonnage, des lois a priori et des coûts qui mènent à des solutions explicites pour la minimisation de (6.1). Cette approche restrictive est techniquement justifiée lorsque les outils de calcul décrits ci-dessous ne sont pas applicables, mais elle est inacceptable en termes subjectifs, car la fonction de coût et la loi a priori devraient être construites en fonction du problème de décision et non pas parce qu'elles fournissent des réponses analytiques, comme nous l'avons souligné dans le Chapitre 3[47].

Ce chapitre a donc pour but d'éviter le recours systématique à des lois a priori et à des coûts simples, en fournissant aux lecteurs une sélection représentative des méthodes d'approximation les plus récentes et les plus sophistiquées pouvant être utilisées lorsque la loi a posteriori ou un estimateur donné n'admettent pas d'expression analytique. Ce chapitre n'est qu'une introduction à ces méthodes ; les lecteurs sont renvoyés à Robert et Casella (2004) pour un traitement plus approfondi.

Bien que les problèmes d'estimation comme la minimisation du coût ou le calcul d'un estimateur MAP puissent aussi être résolus par des techniques de simulation (voir Geyer et Thompson, 1992, Geyer, 1996, Robert et Casella, 1999, Chapitre 5, ou Doucet *et al.*, 2002), nous nous concentrons dans ce chapitre sur les approximations de $\pi(\theta|x)$ et des intégrales correspondantes, parce qu'il s'agit de la pierre angulaire des difficultés de calcul en

[47] Les illustrations classiques ont recours à de tels cas simples, pour permettre une présentation plus claire et concise des points traités, et ce livre a beaucoup fait appel aux familles exponentielles, aux lois a priori conjuguées et aux coûts quadratiques. Néanmoins, une approche plus adaptative, reposant par exemple sur des mélanges des lois a priori conjuguées, devrait être adoptée en pratique.

inférence bayésienne. De plus, si $\pi(\theta|x)$ peut être approchée correctement, il est généralement possible de construire une approximation de $\mathrm{L}(\pi, d|x)$ pour une décision arbitraire d et d'utiliser alors une méthode de minimisation classique.

Nous présentons maintenant une série d'exemples utilisés tout au long de ce chapitre pour illustrer les différentes méthodes de calcul.

Exemple 6.1. Soit $x_1, \ldots, x_n$ un échantillon de $\mathscr{C}(\theta, 1)$, une loi de Cauchy de paramètre de position θ, avec $\theta \sim \mathscr{N}(\mu, \sigma^2)$, où μ et σ^2 sont des hyperparamètres connus. La loi a posteriori de θ est alors

$$\pi(\theta|x_1, \ldots, x_n) \propto e^{-(\theta-\mu)^2/2\sigma^2} \prod_{i=1}^{n} [1 + (x_i - \theta)^2]^{-1},$$

qui ne peut pas être intégrée de façon analytique. Lorsque δ^π est la moyenne a posteriori,

$$\delta^\pi(x_1, \ldots, x_n) = \frac{\int_{-\infty}^{+\infty} \theta e^{-(\theta-\mu)^2/2\sigma^2} \prod_{i=1}^{n} [1 + (x_i - \theta)^2]^{-1} d\theta}{\int_{-\infty}^{+\infty} e^{-(\theta-\mu)^2/2\sigma^2} \prod_{i=1}^{n} [1 + (x_i - \theta)^2]^{-1} d\theta},$$

son calcul nécessite deux intégrations numériques, une pour le numérateur et une autre pour le dénominateur. Le calcul de la variance requiert une intégration supplémentaire. De plus, la structure typiquement *multimodale* de cette loi (voir Exercice 1.27) fait que l'application de techniques d'intégration numérique standard peut nécessiter certains réglages délicats. ‖

Comme on l'a déjà vu auparavant, la difficulté de calcul peut provenir de la fonction de coût choisie, même lorsque la loi a priori est conjuguée.

Exemple 6.2. Soient $x|\theta \sim \mathscr{N}_p(\theta, \sigma^2 I_p)$ et $\theta|\mu, \tau \sim \mathscr{N}_p(\mu, \tau^2 I_p)$, d'hyperparamètres connus μ et τ. La loi a posteriori de θ admet alors une expression simple, puisque

$$\theta|x \sim \mathscr{N}_p\left(\frac{\sigma^2\mu + \tau^2 x}{\sigma^2 + \tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2} I_p\right).$$

Lorsque $||\theta||^2$ est le paramètre d'intérêt, le coût quadratique ramené à l'échelle de l'estimateur usuel est

$$\mathrm{L}(\theta, \delta) = \frac{(\delta - ||\theta||^2)^2}{2||\theta||^2 + p},$$

comme dans Saxena et Alam (1982). Il conduit à l'estimateur de Bayes suivant :

$$\delta^\pi(x) = \frac{\mathbb{E}^\pi[||\theta||^2/(2||\theta||^2 + p)|x]}{\mathbb{E}^\pi[1/(2||\theta||^2 + p)|x]}.$$

Bien que $(\sigma^{-2}+\tau^{-2})||\theta||^2$ soit distribué a posteriori comme une variable aléatoire $\chi^2_p(\lambda)$, avec

$$\lambda = \frac{||\sigma^2\mu+\tau^2 x||^2}{\sigma^2\tau^2(\sigma^2+\tau^2)},$$

δ^π n'admet pas d'expression analytique et une approximation numérique est de nouveau nécessaire. Notons que, dans ce cas, l'intégration numérique est plus compliquée que pour l'Exemple 6.1, car la densité de $\chi^2_p(\lambda)$ (voir l'Appendice A) fait intervenir une fonction de Bessel modifiée, $I_{(p-2)/2}(t)$, qui doit être approchée par une suite de densités du khi deux (centrées) pondérées ou par une approximation en fractions continues (voir l'Exercice 4.36). Une approche alternative est d'intégrer plutôt en θ, mais cela n'est possible que pour de petites valeurs de p. ‖

Les Chapitres 7 et 10 fourniront également des exemples où l'approximation d'estimateurs de Bayes est nécessaire. En effet, la plupart des *estimateurs de Bayes hiérarchiques* ne peuvent pas être calculés de façon analytique ; c'est le cas notamment pour des observations normales (voir le Lemme 10.17) et les modèles graphiques (voir la Note 10.7.1). De plus, une approximation numérique de ces estimateurs peut donner lieu à des complications, en particulier pour des dimensions plus grandes.

Exemple 6.3. Le recours à une *variable auxiliaire* dans un modèle de Student multivarié réduit le nombre d'intégrations à un, comme l'a remarqué Dickey (1968). Rappelons que, si

$$x \sim \mathcal{N}_p(\theta,\sigma^2 I_p), \qquad \theta \sim \mathcal{T}_p(\nu,\mu,\tau^2 I_p),$$

on peut écrire

$$\theta|\xi,x \sim \mathcal{N}_p\left(\xi(x),\frac{\tau^2\sigma^2}{\sigma^2\xi+\tau^2}I_p\right),$$
$$\pi(\xi|x) \propto \frac{\xi^{(p+\nu)/2-1}}{(\xi\sigma^2+\tau^2)^{p/2}}\exp\left\{\frac{-1}{2}\left(\frac{||x-\mu||^2\xi}{\tau^2+\xi\sigma^2}+\xi^2\nu\right)\right\},$$

avec

$$\xi(x) = \frac{\xi\sigma^2\mu+\tau^2 x}{\xi\sigma^2+\tau^2}$$

(voir l'Exemple 10.3). Soit la généralisation suivante :

$$x|\theta,\Lambda \sim \mathcal{N}_p(\theta,\Lambda),$$

lorsque θ et $\Lambda = \mathrm{diag}(\sigma_1^2,\ldots,\sigma_p^2)$ sont inconnus et de lois a priori $(1 \le i \le p)$

$$\theta_i|\sigma_i \sim \mathcal{N}\left(\mu_i,\frac{\sigma_i^2}{n_i}\right), \qquad \sigma_i^2 \sim \mathcal{IG}(\nu_i/2,s_i^2/2),$$

où les n_i, s_i et ν_i sont des hyperparamètres connus. Dans ce cas ($1 \leq i \leq p$),

$$\theta_i|x_i \sim \mathscr{T}\left(\nu_i+1, \frac{x_i+n_i\mu_i}{n_i+1}, \right.$$
$$\left. (\nu_i+1)^{-1}(n_i+1)^{-1}\left[s_i^2+\frac{n_i}{n_1+1}(x_i-\mu_i)^2\right]\right),$$

et le recours à une variable auxiliaire ξ_i pour chaque composante θ_i ne modifie pas la complexité du problème d'estimation, puisque le nombre d'intégrales à calculer reste constant. ‖

Les deux exemples ci-dessous sont paradoxaux, au sens où une expression explicite de l'estimateur de Bayes est disponible, mais ne peut pas être utilisée de façon simple dans la pratique, soit parce qu'elle entraîne une instabilité numérique et donc un manque de fiabilité du résultat (Exemple 6.4), soit parce que le calcul de l'estimateur de Bayes résultant est impossible, car il ne peut pas être effectué en un temps raisonnable pour des tailles d'échantillon réalistes (Exemple 6.5).

Exemple 6.4. Dans le cadre des modèles de *capture-recapture*, nous considérons le modèle temporel (voir la Section 4.3.3) et les lois conjuguées

$$x_i|N,p_i \sim \mathscr{B}(N,p_i),$$
$$\pi(N)=1/N, \quad p_i \sim \mathscr{B}e(\alpha,\beta) \qquad (1 \leq i \leq n).$$

Si x_+ est le nombre d'individus *différents* capturés au moins une fois parmi n captures, la loi a posteriori de N et $p=(p_1,\ldots,p_n)$ est, pour $x=(x_1,\ldots,x_n,x_+)$,

$$\pi(N,p|x) \propto \frac{(N-1)!}{(N-x_+)!}\prod_{i=1}^{n} p_i^{\alpha+x_i-1}(1-p_i)^{\beta+N-x_i-1}$$

et la loi marginale de N se calcule comme

$$\pi(N|x) \propto \frac{(N-1)!}{(N-x_+)!}\prod_{i=1}^{n} B(\alpha+x_i,\beta+N-x_i)$$
$$\propto \frac{(N-1)!}{(N-x_+)!}\prod_{i=1}^{n} \frac{\Gamma(\beta+N-x_i)}{\Gamma(\alpha+\beta+N)}.$$

Par conséquent, la loi a posteriori $\pi(N|x)$ peut s'écrire de façon "explicite",

$$\frac{\frac{(N-1)!}{(N-x_+)!}\prod_{i=1}^{n}\Gamma(\beta+N-x_i)/\Gamma(\alpha+\beta+N)}{\sum_{M=x_+}^{+\infty}\frac{(M-1)!}{(M-x_+)!}\prod_{i=1}^{n}\Gamma(\beta+M-x_i)/\Gamma(\alpha+\beta+M)}. \tag{6.2}$$

En réalité, de par les rapports présents au numérateur et au dénominateur, la formule (6.2) ne nécessite aucune évaluation de la fonction gamma : le recours à la formule récursive $\Gamma(x+1) = x\Gamma(x)$ suffit. Néanmoins, si n est grand, c'est-à-dire si plusieurs captures ont été entreprises, et si, de plus, les tailles de capture résultantes x_i sont très différentes, le calcul de la loi a posteriori (6.2) sera assez difficile. Les quantités (6.2) peuvent beaucoup fluctuer et la règle d'arrêt pour le calcul de la série infinie en (6.2) doit être conçue en conséquence, de crainte qu'on ignore les termes significatifs correspondant aux grandes valeurs de M. De plus, le calcul de la suite (6.2) par la formule de récurrence

$$\frac{\pi(N+1|x)}{\pi(N|x)} = \frac{N}{N+1-x_+} \prod_{i=1}^{n} \frac{\beta+N-x_i}{\alpha+\beta+N},$$

bien que possible, peut être imprécis, car l'erreur d'approximation augmente à chaque étape, en particulier lorsque les x_i sont très différents.

La même critique s'applique au calcul de la moyenne a posteriori

$$\delta^\pi(x) = \frac{\sum_{N=x_+}^{+\infty} \frac{N!}{(N-x_+)!} \prod_{i=1}^{n} \Gamma(\beta+N-x_i)/\Gamma(\alpha+\beta+N)}{\sum_{M=x_+}^{+\infty} \frac{(M-1)!}{(M-x_+)!} \prod_{i=1}^{n} \Gamma(\beta+M-x_i)/\Gamma(\alpha+\beta+M)}\,. \tag{6.3}$$

Par conséquent, même si ces modèles discrets paraissent simples d'un point de vue analytique, les formules explicites ci-dessus ne peuvent être utilisées que pour les exemples les plus triviaux. Lorsque les nombres d'observations et de captures sont importants, il devient nécessaire de recourir à des méthodes numériques alternatives. En outre, l'attrait de telles formules disparaît dans un cadre hiérarchique, car elles ne peuvent pas être utilisées lorsque le couple (α, β) suit une loi a priori (voir George et Robert, 1992). ‖

Exemple 6.5. Soit un échantillon $x_1, \ldots, x_n$ de

$$f(x|\theta) = p\varphi(x; \mu_1, \sigma_1) + (1-p)\varphi(x; \mu_2, \sigma_2), \tag{6.4}$$

c'est-à-dire un mélange de deux lois normales de moyennes μ_i, variances σ_i^2 $(i = 1, 2)$ et poids p $(0 < p < 1)$. Une motivation radiologique de ce modèle a été donnée dans l'Exemple 1.6. Une étude sur un premier ensemble de radiographies des poumons a montré que les images étaient distribuées avec des paramètres qui varient selon la Table 6.1.

Comme première approximation et étant donné l'information fournie par la Table 6.1, une modélisation a priori possible consiste à utiliser des lois a priori "conjuguées" pour $\theta = (\mu_1, \sigma_1^2, p, \mu_2, \sigma_2^2)$,

$$\mu_i|\sigma_i \sim \mathcal{N}(\xi_i, \sigma_i^2/n_i), \qquad \sigma_i^2 \sim \mathcal{IG}(\nu_i/2, s_i^2/2), \qquad p \sim \mathcal{B}e(\alpha, \beta),$$

Tab. 6.1. Paramètres statistiques pour un modèle de radiographie des poumons. (*Source :* Plessis, 1989.)

	μ_1	μ_2	σ_1	σ_2	p
Moyenne	105.33	188.9	32.3	18.2	0.5
Écart type	11.18	7.38	5.62	4.5	0.08

et à calculer la valeur des hyperparamètres ξ_i, n_i, ν_i, s_i et (α, β) à partir de la Table 6.1 par la méthode des moments[48]. En effet, ces lois ne sont pas conjuguées au sens de la Définition 3.7, mais la loi a posteriori correspondante est

$$\pi(\theta|x_1,\ldots,x_n) \propto \prod_{j=1}^{n} \{p\varphi(x_j;\mu_1,\sigma_1) + (1-p)\varphi(x_j;\mu_2,\sigma_2)\}\,\pi(\theta)\,. \tag{6.5}$$

On peut réécrire (6.5) simplement en représentant cette distribution comme une somme pondérée (c'est-à-dire un mélange) de lois conjuguées,

$$\pi(\theta|x_1,\ldots,x_n) = \sum_{\ell=0}^{n} \sum_{(k_t)} \omega(k_t)\pi(\theta|(k_t)), \tag{6.6}$$

où ℓ représente le nombre d'observations attribuées à la première composante et où la seconde somme prend en compte toutes les permutations (k_t) de $\{1,2,\ldots,n\}$ correspondant à une partition différente de $\{x_1,\ldots,x_n\}$ en $\{x_{k_1},\ldots,x_{k_\ell}\}$ et $\{x_{k_{\ell+1}},\ldots,x_{k_n}\}$, caractérisant ainsi les ℓ observations attribuées à la première composante. Le poids a posteriori d'une partition (k_t) est (voir ci-dessous pour la notation)

$$\begin{aligned}\omega(k_t) \propto\ & \frac{\Gamma(\alpha+\ell)\,\Gamma(\beta+n-\ell)\,\Gamma([\nu_1+\ell]/2)}{\left(s_1^2+\hat{s}_1(k_t)+\frac{n_1\ell}{n_1+\ell}(\xi_1-\bar{x}_1(k_t))^2\right)^{(\nu_1+\ell)/2}} \\ & \times \frac{\Gamma([\nu_2+n-\ell]/2)\big/\sqrt{(n_1+\ell)(n_2+n-\ell)}}{\left(s_2^2+\hat{s}_2(k_t)+\frac{n_2(n-\ell)}{n_2+n-\ell}(\xi_2-\bar{x}_2(k_t))^2\right)^{(\nu_2+n-\ell)/2}},\end{aligned}$$

normalisé de telle manière que

$$\sum_{\ell=0}^{n} \sum_{(k_t)} \omega(k_t) = 1.$$

[48]Notons que cet a priori diffère d'une *modélisation bayésienne empirique* (Chapitre 10). En effet, bien que l'a priori résultant ne soit qu'une approximation et que l'hyperparamètre soit estimé par des moyennes classiques, cette loi est fondée sur des observations *précédentes*, ce qui peut être considéré comme une *information a priori*, et non sur l'échantillon observé pour lequel le paramètre θ est inconnu.

Pour une permutation donnée (k_t), la loi a posteriori conditionnelle est

$$\begin{aligned}\pi(\theta|(k_t)) = &\mathcal{N}\left(\xi_1(k_t), \frac{\sigma_1^2}{n_1+\ell}\right) \times \mathcal{IG}((\nu_1+\ell)/2, s_1(k_t)/2)\\ &\times\mathcal{N}\left(\xi_2(k_t), \frac{\sigma_2^2}{n_2+n-\ell}\right) \times \mathcal{IG}((\nu_2+n-\ell)/2, s_2(k_t)/2)\\ &\times\mathcal{B}e(\alpha+\ell, \beta+n-\ell)\,,\end{aligned}$$

où

$$\begin{array}{ll}\bar{x}_1(k_t) = \frac{1}{\ell}\sum_{t=1}^{\ell} x_{k_t}, & \hat{s}_1(k_t) = \sum_{t=1}^{\ell}(x_{k_t}-\bar{x}_1(k_t))^2,\\ \bar{x}_2(k_t) = \frac{1}{n-\ell}\sum_{t=\ell+1}^{n} x_{k_t}, & \hat{s}_2(k_t) = \sum_{t=\ell+1}^{n}(x_{k_t}-\bar{x}_2(k_t))^2\end{array}$$

sont les statistiques habituelles pour les deux sous-échantillons induits par la permutation et

$$\begin{aligned}\xi_1(k_t) &= \frac{n_1\xi_1+\ell\bar{x}_1(k_t)}{n_1+\ell}, \qquad \xi_2(k_t) = \frac{n_2\xi_2+(n-\ell)\bar{x}_2(k_t)}{n_2+n-\ell},\\ s_1(k_t) &= s_1^2+\hat{s}_1^2(k_t)+\frac{n_1\ell}{n_1+\ell}(\xi_1-\bar{x}_1(k_t))^2,\\ s_2(k_t) &= s_2^2+\hat{s}_2^2(k_t)+\frac{n_2(n-\ell)}{n_2+n-\ell}(\xi_2-\bar{x}_2(k_t))^2,\end{aligned}$$

sont les mises à jour a posteriori des hyperparamètres, conditionnellement à la partition (k_t).

Cette décomposition est intéressante, car elle montre que, malgré une formule apparemment inextricable, l'analyse bayésienne de la loi de mélange (6.4) est assez logique. En effet, la loi a posteriori prend en compte *toute* partition possible de l'échantillon, en spécifiant de quelle composante chaque observation est originaire via la permutation correspondante (k_t). Il attribue alors un poids $\omega(k_t)$ à la partition, qui peut être interprété comme la probabilité a posteriori de la partition choisie, et opère comme si chaque observation provenait en réalité de la composante choisie, les lois a posteriori (conditionnelles) $\pi(\theta|(k_t))$ étant identiques aux lois a posteriori habituelles pour (μ_1, σ_1) et (μ_2, σ_2) résultant de l'observation *séparée* de $x_{k_1}, \ldots, x_{k_\ell}$ et $x_{k_{\ell+1}}, \ldots, x_{k_n}$. Des remarques similaires s'appliquent à la loi a posteriori de p, car, conditionnellement à la partition (k_t), cette loi correspond à la loi a posteriori associée à l'observation d'une variable aléatoire binomiale $\mathcal{B}(n,p)$, qui est le nombre d'observations attribuées à la première composante.

La décomposition (6.6) fournit l'estimateur de Bayes suivant de θ :

$$\delta^\pi(x_1,\ldots,x_n) = \sum_{\ell=0}^{n}\sum_{(k_t)}\omega(k_t)\mathbb{E}^\pi[\theta|\mathbf{x},(k_t)],$$

la somme pondérée des estimateurs de Bayes pour chaque partition. Par exemple, l'estimateur de Bayes de μ_1 est

$$\mu_1^\pi(x_1,\ldots,x_n) = \sum_{\ell=0}^{n} \sum_{(k_t)} \omega(k_t)\xi_1(k_t). \tag{6.7}$$

Ces développements sont satisfaisants d'un point de vue théorique, car les estimateurs résultants sont faciles à interpréter et intuitivement convaincants. De façon naturelle, la loi a posteriori prend en compte la possibilité que cette observation ait été générée par la première ou la deuxième composante, puisque l'origine de chaque observation dans l'échantillon est inconnue. Cependant, le calcul pratique de (6.7) implique deux sommes de 2^n termes chacune, ce qui correspond exactement à l'ensemble des partitions différentes de l'échantillon. Il est donc impossible de calculer un estimateur de Bayes de cette façon, pour la plupart des tailles d'échantillon[49]. ‖

L'Exemple 6.5 est représentatif d'un type de modèles statistiques affectés par des problèmes similaires, incluant la plupart des modèles à *données manquantes* (ou *variables latentes*) comme les mélanges, les modèles censurés et la classification (voir Robert et Casella, 1999, Chapitre 9). Ces modèles sont *paradoxaux* au sens où des constructions explicites des estimateurs de Bayes peuvent être formellement disponibles, mais sont inutiles en pratique de par le temps de calcul qu'elles impliquent. De plus, la difficulté de calcul augmente avec la taille de l'échantillon, conduisant à ce qui peut être appelé un paradoxe de l'information, car plus on a d'information, plus il devient difficile de mener une inférence[50] sur θ. Dans de tels cadres, les méthodes d'approximation numérique sont rarement appropriées et des solutions adaptées sont nécessaires, comme celles développées dans les Sections 6.3 et 6.4.

6.2 Méthodes classiques d'approximation

Cette section couvre brièvement quelques techniques classiques qui peuvent faciliter les calculs bayésiens ; la section suivante en revanche traite des méthodes de simulations récentes qui semblent particulièrement adaptées aux exigences de l'approche bayésienne. Une présentation plus détaillée est fournie par Robert et Casella (2004, Chapitres 2-5) ; voir aussi Berger (2000) et Carlin et Louis (2000a) pour une présentation des logiciels bayésiens disponibles.

[49] Par exemple, s'il faut une seconde de temps processeur pour évaluer (6.7) pour un échantillon de taille 20, le calcul de l'estimateur correspondant à un échantillon de taille 40 devrait prendre douze jours.

[50] À strictement parler, la difficulté de calcul grandit toujours avec la taille de l'échantillon, même dans les cas où une statistique exhaustive existe. Cependant, dans le cas de l'Exemple 6.5, cette croissance est tellement rapide (taux exponentiel) qu'elle empêche complètement le calcul même. (De tels problèmes sont appelés *NP-complets* en Recherche opérationnelle.)

6.2.1 Intégration numérique

À partir de la simple méthode de Simpson[51], plusieurs approches ont été conçues en Mathématiques appliquées pour l'approximation numérique d'intégrales. Par exemple, la *quadrature polynomiale* est censée approcher les intégrales liées à des distributions proches de la loi normale (voir Naylor et Smith, 1982, Smith *et al.*, 1985, ou Verdinelli et Wasserman, 1998, pour une introduction détaillée). L'approximation de base est donnée par

$$\int_{-\infty}^{+\infty} e^{-t^2/2} f(t)\, dt \approx \sum_{i=1}^{n} \omega_i f(t_i),$$

où

$$\omega_i = \frac{2^{n-1} n!\, \sqrt{n}}{n^2 [H_{n-1}(t_i)]^2}$$

et t_i est le i-ième zéro du n-ième *polynôme d'Hermite*, $H_n(t)$.

D'autres approximations d'intégrales reliées à la méthode précédente sont disponibles, qui reposent sur différentes bases orthogonales classiques (voir Abramowitz et Stegun, 1964), ou les *ondelettes* (voir la Note 1.8.2 et Müller et Vidakovic, 1999, Chapitre 1), mais ces méthodes requièrent généralement des hypothèses de régularité sur la fonction f, ainsi que des études préliminaires pour déterminer quelle base est la plus adéquate et à quel point cette approximation est précise. Par exemple, des transformations du modèle peuvent être nécessaires pour mettre en pratique l'approximation d'Hermite (voir Naylor et Smith, 1982, et Hills et Smith, 1992) ; Morris (1982) (voir aussi Diaconis et Zabell, 1991) montre comment les lois des familles exponentielles à variance quadratique (Exercices 3.24 et 10.33) peuvent être associées à une base orthogonale particulière (Exercice 6.18).

Cependant, quelle que soit la méthode d'intégration numérique utilisée, sa précision diminue dramatiquement lorsque la dimension de Θ augmente. De façon plus spécifique, l'erreur associée aux méthodes numériques se comporte comme une puissance de la dimension de Θ. En pratique, une règle empirique est que la plupart des méthodes standard ne devraient pas être utilisées pour l'intégration en dimension supérieure à 4, même si ces méthodes continuent à s'améliorer année après année. En effet, la taille de la partie de l'espace non pertinente pour le calcul d'une intégrale donnée augmente considérablement avec la dimension de l'espace. Ce problème est appelé *fléau de la dimension*, voir Robert et Casella (1999, Chapitre 3) pour des détails.

6.2.2 Les méthodes de Monte Carlo

Dans un problème statistique, l'approximation de l'intégrale

[51] Voir Stigler (1986) pour une plus forte connexion entre Simpson (1710-1761) et la Statistique bayésienne.

$$\int_\Theta g(\theta)f(x|\theta)\pi(\theta)\,d\theta, \tag{6.8}$$

doit tirer avantage de la nature particulière de (6.8), à savoir le fait que π soit une densité de probabilité (en supposant qu'il s'agisse d'une loi a priori propre) ou plutôt, que $f(x|\theta)\pi(\theta)$ soit proportionnel à une densité. Une conséquence naturelle de cette perspective est d'utiliser *la méthode de Monte Carlo*, introduite par Metropolis et Ulam (1949) et von Neumann (1951). Par exemple, s'il est possible de produire des variables aléatoires $\theta_1, \ldots, \theta_m$ de loi $\pi(\theta)$, la moyenne

$$\frac{1}{m}\sum_{i=1}^{m} g(\theta_i)f(x|\theta_i) \tag{6.9}$$

converge (presque sûrement) vers (6.8) lorsque m tend vers $+\infty$, selon la Loi des Grands Nombres. De la même façon, si un échantillon iid de θ_i de $\pi(\theta|x)$ peut être simulé, la moyenne

$$\frac{1}{m}\sum_{i=1}^{m} g(\theta_i) \tag{6.10}$$

converge vers

$$\frac{\int_\Theta g(\theta)f(x|\theta)\pi(\theta)\,d\theta}{\int_\Theta f(x|\theta)\pi(\theta)\,d\theta}\,.$$

De plus, si la variance a posteriori $\mathrm{var}(g(\theta)|x)$ est finie, le Théorème Central Limit s'applique à la moyenne (6.10), qui est alors asymptotiquement normale, de variance $\mathrm{var}(g(\theta)|x)/m$. Des régions de confiance peuvent alors se construire à partir de cette approximation normale et, de manière décisive, il découle aussi du Théorème Central Limit que l'ordre de grandeur de l'erreur est $1/\sqrt{m}$ quelle que soit la dimension du problème, au contraire des méthodes numériques.

La mise en œuvre de cette méthode nécessite la production d'une suite iid θ_i par ordinateur, reposant sur un générateur pseudo-aléatoire déterministe imitant la génération de $\pi(\theta)$ ou de $\pi(\theta|x)$ comme suit : un échantillon iid d'une loi uniforme $\mathscr{U}([0,1])$ est généré (voir la Note 6.6.1), puis transformé en variables de la loi d'intérêt (voir Robert et Casella, 2004, Chapitre 2).[52] Les techniques statistiques standard peuvent aussi être utilisées pour déterminer l'erreur d'approximation de (6.8) par la moyenne (6.9).

En réalité, la méthode de Monte Carlo s'applique dans un cadre beaucoup plus général que pour la simulation de π, comme dans le cas ci-dessus. Par exemple, puisque (6.8) peut se représenter de plusieurs manières, il n'est pas nécessaire de simuler les lois $\pi(\cdot|x)$ ou π pour obtenir une bonne approximation

[52]Il n'est pas surprenant que les méthodes de Monte Carlo apparaissent au même moment que les premiers ordinateurs. Ces méthodes ne pouvaient tout simplement pas exister sans ordinateurs et ont en fait contribué aux premiers programmes d'ordinateurs jamais écrits.

de (6.8). En effet, si h est une densité de probabilité telle que supp(h) inclut le support de $g(\theta)f(x|\theta)\pi(\theta)$, l'intégrale (6.8) peut aussi être représentée comme une espérance en h, à savoir

$$\int \frac{g(\theta)f(x|\theta)\pi(\theta)}{h(\theta)} h(\theta)\, d\theta.$$

Cette représentation conduit à la *méthode de Monte Carlo avec fonction d'importance* h : générer $\theta_1, \ldots, \theta_m$ selon h et approcher (6.8) par

$$\frac{1}{m}\sum_{i=1}^{m} g(\theta_i)\omega_i(\theta_i),$$

avec les poids $\omega(\theta_i) = f(x|\theta_i)\pi(\theta_i)/h(\theta_i)$. De nouveau, par la Loi des Grands Nombres, cette approximation converge presque certainement vers (6.8). Et une approximation de $\mathbb{E}^\pi[g(\theta)|x]$ est donnée par

$$\frac{\sum_{i=1}^{m} g(\theta_i)\omega(\theta_i)}{\sum_{i=1}^{m} \omega(\theta_i)}, \tag{6.11}$$

car le numérateur et le dénominateur convergent respectivement vers

$$\int_\Theta g(\theta)f(x|\theta)\pi(\theta)\, d\theta \qquad \text{et} \qquad \int_\Theta f(x|\theta)\pi(\theta)\, d\theta,$$

si supp(h) inclut supp$(f(x|\cdot)\pi)$. Notons que le rapport (6.11) ne dépend d'aucune des constantes de normalisation apparaissant dans $h(\theta)$, $f(x|\theta)$ ou $\pi(\theta)$. L'approximation (6.11) peut par conséquent être utilisée lorsque certaines de ces constantes de normalisation sont inconnues.

Bien que (6.11) converge théoriquement vers $\mathbb{E}^\pi[g(\theta)|x]$ pour toutes les fonctions h vérifiant la condition des supports (Exercice 6.8), le choix de la fonction d'importance est primordial. Tout d'abord, il doit être aisé de simuler selon h, à l'aide d'un générateur pseudo-aléatoire rapide et fiable. (Voir les Exercices 6.9-6.12 pour quelques algorithmes de simulation de lois usuelles. Devroye, 1985, Fishman, 1996, Gentle, 1998, et Robert et Casella, 1999, Chapitre 2, présentent ces méthodes en détail.) De plus, la fonction $h(\theta)$ doit être suffisamment proche de $g(\theta)\pi(\theta|x)$, pour réduire autant que possible la variabilité de (6.11) (Exercice 6.14) ; sinon, la plupart des poids $\omega(\theta_i)$ prendront des valeurs très faibles, et un petit nombre d'entre eux auront une trop forte influence. En effet, si

$$\mathbb{E}^h[g^2(\theta)\omega^2(\theta)]$$

n'est pas finie, la variance de l'estimateur (6.11) est infinie. Bien entendu, la dépendance à g de la fonction d'importance h peut être évitée en proposant des choix génériques comme celui de la loi a posteriori $\pi(\theta|x)$ (qui n'est pas nécessairement le meilleur choix, voir les Exercices 6.13 et 6.14).

Exemple 6.6. (Suite de l'Exemple 6.2) La loi a posteriori de $\eta = ||\theta||^2$ est bien connue, car $\pi(\eta|x)$ est une loi du khi deux décentré $\chi^2_p(\lambda)$ avec coefficient d'échelle $\sigma^2\tau^2/(\sigma^2+\tau^2)$. Simuler un échantillon $\eta_1, \ldots, \eta_m$ de $\pi(\eta|x)$ est trivial : générer

$$\xi_1, \ldots, \xi_n \sim \mathcal{N}(\sqrt{\lambda}, 1), \qquad \zeta_1, \ldots, \zeta_n \sim \mathcal{G}\left(\frac{p-1}{2}, \frac{1}{2}\right)$$

et prendre $\eta_i = \sigma^2\tau^2(\xi_i^2 + \zeta_i)/(\sigma^2 + \tau^2)$ $(i = 1, \ldots, n)$. Nous pouvons alors approcher (6.3) par

$$\hat{\delta}^\pi(x) = \frac{\sum_{i=1}^m \eta_i/(2\eta_i + p)}{\sum_{i=1}^m 1/(2\eta_i + p)}. \tag{6.12}$$

De plus, la variance de (6.12) contrôle la précision de l'approximation (et le choix de m). ‖

Lorsque la loi a posteriori n'est pas disponible, un autre choix simple de fonction d'importance est la loi a priori π. Bien entendu, ceci est intéressant, non pas lorsque π est forcément explicite, mais au moins facile à simuler, par exemple, dans des modèles hiérarchiques où les deux niveaux correspondent à des lois propres. Le même appel à la prudence s'applique de nouveau cependant, puisque π doit être suffisamment proche de $\pi(\theta|x)$ et la variance de l'estimateur (6.11) finie. (Notez que cette condition de finitude est généralement satisfaite puisque $\pi(\theta)$ a souvent des queues plus épaisses que $\pi(\theta|x)$.) Bien évidemment, ce choix est impossible lorsque π est impropre.

Exemple 6.7. (Suite de l'Exemple 6.1) Puisque $\pi(\theta)$ est la loi normale $\mathcal{N}(\mu, \sigma^2)$, il est possible de simuler un échantillon normal $\theta_1, \ldots, \theta_M$ et d'approcher l'estimateur de Bayes par

$$\hat{\delta}^\pi(x_1, \ldots, x_n) = \frac{\sum_{t=1}^M \theta_t \prod_{i=1}^n [1 + (x_i - \theta_t)^2]^{-1}}{\sum_{t=1}^M \prod_{i=1}^n [1 + (x_i - \theta_t)^2]^{-1}}. \tag{6.13}$$

Dans le cas où les x_i sont tous loin de μ, ce choix peut être nuisible puisqu'à la fois le dénominateur et les poids des θ_t dans le numérateur sont petits pour la plupart des θ_t, et l'approximation $\hat{\delta}^\pi$ est par conséquent assez instable ; la Figure 6.1 représente le résultat de cinq cents estimations parallèles suivant (6.13), fondées sur $M = 1\,000$ simulations chacune, via l'écart interquartiles central à 90% des $\hat{\delta}^\pi$ moins la moyenne totale. La variation de δ^π augmente rapidement entre $\mu = 3$ et $\mu = 4$. Cela montre que, lorsque $\mu > 3$, de petits changements dans la simulation des θ_t peuvent produire des variations drastiques de $\hat{\delta}^\pi$. ‖

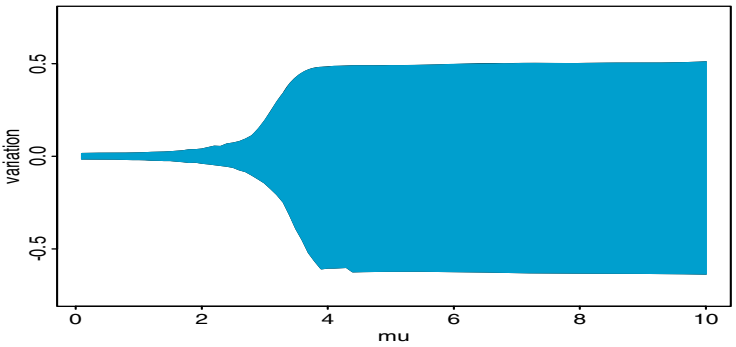

Fig. 6.1. Intervalle de variation à 90% de l'approximation (6.13) lorsque μ varie, pour $n = 10$ observations d'une loi de Cauchy $\mathscr{C}(0,1)$ et $M = 1,000$ simulations de Monte Carlo de θ tirées d'une loi $\mathscr{N}(\mu,1)$.

Exemple 6.8. Soit le modèle

$$x \sim \mathscr{N}_p(\theta, I_p)\,, \quad \theta|c \sim \mathscr{U}_{\{||\theta||^2=c\}} \quad \text{et} \quad c \sim \mathscr{G}(\alpha,\beta)\,.$$

(La justification de ce modèle sera donnée dans l'Exemple 10.26.) Bien que

$$\pi(\theta|x) = \int_0^{+\infty} \pi_1(\theta|x,c)\pi_2(c|x)\,dc$$

conduise à une loi a posteriori et à un estimateur de Bayes tous les deux explicites (voir l'Exemple 10.26), il peut être plus intéressant de générer $c_1,\ldots,c_m$ selon $\mathscr{G}(\alpha,\beta)$, puis les θ_i selon $\mathscr{U}_{\{||\theta||^2=c_i\}}$ $(1 \le i \le m)$ et d'approcher la moyenne a posteriori par

$$\hat{\delta}^\pi(x) = \frac{\sum_{i=1}^m \theta_i \exp\{-||x-\theta_i||^2/2\}}{\sum_{i=1}^m \exp\{-||x-\theta_i||^2/2\}},$$

car cela évite le calcul de fonctions hypergéométriques confluentes. ‖

Lorsque la vraisemblance $\ell(\theta|x)$ peut être normalisée comme une densité, un choix possible de fonction d'importance est $h(\theta) \propto \ell(\theta|x)$. Ce choix a du sens lorsque $\pi(\theta|x)$ est quasi proportionnel à la vraisemblance–comme c'est le cas pour de grandes tailles d'échantillon ou pour des lois a priori presque constantes. Cela peut arriver notamment pour des modèles exponentiels, car, si

$$f(x|\theta) \propto e^{\theta.x-\psi(\theta)},$$

un échantillon $\theta_1,\ldots,\theta_m$ de

$$h(\theta) \propto e^{\theta.x-\psi(\theta)}$$

peut en général être obtenu facilement (voir l'Exercice 6.23 pour une limitation de cette approche).

Une remarque finale sur le choix de la fonction d'importance est qu'il existe généralement un compromis entre des études préliminaires conduisant à une "bonne" fonction h et des algorithmes rapides. Par exemple, lorsque h est choisie parce qu'elle facilite la simulation des θ_i, il faut faire attention à ses queues et s'assurer qu'elles sont plus lourdes que celles de $\pi(\theta|x)$, pour éviter une convergence lente et des variances infinies. D'un autre côté, si h est spécialement réglée pour le calcul d'une intégrale spécifique (Exercice 6.14), elle peut ne pas donner de si bons résultats pour une autre intégrale, même si, en principe, le même échantillon des θ_i peut être utilisé pour le calcul de toute intégrale arbitraire. Cependant, ces difficultés potentielles mises à part, les méthodes d'échantillonnage d'importance constituent un outil très général et finissent souvent par devenir compétitives à l'égard des techniques de Monte Carlo par chaînes de Markov (Section 6.3), comme le montrent par exemple les méthodes de *filtrage particulaire* (voir Doucet *et al.*, 2001 et Cappé *et al.*, 2005) et de Monte Carlo *populationnel* (Cappé *et al.*, 2004, Douc *et al.*, 2005)

Par comparaison avec les méthodes d'intégration numérique, les méthodes de Monte Carlo présentent en effet l'avantage que, une fois l'échantillon $\theta_1, \ldots, \theta_n$ produit, celui-ci peut être utilisé à plusieurs reprises pour tous les objectifs inférentiels, incluant l'obtention des règles de Bayes à partir du coût a posteriori approché

$$\hat{\mathrm{L}}(\pi, d|x) = \frac{1}{m} \sum_{i=1}^{m} \mathrm{L}(\theta_i, d|x).$$

Cependant, si la dimension du problème est petite et si les fonctions à intégrer sont assez régulières, les méthodes d'intégration numérique ont tendance à donner de plus petites erreurs et de meilleurs contrôles de convergence. Des références supplémentaires et une discussion plus détaillée sur les méthodes de Monte Carlo, incluant les techniques améliorées de variables antithétiques et de contrôle, et leurs applications à la statistique bayésienne, peuvent être trouvées dans Robert et Casella (2004) et Chen *et al.* (2000).

6.2.3 L'approximation analytique de Laplace

Lorsque la fonction à intégrer dans (6.8) est assez régulière, il existe une solution alternative analytique–mais asymptotique–aux simulations de Monte Carlo. Cette méthode a été introduite par Laplace et est par conséquent appelée *approximation de Laplace*. Soit une espérance a posteriori

$$\mathbb{E}^{\pi}[g(\theta)|x] = \frac{\int_{\Theta} g(\theta) f(x|\theta) \pi(\theta)\, d\theta}{\int_{\Theta} f(x|\theta)\pi(\theta)\, d\theta}.$$

Ce rapport d'intégrales peut s'écrire

$$\mathbb{E}^\pi[g(\theta)|x] = \frac{\int_\Theta b_N(\theta)\exp\{-nh_N(\theta)\}\,d\theta}{\int_\Theta b_D(\theta)\exp\{-nh_D(\theta)\}\,d\theta}, \tag{6.14}$$

où la dépendance en x est supprimée par souci de simplicité et où n est normalement la taille de l'échantillon (bien qu'il puisse parfois correspondre à la variance a priori inverse, comme dans Robert (1993b) ou dans l'Exemple 6.11). Lorsque $h_N(\theta) = h_D(\theta)$, $\mathbb{E}^\pi[g(\theta)|x]$ s'écrit sous une *forme standard*; lorsque $b_N(\theta) = b_D(\theta)$, l'espérance a posteriori (6.14) est écrite sous *forme exponentielle complète*, pour reprendre la terminologie de Tierney et Kadane (1986). Pour une fonction donnée h admettant un minimum unique $\hat\theta$, le *développement de Laplace* d'une intégrale générale est donné par

$$\int b(\theta)e^{-nh(\theta)}d\theta = \sqrt{2\pi}\sigma e^{-n\hat h}\left\{\hat b + \frac{1}{2n}\left[\sigma^2\hat b'' - \sigma^4\hat b'\hat h''' + \frac{5}{12}\hat b(\hat h''')^2\sigma^6 - \frac{1}{4}\hat b\hat h^{(4)}\sigma^4\right]\right\} + \mathrm{O}(n^{-2}),$$

où $\hat b$, $\hat h$, etc., sont les valeurs prises par b, h et leurs dérivées pour $\theta = \hat\theta$, et $\sigma^2 = [h''(\hat\theta)]^{-1}$ (voir Olver, 1974, et Schervish, 1995). Cette approximation du deuxième ordre ne nécessite le calcul que des deux premières dérivées de g, par opposition à une approche similaire proposée par Lindley (1980). En plus, si on suppose que h_N et h_D satisfont $\hat h_N - \hat h_D = \mathrm{O}(n^{-1}), \ldots, \hat h_N^{(4)} - \hat h_D^{(4)} = \mathrm{O}(n^{-1})$ (comme c'est bien entendu le cas pour la forme standard), le développement de Laplace conduit à l'approximation suivante de $\mathbb{E}^\pi[g(\theta)|x]$ (avec $\hat b_D = b_D(\hat\theta_D)$, $\hat b_N = b_N(\hat\theta_N)$, et ainsi de suite) :

Lemme 6.9. *Si* $\hat b_D \neq 0$,

$$\frac{\int_\Theta b_N(\theta)\exp\{-nh_N(\theta)\}\,d\theta}{\int_\Theta b_D(\theta)\exp\{-nh_D(\theta)\}\,d\theta} = \frac{\sigma_N}{\sigma_D}e^{-n(\hat h_N - \hat h_D)}\left[\frac{\hat b_N}{\hat b_D} + \frac{\sigma_D^2}{2n\hat b_D^2}\left\{\hat b_D\hat b_N'' - \hat b_N\hat b_D'' - \sigma_D^2\hat h_D'''(\hat b_D\hat b_N' - \hat b_N\hat b_D')\right\}\right] + \mathrm{O}(n^{-2}).$$

Une démonstration de ce résultat est donnée dans Tierney *et al.* (1989) (voir aussi l'Exercice 6.17). Le Lemme 6.9 implique alors le développement suivant pour les deux formes du rapport (6.14) :

Corollaire 6.10. *Lorsque* $\mathbb{E}^\pi[g(\theta)|x]$ *s'écrit de façon standard,*

$$\mathbb{E}^\pi[g(\theta)|x] = \hat g + \frac{\sigma_D^2\hat b_D'\hat g'}{n\hat b_D} + \frac{\sigma_D^2\hat g''}{2n} - \frac{\sigma_D^4\hat h'''\hat g'}{2n} + \mathrm{O}(n^{-2}). \tag{6.15}$$

Pour la forme exponentielle complète, si g *est positive et* $g(\hat\theta_D)$ *est uniformément bornée (en* n*) par des constantes strictement positives,*

$$\mathbb{E}^\pi[g(\theta)|x] = \frac{\hat b_N}{\hat b_D}\frac{\sigma_N^2}{\sigma_D^2}e^{-n(\hat h_n - \hat h_D)} + \mathrm{O}(n^{-2}). \tag{6.16}$$

Preuve. Pour la forme standard, $h_N = h_D$; donc, $b_N = gb_D$, $\hat{\theta}_D = \hat{\theta}_N$. Par conséquent,

$$\frac{\hat{b}_D\hat{b}'_N - \hat{b}_N\hat{b}'_D}{\hat{b}_D^2} = \left(\frac{b_N}{b_D}\right)'\Big|_{\theta=\hat{\theta}_D} = \hat{g}'$$

et

$$\frac{\hat{b}_D\hat{b}''_N - \hat{b}_N\hat{b}''_D}{\hat{b}_D^2} = \hat{g}'' + 2\frac{\hat{b}'_D}{\hat{b}_D}\hat{g}'.$$

Le résultat découle alors du Lemme 6.9.

Dans le cas exponentiel complet, posons $h_N = h_D - (1/n)\log(g)$. Puisque nous supposons que $g(\hat{\theta}_D) \geq c > 0$ pour tout n, $\hat{\theta}_N - \hat{\theta}_D = \mathrm{O}(n^{-1})$. Et puisque $b_D = b_N$, cela implique $\hat{b}_N^{(i)} - \hat{b}_D^{(i)} = \mathrm{O}(n^{-1})$ $(i = 0, 1, 2)$. Les termes additionnels dans le Lemme 6.9 peuvent donc être ignorés. □

Le Corollaire 6.10 montre clairement l'avantage de l'interprétation exponentielle complète de (6.14), qui évite le calcul des dérivées première et deuxième, $\hat{g}'$ et $\hat{g}''$, apparaissant dans (6.15). Notons que (6.16) peut aussi s'écrire

$$\mathbb{E}^\pi[g(\theta)|x] = \frac{\sigma_N^2}{\sigma_D^2}\frac{g(\hat{\theta}_N)f(x|\hat{\theta}_N)\pi(\hat{\theta}_N)}{f(x|\hat{\theta}_D)\pi(\hat{\theta}_D)} + \mathrm{O}(n^{-2}).$$

L'hypothèse sur g, à savoir que g est positive et bornée, en $\hat{\theta}_D$, par des constantes strictement positives, est cependant assez restrictive. En effet, la décomposition habituelle $g = g^+ - g^-$ ne marche pas dans ce cadre. Tierney *et al.* (1989) surmontent cet inconvénient en évaluant d'abord la fonction génératrice des moments de $g(\theta)$,

$$M(s) = \mathbb{E}^\pi[\exp\{sg(\theta)\}|x],$$

bien entendu positive, par $\hat{M}(s)$ via (6.16). Ils calculent $\mathbb{E}^\pi[g(\theta)|x]$ comme

$$\mathbb{E}^\pi[g(\theta)|x] = \frac{d}{ds}(\log\hat{M}(s))\big|_{s=0} + \mathrm{O}(n^{-2}).$$

Ces auteurs ont aussi établi le résultat plutôt surprenant que cette approche fournit le développement standard (6.15) sans nécessiter une évaluation des première et deuxième dérivées de g (voir l'Exercice 6.18).

Exemple 6.11. (Tierney *et al.*, 1989) Soit $\pi(\theta|x)$ une loi $\mathscr{B}e(\alpha, \beta)$; l'espérance a posteriori de θ est alors

$$\delta^\pi(x) = \frac{\alpha}{\alpha+\beta}.$$

Ce calcul exact peut être comparé aux approximations (6.15),

$$\delta^\pi(x) = \frac{\alpha^2 + \alpha\beta + 2 - 4\alpha}{(\alpha+\beta-2)^2} + \mathrm{O}((\alpha+\beta)^{-2}),$$

et (6.16),

$$\delta^\pi(x) = \frac{\alpha}{\alpha+\beta-1}\left(\frac{\alpha}{\alpha-1}\right)^{\alpha-0.5}\left(\frac{\alpha+\beta-2}{\alpha+\beta-1}\right)^{\alpha+\beta-0.5} + \mathrm{O}((\alpha+\beta)^{-2}).$$

Notant $p = \alpha/(\alpha+\beta)$ et $n = \alpha+\beta$, l'erreur d'approximation est

$$\Delta^S = 2\frac{1-2p}{n^2} + \mathrm{O}(n^{-3})$$

dans le cas standard, et

$$\Delta^E = 2\frac{1-13p^2}{12pn^2} + \mathrm{O}(n^{-3})$$

dans le cas exponentiel complet. Le deuxième développement est alors meilleur pour les valeurs moyennes de p. ‖

Nous renvoyons les lecteurs à Leonard (1982), Tierney et Kadane (1986), Tierney *et al.* (1989) et Kass et Steffey (1989) pour des résultats additionnels et des commentaires. Une réserve faite dans Smith *et al.* (1985) sur les approximations de Laplace est qu'elles ne sont justifiées que de *façon asymptotique* ; les vérifications spécifiques menées dans différentes publications ne peuvent fournir une justification globale de la méthode, même si elles semblent donner des résultats assez satisfaisants dans la plupart des cas. D'autres critiques de cette approche sont que

(1) les méthodes analytiques impliquent toujours des études préliminaires délicates sur la régularité de la fonction intégrée, ce qui n'est pas forcément faisable ;

(2) la loi a posteriori doit être assez semblable à la loi normale (pour laquelle l'approximation de Laplace est exacte) ; et

(3) de telles méthodes ne peuvent pas être utilisées dans des cas comme ceux de l'Exemple 6.5, où le calcul de l'estimateur du maximum de vraisemblance est assez difficile.

Des extensions de la méthode de Laplace à des *approximations de point-selle* sont passées en revue dans Kass (1989) (voir aussi Rousseau, 1997, 2000).

6.3 Méthodes de Monte Carlo par chaînes de Markov

Nous considérons dans cette section une méthode de Monte Carlo plus générale, permettant d'approcher la génération de variables aléatoires d'une loi a posteriori $\pi(\theta|x)$ lorsque cette loi ne peut pas être simulée directement. L'avantage de cette méthode sur les méthodes de Monte Carlo classiques décrites dans la Section 6.2.2 est qu'elle ne nécessite pas la construction précise

d'une fonction d'importance, puisqu'elle prend en compte les caractéristiques de $\pi(\theta|x)$. Cette extension, appelée *Monte Carlo par chaînes de Markov* (et abrégée en MCMC), a des applications presque illimitées, même si ses performances varient largement, selon la complexité du problème. Elle tire son nom de l'idée que, pour produire des approximations acceptables d'intégrales et d'autres fonctions dépendant d'une loi d'intérêt, il suffit de générer une *chaîne de Markov* $(\theta^{(m)})_m$ de loi limite la loi d'intérêt[53]. Cette idée d'utiliser le comportement limite d'une chaîne de Markov apparaît à la même époque que la technique de Monte Carlo originelle, au moins dans la littérature de Physique particulaire (Metropolis *et al.*, 1953), mais elle nécessite une puissance de calcul qui n'était alors pas suffisament grande pour être appréciée dans sa globalité.

Après une brève discussion sur l'intérêt de l'utilisation d'une chaîne de Markov en simulation (Section 6.3.1), nous présenterons les deux types de techniques les plus importantes conçues pour créer des chaînes de Markov de loi stationnaire donnée, à savoir les algorithmes de Metropolis-Hastings (Section 6.3.2) et l'échantillonnage de Gibbs (Sections 6.3.3-6.3.6). Nous renvoyons les lecteurs à Gilks *et al.* (1996) et Robert et Casella (2004) pour des perspectives plus larges sur ce sujet.

6.3.1 Les MCMC en pratique

Le paradoxe apparent d'une simulation par chaînes de Markov est qu'il semble que nous devions recourir *deux fois* à un argument asymptotique : premièrement, la chaîne doit converger vers sa loi stationnaire ; deuxièmement, des moyennes empiriques comme (6.9) doivent converger vers l'espérance correspondante $\mathbb{E}^\pi[g(\theta)|x]$. Nous expliquons maintenant pourquoi, grâce au Théorème Ergodique, ceci n'est pas le cas.

Si les chaînes de Markov $(\theta^{(m)})_m$ produites par des algorithmes MCMC sont irréductibles, c'est-à-dire si elles peuvent visiter (avec probabilité non nulle) tout ensemble A tel que $\pi(A|x) > 0$, alors, de par leur nature même, ces chaînes sont *récurrentes positives,* de loi stationnaire $\pi(\theta|x)$, c'est-à-dire que le nombre moyen de visites d'un ensemble arbitraire A de mesure positive est infini. Ces chaînes de Markov sont aussi *ergodiques*, ce qui signifie que la loi de $\theta^{(m)}$ converge vers $\pi(\cdot|x)$ pour presque toute valeur initiale $\theta^{(0)}$; en d'autres termes, l'influence de la valeur initiale disparaît. (Sous des conditions assez générales, les chaînes MCMC sont même *récurrentes au sens de Harris*, ce qui implique que le "presque" ci-dessus disparaît.)

[53] Cette section minimise le recours à la théorie des chaînes de Markov, bien que certaines notions comme l'*ergodicité* ne puissent pas être omises. Nous renvoyons les lecteurs à Meyn et Tweedie (1993) pour une introduction profonde et pédagogique sur ce sujet. Voir aussi Robert et Casella (1999, Chapitre 4) pour un traitement plus expéditif de ces notions, nécessaires pour la compréhension des méthodes MCMC.

Par conséquent, pour k suffisamment grand, le $\theta^{(k)}$ résultant est distribué approximativement selon $\pi(\theta|x)$, quelle que soit la valeur initiale $\theta^{(0)}$. Dans la pratique, le problème est alors de déterminer ce que signifie un "grand" k, car il détermine le nombre de simulations à effectuer : s'agit-il de 200 ou 10^{10} simulations ? La vitesse de convergence, c'est-à-dire le taux de décroissance de la différence (distance) entre la loi de $\theta^{(k)}$ et sa limite, apporte une réponse à ce problème, mais jusqu'ici elle a été surtout étudiée d'un point de vue théorique (voir Roberts et Tweedie, 2005). De plus, ce taux de convergence dépend souvent du point de départ (sauf si la chaîne est *uniformément ergodique*) et un nombre k d'itérations donné ne fournit pas la même qualité d'approximation pour différentes valeurs de $\theta^{(0)}$. Il existe donc des obstacles pratiques à la simulation par chaînes de Markov, puisqu'on ignore le plus souvent si la chaîne a itéré suffisamment longtemps. Mais, comme l'ont détaillé Robert et Casella (2004, Chapitre 12), il existe désormais des tests de diagnostic et un logiciel correspondant, `CODA` (voir la Note 6.6.2), qui fournissent différents indicateurs de stationnarité de la chaîne et limitent en partie cette difficulté.

Une fois $\theta_1 = \theta^{(k)}$ généré, une façon naïve de construire un échantillon iid $\theta_1, \ldots, \theta_m$ suivant $\pi(\theta|x)$ est d'utiliser le même algorithme avec une autre valeur initiale $\theta_2^{(0)}$ et une autre séquence de k transitions de Markov afin d'obtenir θ_2, et ainsi de suite jusqu'à θ_m. Comme nous l'avons montré ci-dessous, la vitesse de convergence dépend souvent de la valeur initiale, et il est donc préférable (en termes de convergence) de prendre la valeur actuelle $\theta^{(k)}$ comme nouvelle valeur initiale, même si cela introduit de la dépendance entre les θ_i. Cependant, l'indépendance n'est pas fondamentale lorsqu'on s'intéresse principalement à des fonctionnelles de $\pi(\theta|x)$, car le *Théorème Ergodique* implique que la moyenne

$$\frac{1}{K}\sum_{k=1}^{K} g(\theta^{(k)})$$

converge vers $\mathbb{E}^{\pi}[g(\theta)|x]$ (du moment que $\mathbb{E}^{\pi}[|g(\theta)||x]$ est fini) lorsque K tend vers l'infini (voir Meyn et Tweedie, 1993). L'influence de la valeur de départ disparaît donc aussi dans la moyenne (d'où l'ergodicité). De plus, cette propriété est aussi satisfaite par toute sous-suite de $(\theta^{(k)})$.

Le Théorème Ergodique résout donc le paradoxe des deux asymptotiques mentionné au début de cette section, car il étend la Loi des Grands Nombres à des suites dépendant de variables aléatoires et supprime le besoin de produire un échantillon iid, qui serait, de toute manière, seulement approximatif si nous utilisions la méthode proposée ci-dessus. En effet, comme l'a noté Geyer (1992), la théorie des chaînes de Markov ne donne pas d'indication générale sur le fait que la stationnarité soit atteinte, car, d'un point de vue mathématique, ceci n'est qu'une propriété asymptotique de la chaîne[54]. Par conséquent, il vaut mieux considérer une *seule suite* $(\theta^{(k)})$, puisque chaque

[54]Une exception est fournie par les cas du *renouvellement* et de l'échantillonnage *exact* (voir Robert et Casella, 2004, Chapitre 13), où il est possible d'exhiber des k

étape de simulation nous rapproche (en probabilité) d'une réalisation de la loi stationnaire, $\pi(\theta|x)$. De plus, une simulation reposant sur de multiples points de départ entraîne un gaspillage considérable, puisque la plupart des valeurs simulées sont rejetées. Cependant, le recours à des chaînes multiples est assez utile pour l'étude de la convergence d'une chaîne de Markov et apparaît donc fréquemment dans des techniques de contrôle, comme dans la méthode *within-between* de Gelman et Rubin (1992) (voir Robert et Casella, 2004, Section 12.3.4).

Lorsque cela est nécessaire, une quasi-indépendance peut être obtenue par échantillonnage par paquets, c'est-à-dire en ne retenant qu'un point de la chaîne toutes les t itérations, pour un échantillon simulé efficacement, avec, par exemple, $t = 5$ ou $t = 10$. Raftery et Lewis (1992a,b) proposent une détermination plus complexe de la taille du paquet t, qui est induite par la chaîne et fondée sur une "binarisation" de cette chaîne. (Voir Robert et Casella, 2004, Section 12.3.4, pour une évaluation critique de cette méthode, implémentée dans le logiciel `CODA`.)

6.3.2 Algorithmes de Metropolis-Hastings

Une fois acquis le principe d'utilisation d'une chaîne de Markov de loi stationnaire π–plutôt que des variables iid distribuées exactement selon π– pour approcher des quantités comme (6.8), la mise en œuvre de ce principe nécessite la construction d'un mécanisme de génération pour produire de telles chaînes de Markov. De façon étonnante, un algorithme quasi universel satisfaisant cette contrainte existe : il a été développé par Metropolis *et al.* (1953), au départ pour la Physique particulaire (et la bombe H...), et généralisé par Hastings (1970) dans un cadre plus statistique (et plus pacifique). En réalité, il s'applique à une grande variété de problèmes, car sa principale restriction est que la loi d'intérêt soit connue à une constante près, mais nous verrons plus tard que cette contrainte peut être levée de plusieurs façons.

Dans sa version moderne, l'*algorithme de Metropolis-Hastings* peut être décrit de la façon suivante. Pour une densité donnée $\pi(\theta)$, connue à un facteur de normalisation près, et une densité conditionnelle $q(\theta'|\theta)$, l'algorithme génère la chaîne $(\theta^{(m)})_m$ comme suit :

ALGORITHME 6.1. –**Algorithme de Metropolis-Hastings**–

Itération 0 : Initialiser avec une valeur arbitraire $\theta^{(0)}$

Itération m : Mettre à jour $\theta^{(m)}$ par $\theta^{(m+1)}$ $(m = 1, 2, \ldots)$, de la façon suivante :

a) Générer $\xi \sim q(\xi|\theta^{(m)})$

tels que $\theta^{(k)}$ soit exactement distribué suivant la loi stationnaire. Voir aussi Hobert et Robert (2004).

b) Poser

$$\varrho(\theta^{(m)},\xi) = \frac{\pi(\xi)\, q(\theta^{(m)}|\xi)}{\pi(\theta^{(m)})\, q(\xi|\theta^{(m)})} \wedge 1$$

c) Prendre

$$\theta^{(m+1)} = \begin{cases} \xi & \text{avec probabilité } \varrho(\theta^{(m)},\xi), \\ \theta^{(m)} & \text{sinon.} \end{cases}$$

La loi de densité $\pi(\theta)$ est souvent appelée *loi cible* ou *loi objet*, tandis que la loi de densité $q(\cdot|\theta)$ est dite *loi de proposition*. Une propriété stupéfiante de cet algorithme est d'autoriser un nombre infini de lois de proposition produisant toutes une chaîne de Markov convergeant vers la loi d'intérêt.

Théorème 6.12. *Si la chaîne $(\theta^{(m)})_m$ est irréductible, c'est-à-dire si, pour tout sous-ensemble A tel que $\pi(A) > 0$, il existe M tel que $P_{\theta^{(0)}}(\theta^{(M)} \in A) > 0$, alors π est la loi stationnaire de la chaîne. Si de plus la chaîne est apériodique, elle est aussi ergodique de loi limite π, pour presque toute valeur initiale $\theta^{(0)}$, au sens où*

$$\lim_{m\to\infty} \sup_A \left| P_{\theta^{(0)}}(\theta^{(m)} \in A) - \pi(A) \right| = 0 \qquad (\pi\, p.s.)$$

La propriété au cœur de ce résultat est la *condition d'équilibre ponctuel*, c'est-à-dire le fait que le noyau de transition de la chaîne de Markov associée à l'algorithme ci-dessus, noté $K(\theta'|\theta)$, satisfasse

$$\pi(\theta)K(\theta'|\theta) = \pi(\theta')K(\theta|\theta')\,, \tag{6.17}$$

ce qui se vérifie aisément en écrivant le noyau de l'algorithme de Metropolis-Hastings

$$K(\theta'|\theta) = \varrho(\theta,\theta')q(\theta'|\theta) + \int [1-\varrho(\theta,\xi)]q(\xi|\theta)d\xi\; \delta_\theta(\theta')\,,$$

où δ est la masse de Dirac.

Lorsqu'on intègre les deux côtés de (6.17) en θ, le terme de droite donne $\pi(\theta')$, car $K(\theta|\theta')$ est une densité (conditionnelle) en θ ; le terme de gauche donne la densité de la chaîne de Markov après une étape, lorsque $\theta^{(0)} \sim \pi$. Par conséquent, la loi π est bien stationnaire pour le noyau de transition $K(\theta'|\theta)$. (Voir l'Exercice 6.20 et Robert et Casella, 2004, Section 6.2, pour plus de détails.)

La condition d'irréductibilité du Théorème 6.12 est bien entendu une condition nécessaire pour que la chaîne explore le support de π. Des conditions suffisantes pour l'irréductibilité sont, par exemple, que le support de $q(\cdot|\theta)$ contienne le support de π pour tout θ ou, plus généralement, que le

support de $q(\cdot|\theta)$ contienne un voisinage de θ de rayon constant (voir Robert et Casella, 1999, Lemme 6.2.7).

Tandis que le Théorème 6.12 donne une condition formelle pour que la chaîne converge, ce qui couvre une immense catégorie de lois proposées, la sélection pratique de cette loi est beaucoup plus délicate, car un faible chevauchement entre les supports de π et $q(\cdot|\theta)$ peut considérablement ralentir la convergence.

Exemple 6.13. Les *lois de Weibull* sont utilisées abondamment en fiabilité et dans d'autres applications en ingénierie, en partie à cause de leur capacité à décrire différents comportements de taux de risque et en partie pour des raisons historiques. Puisqu'elles n'appartiennent à aucune famille exponentielle, étant de la forme

$$f(x) \propto \alpha\eta x^{\alpha-1} e^{-x^\alpha \eta}, \tag{6.18}$$

elles ne peuvent pas conduire à des lois a posteriori explicites pour les paramètres α et η. Pour $\theta = (\alpha, \eta)$, considérons la loi a priori (propre)

$$\pi(\theta) \propto e^{-\alpha} \eta^{\beta-1} e^{-\xi\eta}$$

et des observations $x_1, \ldots, x_n$ de (6.18). Un algorithme de Metropolis-Hastings pour la simulation de $\pi(\theta|x_1, \ldots, x_n)$ peut se fonder sur la loi conditionnelle

$$q(\theta'|\theta) = \frac{1}{\alpha\eta} \exp\left(-\frac{\alpha'}{\alpha} - \frac{\eta'}{\eta}\right),$$

c'est-à-dire sur deux lois exponentielles indépendantes de moyennes α et η, versions exponentielles des marches aléatoires (voir ci-dessous). La probabilité d'acceptation résultante est alors

$$\varrho = 1 \wedge \left(\frac{\eta'}{\eta}\right)^\beta \frac{\alpha'}{\alpha} \left(\prod_{i=1}^n x_i\right)^{\alpha'-\alpha} \prod_{i=1}^n e^{x_i^\alpha - x_i^{\alpha'}} e^{-\alpha/\alpha' - \eta/\eta' + \alpha'/\alpha + \eta'/\eta},$$

si $(\alpha', \eta') = \theta'$ est la valeur simulée et $(\alpha, \eta) = \theta$ est la valeur courante des paramètres. ‖

Dès Hastings (1970), le choix le plus courant pour q est une *marche aléatoire*, où $q(\theta'|\theta)$ est de la forme $f(||\theta' - \theta||)$. La valeur proposée ξ dans l'algorithme de Metropolis-Hastings est alors de la forme

$$\xi = \theta^{(m)} + \varepsilon\,,$$

où ε est une variable aléatoire de loi symétrique f. L'idée naturelle sur laquelle repose ce choix est de perturber aléatoirement la valeur courante de la chaîne, tout en restant aux alentours de ce point, et de voir si la nouvelle valeur ξ est vraisemblable pour la loi d'intérêt. Pour ce mécanisme de proposition en marche aléatoire, le rapport d'acceptation de Metropolis-Hastings est

$$\varrho = \frac{\pi(\xi)}{\pi(\theta^{(m)})} \wedge 1.$$

La chaîne $(\theta^{(m)})_m$ restera donc plus longtemps en un point donné ξ si la valeur a posteriori correspondante $\pi(\xi)$ est supérieure et, inversement, des points ξ tels que $\pi(\xi) = 0$ ne seront jamais visités. Des choix standard pour q sont les lois uniformes, normales ou de Cauchy. (Notons que l'Exemple 6.13 est bien un cas particulier de l'algorithme de Metropolis-Hastings à marche aléatoire, car la proposition est une marche aléatoire en $(\log\alpha, \log\eta)$.)

Exemple 6.14. Pour $\theta, x \in \mathbb{R}^2$, soit la loi normale modifiée

$$\pi(\theta|x) \propto \exp\{-||\theta - x||^2/2\} \prod_{i=1}^{p} \exp\left\{\frac{-1}{||\theta - \mu_i||^2}\right\},$$

où les μ_i agissent comme des points répulsifs, c'est-à-dire des valeurs improbables (ou interdites) de θ. Un algorithme de Metropolis-Hastings à marche aléatoire fondé sur une proposition $\mathcal{N}_2(0, 0.2\, I_2)$ conduit au résultat représenté par la Figure 6.2 pour $x = 0$ et $p = 15$. Les μ_j, qui sont représentés par des croix, sont correctement évités par la chaîne de Markov, qui retrouve aussi la forme de la densité normale. ‖

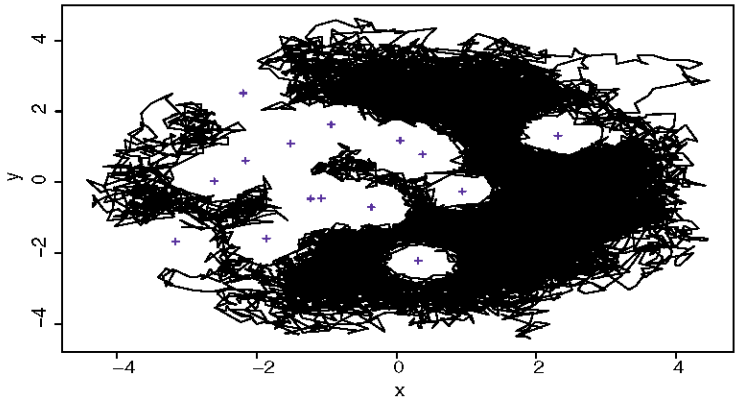

Fig. 6.2. Trajet de la chaîne de Markov $(\theta^{(m)})_m$ sur la surface a posteriori de $\pi(\theta|x)$ dans l'Exemple 6.14 et les points répulsifs μ_j indiqués par des croix, pour $x = 0$ et $p = 15$ (5 000 itérations).

Clairement, cet algorithme est applicable dans une grande généralité et, de plus, a des contraintes de calibration limitées, car la loi des perturbations peut

être choisie de façon quasi indépendante de la vraie densité π. (En effet, cette loi dépend d'un facteur d'échelle qui devrait seulement être réglé en fonction du taux d'acceptation moyen de l'algorithme[55] ; voir Robert et Casella, 2004, Section 7.5 et Note 7.8.4.) Bien qu'il ne puisse pas vérifier des propriétés de convergence plus fortes que la *convergence géométrique* à cause des propriétés de queues longues du mécanisme de proposition (voir Mengersen et Tweedie, 1996), l'algorithme de Metropolis-Hastings à marche aléatoire apparaît encore comme étant le "passe-partout" des techniques MCMC.

Un autre type de mécanisme de proposition, ressemblant plus aux techniques de Monte Carlo standard, est *le mécanisme indépendant*, où la densité $q(\cdot|\theta)$ ne dépend pas de θ,

$$q(\theta'|\theta) = h(\theta') \,.$$

(Puisque la valeur proposée peut être rejetée avec une probabilité positive, l'algorithme produit néanmoins une chaîne de Markov.) Bien que leurs propriétés théoriques soient souvent meilleures que celles de l'algorithme de Metropolis-Hastings à marche aléatoire (voir Mengersen et Tweedie, 1996), ces méthodes ont des applications plus limitées, car le mécanisme de proposition h doit ressembler dans un certain sens à la loi cible π. La loi proposée est parfois la loi a priori ou est fondée sur un développement asymptotique de la loi π, par exemple, une *approximation par point-selle* (Robert et Casella, 1999, Exemple 6.3.4) ou sur un algorithme d'acceptation-rejet approximatif comme dans l'algorithme ARMS de Gilks *et al.* (1995) (voir aussi l'Exercice 6.12). (Notons la similitude avec la méthode d'échantillonnage d'importance de la Section 6.2.2 : le choix du mécanisme de proposition de la loi h est fondamental pour la mise en œuvre pratique de la méthode.)

6.3.3 L'échantillonnage de Gibbs

La technique de Metropolis-Hastings présentée dans la section précédente est attrayante de par son universalité, mais, d'un autre côté, le manque de connexion entre le mécanisme de proposition q et la loi cible π peut être néfaste pour les propriétés de convergence de la méthode et, dans la pratique, peut facilement empêcher la convergence si la probabilité d'atteindre des parties éloignées du support de la loi π est trop petite. L'approche de l'*échantillonnage de Gibbs*, qui repose sur une perspective différente, est pour sa part fondée sur la loi π. Cette méthode tire son nom des champs aléatoires de Gibbs, où elle a été utilisée pour la première fois par Geman et Geman (1984) ; voir Robert et Casella (2004, Note 10.6.1) pour un bref compte-rendu des débuts de l'échantillonnage de Gibbs.

[55] Le facteur d'échelle 0.02 dans l'Exemple 6.14 a délibérément été choisi trop petit, pour mieux illustrer la façon dont la chaîne de Markov évite les points répulsifs μ_i. En pratique, un facteur d'échelle petit peut conduire à des problèmes d'irréductibilité si la chaîne de Markov n'arrive pas à franchir des zones de très faible probabilité pour joindre deux régions (modales) de forte probabilité.

D'un point de vue général, l'échantillonnage de Gibbs tire profit des *structures hiérarchiques* d'un modèle, par exemple lorsque celui-ci peut s'écrire sous la forme

$$\pi(\theta|x) = \int \pi_1(\theta|x,\lambda)\pi_2(\lambda|x)\,d\lambda. \tag{6.19}$$

L'idée est alors de simuler la loi jointe $\pi_1(\theta|x,\lambda)\pi_2(\lambda|x)$, afin d'obtenir $\pi(\theta|x)$ comme la loi marginale. Bien entendu, lorsque les deux lois $\pi_1(\theta|x,\lambda)$ et $\pi_2(\lambda|x)$ sont connues et peuvent être simulées, la génération de θ de $\pi(\theta|x)$ est équivalente à la génération de λ de $\pi_2(\lambda|x)$, puis de θ de $\pi_1(\theta|x,\lambda)$.

Exemple 6.15. (Casella et George, 1992) Soit $(\theta,\lambda) \in \mathbb{N} \times [0,1]$ et

$$\pi(\theta,\lambda|x) \propto \binom{n}{\theta} \lambda^{\theta+\alpha-1}(1-\lambda)^{n-\theta+\beta-1},$$

où les paramètres α et β dépendent en réalité de x. Ce modèle peut s'écrire de façon hiérarchique (6.19), avec $\pi_1(\theta|x,\lambda)$ une loi binomiale, $\mathscr{B}(n,\lambda)$, et $\pi_2(\lambda|x)$ une loi bêta, $\mathscr{B}e(\alpha,\beta)$. La loi marginale de θ est alors

$$\pi(\theta|x) = \binom{n}{\theta} \frac{B(\alpha+\theta,\beta+n-\theta)}{B(\alpha,\beta)},$$

c'est-à-dire une *loi bêta-binomiale*. Cette loi marginale n'est pas particulièrement facile à utiliser. Par exemple, le calcul de $\mathbb{E}[\theta/(\theta+1)|x]$, ou de la loi a posteriori de $\eta = \exp(-\theta^2)$, ne peut pas être fait explicitement et peut nécessiter des approximations numériques complexes lorsque α, β et n sont grands. Par conséquent, en fonction du problème inférentiel, il peut être plus avantageux de tirer profit de la décomposition hiérarchique ci-dessus et de simuler $(\lambda^{(1)},\theta^{(1)}),\ldots,(\lambda^{(m)},\theta^{(m)})$ avec $\lambda^{(i)} \sim \mathscr{B}e(\alpha,\beta)$ et $\theta^{(i)} \sim \mathscr{B}(n,\lambda^{(i)})$; par exemple, $\mathbb{E}[\theta/(\theta+1)|x]$ peut alors être approchée par

$$\frac{1}{m}\sum_{i=1}^{m} \frac{\theta^{(i)}}{\theta^{(i)}+1}.$$

(On remarquera que, dans ce cas, l'utilisation d'un algorithme MCMC n'est pas utile.) ‖

Cependant, et par contraste avec l'Exemple 6.15, la loi marginale $\pi_2(\lambda|x)$ n'est pas toujours disponible (sous forme analytique ou algorithmique) et la méthode classique de Monte Carlo par simulation directe ne peut pas être mise en œuvre. Il est en fait plus fréquent que les deux *lois a posteriori conditionnelles*, $\pi_1(\theta|x,\lambda)$ et $\pi_2(\lambda|x,\theta)$, puissent être simulées. Puisqu'elles sont suffisamment informatives sur la loi jointe, $\pi(\theta,\lambda|x)$, et puisque $\pi(\theta,\lambda|x)$ peut être obtenu à partir de ces densités conditionnelles (voir les Exercices 6.26 et 6.27), il semble conceptuellement possible de fonder un algorithme de simulation de $\pi(\theta|x)$ sur ces lois conditionnelles uniquement.

Exemple 6.16. (Suite de l'Exemple 6.4) Pour le modèle de capture-recapture temporel, les deux lois a posteriori conditionnelles sont $(1 \leq i \leq n)$

$$p_i|x, N \sim \mathscr{B}e(\alpha + x_i, \beta + N - x_i)$$
$$N - x_+|x, p \sim \mathscr{N}eg(x_+, \varrho),$$

avec

$$\varrho = 1 - \prod_{i=1}^{n}(1 - p_i).$$

En revanche, la loi marginale a posteriori $\pi_2(p|x)$ ne peut pas être obtenue explicitement ou simulée directement. ‖

Une première technique d'*échantillonnage de Gibbs*, d'abord appelée *augmentation des données* parce que utilisée dans ce contexte, a été introduite par Tanner et Wong (1987) afin de tirer profit des lois conditionnelles selon l'algorithme itéré suivant :

ALGORITHME 6.2. –**Échantillonnage de Gibbs bivarié**–
Initialisation : Commencer par une valeur arbitraire $\lambda^{(0)}$.
Itération t **:** pour $\lambda^{(t-1)}$ donné, générer
a. $\theta^{(t)}$ selon $\pi_1(\theta|x, \lambda^{(t-1)})$
b. $\lambda^{(t)}$ selon $\pi_2(\lambda|x, \theta^{(t)})$.

Il est alors simple de montrer que $\pi(\theta, \lambda|x)$ est la loi stationnaire de la transition ci-dessus : si $(\theta^{(i-1)}, \lambda^{(i-1)})$ est distribué selon la loi jointe, $\lambda^{(i-1)}$ est distribué selon la loi marginale $\pi_2(\lambda|x)$ et, par conséquent, $(\theta^{(i)}, \lambda^{(i-1)})$ est toujours distribué selon la loi jointe. (En réalité, il faut s'assurer que le support de la loi jointe soit égal au produit cartésien des supports de π_1 et π_2 ; voir Robert et Casella, 2004, Exemple 10.7, pour un contre-exemple.) Le même raisonnement s'applique à la deuxième étape de l'algorithme et la chaîne $(\theta^{(t)}, \lambda^{(t)})$ est ergodique de loi limite π. De plus, la structure duale de l'algorithme ci-dessus conduit à de bonnes propriétés de convergence, comme l'ont montré Diebolt et Robert (1994) :

Lemme 6.17. *Si $\pi_1(\theta|x, \lambda) > 0$ sur Θ ($\pi_2(\lambda|x, \theta) > 0$ sur Λ, respectivement), les deux suites $(\theta^{(m)})$ et $(\lambda^{(m)})$ sont des chaînes de Markov ergodiques de lois invariantes $\pi(\theta|x)$ et $\pi(\lambda|x)$, respectivement.*

De plus, on peut montrer que, si la convergence est uniformément géométrique pour une des deux chaînes, par exemple si elle prend ses valeurs dans un espace fini, la convergence vers la loi stationnaire est aussi uniformément géométrique pour l'*autre* chaîne. Cette propriété est connue sous le nom de *principe de dualité* (voir l'Exercice 6.28).

Exemple 6.18. (Suite de l'Exemple 6.15) Les lois conditionnelles sont

$$\theta|x,\lambda \sim \mathscr{B}(n,\lambda), \qquad \lambda|x,\theta \sim \mathscr{B}e(\alpha+\theta,\beta+n-\theta)$$

et rendent possible la mise en œuvre de l'échantillonnage de Gibbs, même s'il n'est pas nécessaire dans ce contexte. La Figure 6.3 donne une comparaison de l'histogramme d'un échantillon de cinq mille observations obtenues par échantillonnage par paquets (avec $t = 10$), et l'histogramme d'un échantillon de cinq mille observations θ simulées directement de la loi bêta-binomiale. La forte ressemblance entre les deux montre que l'approximation par l'échantillonnage de Gibbs est tout à fait acceptable. ‖

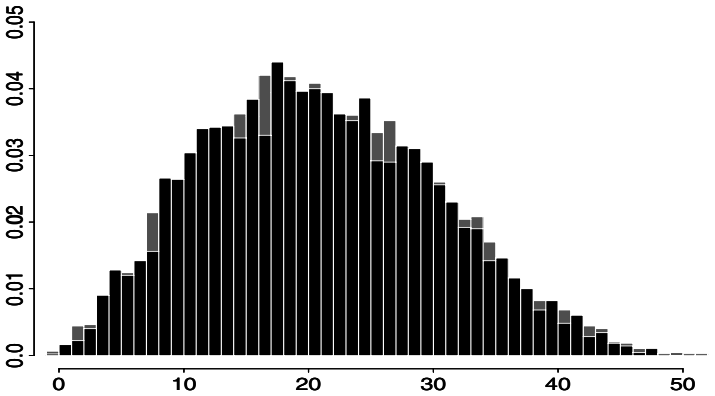

Fig. 6.3. Histogrammes d'échantillons de taille 5 000 de la loi bêta-binomiale de paramètres $n = 54$, $\alpha = 3.4$, et $\beta = 5.2$: *(gris foncé)* simulé directement ; *(gris clair)* obtenu par échantillonnage de Gibbs.

6.3.4 Rao-Blackwellisation

Comme nous avons discuté dans la Section 6.3.1, l'échantillon $\theta^{(1)},\ldots,\theta^{(m)}$ produit par échantillonnage de Gibbs peut être utilisé de la même façon que celui obtenu par la méthode classique de Monte Carlo, mais Gelfand et Smith (1990) remarquent que la structure conditionnelle de l'algorithme d'échantillonnage et l'échantillon dual, $\lambda^{(1)},\ldots,\lambda^{(m)}$, devraient être exploités. En effet, si la quantité d'intérêt est $\mathbb{E}^\pi[g(\theta)|x]$, on peut utiliser la moyenne des espérances conditionnelles

$$\delta_2 = \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}^\pi[g(\theta)|x,\lambda^{(m)}],$$

lorsque celles-ci peuvent être calculées facilement, plutôt que d'utiliser la moyenne directe

$$\delta_1 = \frac{1}{m}\sum_{i=1}^{m} g(\theta^{(i)}).$$

Cette modification est fondée sur le théorème de Rao-Blackwell (voir le Théorème 2.20). Si les $\lambda^{(i)}$ et les $\theta^{(i)}$ étaient indépendants,

$$\begin{aligned}\mathbb{E}^{\pi}\left[(\delta_1 - \mathbb{E}^{\pi}[g(\theta)|x])^2|x\right] &= \frac{1}{m}\mathrm{var}^{\pi}(g(\theta)|x)\\ &\geq \frac{1}{m}\mathrm{var}^{\pi}\left(\mathbb{E}^{\pi}[g(\theta)|x,\lambda]|x\right)\\ &= \mathbb{E}^{\pi}\left[(\delta_2 - \mathbb{E}^{\pi}[g(\theta)|x,\lambda])^2\big|x\right].\end{aligned}$$

Liu *et al.* (1994) montrent que cette inégalité est aussi toujours vérifiée pour l'Algorithme 6.2 de Gibbs bivarié, car $\mathrm{cov}(\theta^{(0)},\theta^{(m)})$ est alors positive et décroît en m (Exercice 6.30). L'estimateur δ_2, baptisé *Rao-Blackwellisation*, domine donc δ_1. (Mais cette domination ne s'étend pas nécessairement à d'autres techniques MCMC, voir Liu *et al.*, 1995, et Geyer, 1995.)

Exemple 6.19. (Casella et George, 1992) Soient les lois conditionnelles suivantes (x est omis des notations) :

$$\begin{aligned}\pi(\theta|\lambda) &\propto \lambda e^{-\theta\lambda}, \qquad 0<\theta<B,\\ \pi(\lambda|\theta) &\propto \theta e^{-\lambda\theta}, \qquad 0<\lambda<B.\end{aligned}$$

La loi marginale de θ (ou de λ) ne peut pas être calculée, mais les lois conditionnelles sont faciles à simuler, car ce sont des exponentielles tronquées. Puisque $\mathbb{E}^{\pi}[\theta|\lambda] \simeq 1/\lambda$ pour B grand, $\mathbb{E}^{\pi}[\theta|x]$ peut être approché par

$$\frac{1}{m}\sum_{i=1}^{m}\theta_i \qquad \text{ou} \qquad \frac{1}{m}\sum_{i=1}^{m}\frac{1}{\lambda_i}.$$

Pour cet exemple particulier, la symétrie complète entre les deux lois conditionnelles implique que les deux estimateurs ont exactement les mêmes propriétés probabilistes, en plus de converger vers la même valeur. ‖

Le même argument nous conduit à proposer l'approximation de la densité a posteriori $\pi(\theta|x)$ par la moyenne des densités conditionnelles

$$\frac{1}{m}\sum_{i=1}^{m}\pi(\theta|x,\lambda_i),$$

plutôt que par les méthodes *d'estimation non paramétrique par noyau* standard (voir Tanner et Wong, 1987, et Gelfand et Smith, 1990).

6.3.5 L'échantillonnage de Gibbs général

Une généralisation de l'Algorithme 6.2 de Gibbs bivarié consiste à considérer plusieurs groupes de paramètres, $\theta, \lambda_1, \ldots, \lambda_p$, tels que

$$\pi(\theta|x) = \int \ldots \int \pi(\theta, \lambda_1, \ldots, \lambda_p|x)\, d\lambda_1 \cdots d\lambda_p\,. \tag{6.20}$$

Cette généralisation correspond par exemple à l'introduction de niveaux additionnels dans le modèle hiérarchique (6.19), pour des raisons de modélisation ou de simulation, ou de décomposition de l'hyperparamètre λ ou du paramètre θ en des composantes de plus petites dimensions.

Comme expliqué dans la Section 6.3.3 à propos du procédé de Gibbs bivarié, l'échantillonnage de Gibbs fournit des simulations de la loi jointe $\pi(\theta, \lambda_1, \ldots, \lambda_p|x)$, lorsque certaines des lois conditionnelles associées à π sont disponibles. Bien entendu, lorsque $\pi(\theta|x)$ se décompose elle-même en lois conditionnelles, il n'y a pas besoin d'introduire des paramètres additionnels λ_i $(1 \leq i \leq p)$.

Exemple 6.20. (Suite de l'Exemple 6.15) Si la taille de la population n suit une loi a priori de Poisson, $\mathscr{P}(\xi)$, la loi a posteriori jointe est

$$\pi(\theta, \lambda, n|x) \propto \binom{n}{\theta} \lambda^{\theta+\alpha-1}(1-\lambda)^{n-\theta+\beta-1} e^{-\xi}\frac{\xi^n}{n!}$$

et la loi marginale de θ ne peut pas être calculée. En revanche, les lois conditionnelles complètes ont des expressions explicites, car

$$\begin{aligned} \theta|x, \lambda, \xi &\sim \mathscr{B}(n, \lambda), \\ \lambda|x, \theta, \xi &\sim \mathscr{B}e(\theta+\alpha, n-\theta+\beta), \\ n-\theta|x, \theta, \lambda &\sim \mathscr{P}(\xi(1-\lambda)). \end{aligned}$$

La simulation de ces trois lois conditionnelles est donc possible. ‖

Exemple 6.21. (Tanner et Wong, 1987) Soit un modèle multinomial

$$y \sim \mathscr{M}_5\left(n; a_1\mu + b_1, a_2\mu + b_2, a_3\eta + b_3, a_4\eta + b_4, c(1-\mu-\eta)\right),$$

paramétré par μ et η, où

$$0 \leq a_1 + a_2 = a_3 + a_4 = 1 - \sum_{i=1}^{4} b_i = c \leq 1$$

et $c, a_i, b_i \geq 0$ sont connus. Ce modèle correspond à un échantillonnage selon

$$x \sim \mathscr{M}_9(n; a_1\mu, b_1, a_2\mu, b_2, a_3\eta, b_3, a_4\eta, b_4, c(1-\mu-\eta)),$$

et à un regroupement de certaines composantes :

$$y_1 = x_1 + x_2, \quad y_2 = x_3 + x_4, \quad y_3 = x_5 + x_6, \quad y_4 = x_7 + x_8, \; y_5 = x_9.$$

Une loi a priori conjuguée pour (μ, η) et le modèle en x est la loi de Dirichlet $\mathscr{D}(\alpha_1, \alpha_2, \alpha_3)$,

$$\pi(\mu, \eta) \propto \mu^{\alpha_1 - 1} \eta^{\alpha_2 - 1} (1 - \eta - \mu)^{\alpha_3 - 1},$$

où $\alpha_1 = \alpha_2 = \alpha_3 = 1/2$ correspond à un modèle non informatif. Dans ce cadre, la loi a posteriori de (μ, η) ne peut être obtenue de façon explicite. Cependant, si nous introduisons les *données manquantes* $z = (x_1, x_3, x_5, x_7)$, qui ne sont pas observées (et donc bien manquantes), x est en relation bijective avec (y, z) et

$$\begin{aligned} \pi(\eta, \mu | y, z) &= \pi(\eta, \mu | x) \\ &\propto \mu^{z_1} \mu^{z_2} \eta^{z_3} \eta^{z_4} (1 - \eta - \mu)^{y_5 + \alpha_3 - 1} \mu^{\alpha_1 - 1} \eta^{\alpha_2 - 1}, \end{aligned}$$

où nous désignons les coordonnées de z par (z_1, z_2, z_3, z_4). Par conséquent,

$$\mu, \eta | y, z \sim \mathscr{D}(z_1 + z_2 + \alpha_1, z_3 + z_4 + \alpha_2, y_5 + \alpha_3).$$

De plus,

$$\begin{aligned} z_i | y, \mu, \eta &\sim \mathscr{B}\left(y_i, \frac{a_i \mu}{a_i \mu + b_i}\right) \qquad (i = 1, 2), \\ z_i | y, \mu, \eta &\sim \mathscr{B}\left(y_i, \frac{a_i \eta}{a_i \eta + b_i}\right) \qquad (i = 3, 4). \end{aligned}$$

En définissant $\theta = (\mu, \eta)$ et $\lambda = z$, il apparaît donc que certaines lois conditionnelles peuvent être simulées dans ce cadre. Notons que les données manquantes z n'apparaissent pas dans la formulation originelle du problème et sont peut-être artificielles, au sens où le modèle considéré ne correspond pas nécessairement à un modèle multinomial global. Cependant, ces données manquantes facilitent considérablement la simulation des θ tout en préservant leur loi marginale. D'autres modèles à données manquantes présentent le même avantage. ‖

Dans ce cadre hiérarchique général, la mise en œuvre de l'échantillonnage de Gibbs peut être faite de plusieurs façons. Si la décomposition de (θ, λ) en $(\theta, \lambda_1, \ldots, \lambda_p)$ correspond à une décomposition du modèle selon ses niveaux hiérarchiques, c'est-à-dire

$$\pi(\theta | x) = \int .. \int \pi_1(\theta | \lambda_1, x) \pi_2(\lambda_1 | \lambda_2) .. \pi_{p+1}(\lambda_p) \, d\lambda_1 \cdots d\lambda_p, \quad (6.21)$$

il semble logique de simuler selon les lois conditionnelles

$$\begin{aligned}
\pi(\theta|x,\lambda_1,\ldots,\lambda_p) &= \pi_1(\theta|\lambda_1,x),\\
\pi(\lambda_i|x,\theta,(\lambda_j)_{j\neq i}) &= \pi(\lambda_i|\lambda_{i-1},\lambda_{i+1}) \qquad (1<i<p),\\
\pi(\lambda_1|x,\theta,(\lambda_j)_{j\neq 1}) &= \pi(\lambda_1|\theta,\lambda_2),\\
\pi(\lambda_p|x,\theta,(\lambda_j)_{j\neq p}) &= \pi(\lambda_p|\lambda_{p-1}),
\end{aligned} \tag{6.22}$$

quelles que soient les dimensions de θ et λ_j (Exercice 6.32). Dans l'Exemple 6.21 notamment, (μ,η) pourrait être généré conditionnellement à (y,z) selon une loi de Dirichlet et z conditionnellement à (μ,η).

Un algorithme alternatif également proposé par Gelfand et Smith (1990) est l'échantillonneur de Gibbs *direction par direction*, qui ne prend pas en compte les divisions hiérarchiques et ne considère que les paramètres unidimensionnels, afin de les générer conditionnellement aux autres paramètres.

Exemple 6.22. (Suite de l'Exemple 6.21) Puisque

$$\begin{aligned}
\frac{\mu}{1-\eta}|y,z,\eta &\sim \mathscr{B}e(z_1+z_2+\alpha_1, y_5+\alpha_3),\\
\frac{\eta}{1-\mu}|y,z,\mu &\sim \mathscr{B}e(z_3+z_4+\alpha_2, y_5+\alpha_3),
\end{aligned}$$

cette version de l'échantillonnage de Gibbs conduit à une simulation itérative de

$$\begin{aligned}
\mu^{(t)} &\sim (1-\eta^{(t-1)})\mathscr{B}e\left(z_1^{(t-1)}+z_2^{(t-1)}+\alpha_1, y_5+\alpha_3\right),\\
\eta^{(t)} &\sim (1-\mu^{(t)})\mathscr{B}e\left(z_3^{(t-1)}+z_4^{(t-1)}+\alpha_2, y_5+\alpha_3\right),\\
z_j^{(t)} &\sim \mathscr{B}\left(y_j, \frac{a_j\mu^{(t)}}{a_j\mu^{(t)}+b_j}\right) \qquad (j=1,2),\\
z_j^{(t)} &\sim \mathscr{B}\left(y_j, \frac{a_j\eta^{(t)}}{a_j\eta^{(t)}+b_j}\right) \qquad (j=3,4).
\end{aligned} \tag{6.23}$$

La différence avec la simulation de (μ,η,z) dans l'Exemple 6.21 est donc mineure. ‖

La formulation générale de l'*algorithme d'échantillonnage de Gibbs* pour une loi jointe $\pi(\theta_1,\ldots,\theta_p)$, de lois conditionnelles complètes $\pi_1,\ldots,\pi_p$ est exposée ci-dessous.

ALGORITHME 6.3. –**Échantillonnage de Gibbs**–
Pour $(\theta_1^{(t)},\ldots,\theta_p^{(t)})$ donnés, simuler

1. $\theta_1^{(t+1)} \sim \pi_1(\theta_1|\theta_2^{(t)},\ldots,\theta_p^{(t)})$,
2. $\theta_2^{(t+1)} \sim \pi_2(\theta_2|\theta_1^{(t+1)},\theta_3^{(t)},\ldots,\theta_p^{(t)})$,

$$\vdots$$

p. $\theta_p^{(t+1)} \sim \pi_p(\theta_p|\theta_1^{(t+1)}, \ldots, \theta_{p-1}^{(t+1)})$.

La validation de l'algorithme de Gibbs bivarié ci-dessus s'étend à ce cas : la loi jointe π est stationnaire à chaque étape de cet algorithme, car les π_j sont des lois conditionnelles complètes de π. Sous la *contrainte de positivité* que le support de π est le produit cartésien des supports des π_i, la chaîne résultante est ergodique.

Comparé à l'algorithme de Metropolis-Hastings, le nombre de versions de l'échantillonnage de Gibbs est faible et, de plus, les différences entre les propriétés de convergence sont souvent mineures. L'approche de (6.22) (aussi appelée *échantillonnage de substitution* dans Gelfand et Smith, 1990) devrait malgré tout être préférable à une approche direction par direction, car elle respecte la structure hiérarchique initiale du modèle et converge souvent plus rapidement vers la loi stationnaire (voir Liu *et al.*, 1994, 1995, et Roberts et Sahu, 1997). L'échantillonnage de Gibbs bivarié est le seul cas d'échantillonnage de Gibbs produisant une chaîne de Markov pour à la fois $(\theta^{(t)})$ et $(\lambda^{(t)})$; dans tout autre procédé, les sous-chaînes ne sont pas des chaînes de Markov (Exercice 6.33).

Cependant, pour être capable d'utiliser l'échantillonnage de Gibbs bivarié ou même l'échantillonnage de substitution, on a besoin des lois conditionnelles pour tout niveau hiérarchique (comme $\pi(\eta, \mu|y, z)$ dans l'Exemple 6.21) et celles-ci peuvent être plus difficiles à calculer que les lois conditionnelles complètes (voir l'Exercice 6.50). De plus, l'échantillonnage de Gibbs ne requiert pas en réalité que les θ_i soient unidimensionnels et le choix de la décomposition peut alors être entièrement fondé sur des raisons de simulation. Notons aussi que, lorsque des lois conditionnelles, comme $\pi(\theta|x, \lambda_{i_0})$, peuvent être simulées, il est bien entendu préférable d'utiliser ces lois, car elles augmentent la vitesse de convergence en réduisant la dépendance en les autres paramètres. (Cette technique est appelée *regroupement*; voir par exemple Roberts et Sahu, 1997.) Une dernière remarque importante en pratique est que, chaque fois que la simulation d'une loi conditionnelle donnée $\pi_i(\theta_i|\theta_j,\, j \neq i)$ est difficile, cette étape de simulation peut être remplacée par une *seule* étape de Metropolis-Hastings de loi cible $\pi_i(\theta_i|\theta_j,\, j \neq i)$. Ceci peut sembler constituer un mécanisme d'approximation rudimentaire, mais ce n'est pas le cas : le remplacement d'une simulation de $\pi_i(\theta_i|\theta_j,\, j \neq i)$ par une étape de Metropolis-Hastings ne modifie pas la loi stationnaire de la chaîne, et est donc entièrement valable d'un point de vue MCMC.

Exemple 6.23. (Suite de l'Exemple 6.16) Lorsque N, la taille de la population, est le paramètre d'intérêt, l'échantillonnage de Gibbs fournit un échantillon $N_1, \ldots, N_m$, partant de la valeur initiale de $p = (p_1^{(0)}, \ldots, p_n^{(0)})$, en

simulant itérativement

$$N^{(j)} - x_+|x, p^{(j-1)} \sim \mathcal{N}eg(x_+, \varrho^{(j-1)}),$$
$$p_i^{(j)}|x, N^{(j)} \sim \mathcal{B}e(\alpha + x_i, \beta + N^{(j)} - x_i) \qquad (1 \leq i \leq n).$$

(Il s'agit en fait d'un cas d'échantillonnage de Gibbs bivarié.) L'échantillon $N_1, \ldots, N_m$ est alors obtenu en prenant $N_1 = N^{(k_0+T)}$, $N_2 = N^{(k_0+2T)}$, ..., $N_m = N^{(k_0+mT)}$, où k_0 représente le temps de "chauffe", c'est-à-dire le nombre de répétitions pour devenir raisonnablement proche de la stationnarité, et T est la taille du paquet, c'est-à-dire le nombre de répétitions pour accomplir l'indépendance approximative entre les points de l'échantillon. L'échantillonnage de Gibbs fournit simultanément un échantillon $p^1, \ldots, p^m$. L'espérance $\mathbb{E}^\pi[N|x]$ peut alors être approchée par

$$\hat{\delta}^\pi(x) = \frac{1}{m}\sum_{t=1}^{m} \mathbb{E}^\pi[N|x, p^t]$$
$$= \frac{1}{m}\sum_{t=1}^{m} \left(1 - \prod_{i=1}^{n}(1 - p_i^t)\right)^{-1} x_+,$$

selon l'argument de "Rao-Blackwellisation" mentionné ci-dessus. George et Robert (1992) fournissent des extensions hiérarchiques dans ce cadre en considérant différentes familles de lois a priori pour les hyperparamètres (α, β) qui deviennent eux-mêmes aléatoires. ‖

Une comparaison générale entre algorithmes de Metropolis-Hastings et échantillonnage de Gibbs n'a pas de sens : suivant le problème considéré et le choix de lois proposées ou de décompositions hiérarchiques, un algorithme peut converger plus rapidement qu'un autre. Le seul avertissement que nous pouvons fournir ici est que, contrairement à une croyance répandue, l'échantillonnage de Gibbs n'est pas nécessairement une solution optimale. En effet, même si cet algorithme se construit directement à partir de la loi cible π et ne fait donc pas intervenir un apport subjectif de l'expérimentateur, le fait qu'il mette à jour une composante de la chaîne (ou un bloc) à la fois peut affaiblir de beaucoup ses propriétés de convergence si la loi a un support très étroit ou multimodal. Au contraire, un algorithme de Metropolis-Hastings utilisant un mécanisme de proposition à marche aléatoire peut être inefficace si la forme ou l'échelle de la loi proposée ne sont pas ajustées au support de π ; en revanche, cette approche peut aussi permettre de grands sauts pouvant atteindre des modes plus éloignés de π. Nous pourrions qualifier les échantillonneurs de Gibbs d'algorithmes *locaux* et les techniques de Metropolis-Hastings à marche aléatoire d'algorithmes *globaux* au sens où, grossièrement, les premiers fournissent souvent une meilleure image des alentours du point de départ, tandis que les seconds explorent le support de π sur une plus large échelle (voir Besag, 2000, pour une discussion plus détaillée). La meilleure solution à ce dilemme est alors de profiter des caractéristiques positives de ces

différents échantillonneurs en les combinant en un algorithme *hybride* incorporant différentes étapes MCMC, de façon déterministe ou aléatoire.

6.3.6 L'échantillonnage par tranche

L'échantillonnage de Gibbs peut apparaître à ce stade comme une méthode MCMC particulière qui ne peut être utilisée que dans un cadre relativement restrictif : il fait intervenir des structures hiérarchiques, comme dans (6.19) et ne s'applique donc pas à des problèmes unidimensionnels ; il nécessite la connaissance des lois conditionnelles complètes et ne peut donc s'appliquer à des modèles complexes.

Cette perception de l'échantillonnage de Gibbs est erronée : comme nous allons le voir tout de suite, cette méthode s'applique aussi à des problèmes unidimensionnels, elle ne requiert pas une simulation des lois conditionnelles complètes, et elle s'applique aux mêmes modèles que les autres méthodes MCMC. En fait, la décomposition hiérarchique (6.19) n'est pas particulièrement restrictive. En effet, de nombreuses lois (des observations ou des paramètres) peuvent s'écrire comme des *mélanges cachés*, pour un paramètre λ totalement artificiel (voir la Note 6.6.3). Par conséquent, même lorsqu'une structure hiérarchique n'apparaît pas dans le problème original, elle peut souvent être réintroduite pour améliorer le calcul des estimateurs de Bayes ou même le choix de la loi a priori.

La généralité de l'échantillonnage de Gibbs est mise en évidence dans la version particulière dite de *l'échantillonnage par tranche* (Wakefield *et al.*, 1991, Besag et Green, 1993, et Damien *et al.*, 1999). Considérons une loi $\pi(\theta)$ sur un ensemble général Θ, uni- ou multidimensionnel, et réécrivons π comme le produit

$$\pi(\theta) = \prod_{i=1}^{k} \varpi_i(\theta), \tag{6.24}$$

où les ϖ_i sont des fonctions positives, mais non nécessairement des densités. Alors $\pi(\theta)$ peut s'écrire comme la loi marginale

$$\pi(\theta) = \int \prod_{i=1}^{k} \mathbb{I}_{0\le\omega_i\le\varpi_i(\theta)}\, d\,\omega_1 \cdots d\,\omega_k\,.$$

L'*échantillonnage par tranche* correspondant s'obtient directement :

ALGORITHME 6.4. –**Échantillonnage par tranche**–
À l'itération t, simuler

1. $\omega_1^{(t+1)} \sim \mathcal{U}_{[0,\varpi_1(\theta^{(t)})]}$

⋮

k. $\omega_k^{(t+1)} \sim \mathscr{U}_{[0,\varpi_k(\theta^{(t)})]}$

k+1. $\theta^{(t+1)} \sim \mathscr{U}_{A^{(t+1)}}$, avec

$$A^{(t+1)} = \{\xi;\ \varpi_i(\xi) \geq \omega_i^{(t+1)},\ i = 1, \ldots, k\}.$$

Les ω_j sont un type particulier de *variables auxiliaires*, sans signification pour le problème statistique considéré. Notons qu'il existe de nombreuses représentations possibles (6.24) pour la même loi π, notamment le cas simple

$$\pi(\theta) = \int_0^1 \mathbb{I}_{0\leq\omega\leq\pi(\theta)} d\omega\,,$$

et que le choix d'une représentation est purement dicté par son caractère pratique. En fait, la dernière étape (k+1) dans l'algorithme ci-dessus peut être délicate à mettre en œuvre, puisque l'ensemble $A^{(t)}$ est souvent difficile à construire, mais cette décomposition montre que l'échantillonnage de Gibbs peut fournir, au moins formellement, une représentation de toutes les lois (voir Roberts et Rosenthal, 1998, Tierney et Mira, 1998 et Mira *et al.*, 2001, pour des propriétés théoriques de l'échantillonnage par tranche.)

Exemple 6.24. (Suite de l'Exemple 6.13) La loi jointe de (α, η) étant

$$\pi(\alpha, \eta | x_1, \ldots, x_n) \propto \alpha^n \eta^{n+\beta-1} \left(\prod_{i=1}^n x_i\right)^\alpha \exp\left\{-\eta \sum_{i=1}^n x_i^\alpha - \alpha - \xi\eta\right\}\,,$$

la loi conditionnelle $\pi_1(\eta|\alpha, x_1, \ldots, x_n)$ est tout simplement la loi

$$\mathscr{G}(\beta + n, \xi + \sum_i x_i^\alpha)$$

qui est facile à simuler. La loi conditionnelle $\pi_2(\alpha|\eta, x_1, \ldots, x_n)$ est beaucoup plus complexe à cause de la partie exponentielle faisant intervenir les x_i^α. Si nous écrivons cette loi comme $\alpha^n \chi^\alpha \exp(-\eta \sum_{i=1}^n x_i^\alpha)$, nous pouvons l'exprimer comme la loi marginale (en α) de

$$\alpha^n \mathbb{I}_{0\leq\omega_0\leq\chi^\alpha} \prod_{i=1}^n \mathbb{I}_{0\leq\omega_i\leq\exp(-\eta x_i^\alpha)}\,.$$

La loi conditionnelle de α sachant η et les ω_i est alors proportionnelle à

$$\alpha^n \mathbb{I}_{\alpha\log(\chi)\leq\log(\omega_0)} \prod_{i=1}^n \mathbb{I}_{\alpha\log(x_i)\leq\log\{-\log(\omega_i)/\eta\}}\,,$$

c'est-à-dire une simple loi puissance α^n sur un intervalle $(\underline{\alpha}, \overline{\alpha})$. L'échantillonnage de Gibbs de la loi a posteriori de Weibull s'obtient alors par simulation itérative des η, des ω_i et des α. ‖

Exemple 6.25. (Suite de l'Exemple 6.5) Puisque la loi a posteriori de $\theta = (\mu_1, \sigma_1^2, p, \mu_2, \sigma_2^2)$ admet une expression analytique,

$$\pi(\theta|x) \propto \tilde{\pi}(\theta|x) = \pi(\theta) \prod_{i=1}^{n} \{p\varphi(x_i; \mu_1, \sigma_1) + (1-p)\varphi(x_i; \mu_2, \sigma_2)\} \ ,$$

un échantillonneur par tranche formel admettant une variable auxiliaire unique ω peut être proposé, avec $\theta \sim \mathscr{U}_{\tilde{\pi}(\theta|x) \geq \omega}$. Mais il est impossible de simuler cette loi uniforme, puisque la contrainte $\tilde{\pi}(\theta|x) \geq \omega$ ne peut pas être transformée en une contrainte sur θ. Une version utilisable de l'échantillonnage par tranche dans ce cadre peut se construire en introduisant plutôt n variables auxiliaires ω_i de telle manière que $\tilde{\pi}(\theta|x)$ s'écrive comme la loi marginale de

$$\pi(\theta) \prod_{i=1}^{n} \mathbb{I}_{p\varphi(x_i;\mu_1,\sigma_1)+(1-p)\varphi(x_i;\mu_2,\sigma_2) \geq \omega_i \geq 0} \cdot$$

Bien que la loi jointe de θ conditionnelle aux ω_i ne soit pas toujours disponible, les lois conditionnelles complètes des paramètres μ_1, σ_1^2, p, μ_2 et σ_2^2 sont simples à simuler. (Comme nous le verrons dans la Section 6.4, qui traite des mélanges, l'échantillonneur de Gibbs initialement proposé pour ce modèle repose aussi sur la simulation de n variables auxiliaires.) ‖

6.3.7 L'impact des méthodes MCMC sur la statistique bayésienne

Cette section a présenté très brièvement les bases des méthodes MCMC, et donné quelques illustrations tirées des problèmes de calcul bayésien. Il est important de souligner à ce stade que l'apparition de ces outils MCMC en statistique bayésienne a eu un effet "dévastateur" ! En effet, elle a radicalement modifié la façon dont les gens travaillent avec des modèles et des hypothèses a priori, permettant de prendre en compte des structures beaucoup plus complexes, comme par exemple dans le cas des *modèles graphiques* où les relations entre variables ne sont définies qu'à un niveau local, la loi jointe étant impossible à concevoir (voir Cowell *et al.*, 1999, et Note 10.7.1). De même, *les modèles à variables latentes* comme les modèles de chaînes de Markov cachées ou à volatilité stochastique, peuvent désormais être correctement analysés (voir la Note 6.6.5 et Robert et Casella, 1999, Chapitre 9) alors que seules des approximations grossières étaient disponibles par le passé, un changement qui a eu un impact immense en traitement du signal bayésien, en économétrie et en finance mathématique.

La "dévastation" mentionnée ci-dessus concerne aussi les structures rigides autrefois imposées par la contrainte d'un traitement analytique ; par exemple, le recours à des lois conjuguées n'est plus indispensable, même si celles-ci restent très utiles comme lois a priori de base pour les différents niveaux d'une

modélisation hiérarchique (voir le Chapitre 10). De même, des représentations beaucoup plus flexibles peuvent être proposées dans le domaine du choix de modèle, comme nous le verrons au Chapitre 7, où la possibilité de prendre en compte de nombreux modèles simultanément incite le statisticien à passer des tests au sens strict au *moyennage de modèles*, les modèles les plus probables obtenant les poids les plus élevés mais sans écarter aucun modèle a priori ; voir aussi Berger (2000), Cappé et Robert (2000) et Gelfand (2000) pour des revues sur l'impact des méthodes MCMC.

Comme toujours, un accroissement significatif de la facilité à utiliser une technique donnée s'accompagne d'un accroissement proportionnel des possibilités de détournements de cette technique. Dans le cas de l'analyse bayésienne, cela signifie que l'impact d'une modélisation a priori est plus difficile à évaluer à partir des lois conditionnelles utilisées en échantillonnage de Gibbs. Pis, la loi a posteriori peut être impropre (Section 1.5) sans que son utilisateur en soit conscient (voir la Note 6.6.4). Mais ces défauts ne peuvent pas se comparer avec les conséquences sur la portée et le nombre d'applications bayésiennes rencontrées depuis dans la littérature, incluant la résolution de problèmes inférentiels jamais considérés auparavant.

6.4 Estimation bayésienne de mélanges

Nous concluons ce chapitre en montrant comment les méthodes MCMC permettent le calcul d'estimateurs de Bayes des paramètres d'un mélange de lois normales considéré dans l'Exemple 6.5. L'extension à d'autres mélanges de lois appartenant à une famille exponentielle ou à des modèles à chaînes de Markov cachées est triviale (voir Gruet *et al.*, 1999, et Robert et Casella, 2004, Notes 9.7.1 et 14.6.3). Comme nous l'avons détaillé dans la Section 6.1, une analyse bayésienne d'un modèle de mélange mène au paradoxe de l'information suivant : un estimateur explicite est disponible et est justifiable intuitivement, mais il ne peut pas être calculé lorsque le nombre d'observations devient trop grand. De plus, les estimateurs du maximum de vraisemblance des paramètres de (6.4) ne sont pas clairement définis, la résolution des équations de vraisemblance est difficile et les approximations analytiques des estimateurs de Bayes posent problème (voir Crawford *et al.*, 1992 , pour une approche reposant sur l'approximation de Laplace). De même, un traitement Monte Carlo standard des modèles de mélanges est ardu même si Casella *et al.* (2000) ont proposé une méthode fondée sur l'échantillonnage d'importance dans un cadre conjugué (Exercice 6.42) ; voir la Note 6.6.6 pour de plus amples références et des détails sur les débuts de l'estimation de mélanges.

L'échantillonnage de Gibbs pour les mélanges repose sur une représentation *par données manquantes*, comme dans Dempster *et al.* (1977), afin de construire une structure hiérarchique similaire à (6.19). Soit

$$x \sim f(x|\theta) = \sum_{i=1}^{k} p_i \varphi(x; \mu_i, \sigma_i), \tag{6.25}$$

un mélange de k lois normales de moyennes μ_i et variances σ_i^2 $(1 \leq i \leq k)$, avec $\sum_i p_i = 1$ $(p_i > 0)$. Pour un échantillon $x_1, \ldots, x_n$ donné de (6.25), on définit les valeurs manquantes z_j $(1 \leq j \leq n)$ comme les *vecteurs d'indicatrices de composantes* des x_j, c'est-à-dire

$$z_{ij} = \begin{cases} 1 & \text{si } x_j \sim \varphi(x; \mu_i, \sigma_i), \\ 0 & \text{sinon,} \end{cases}$$

et $\sum_i z_{ij} = 1$. Ce vecteur peut aussi être considéré comme un paramètre supplémentaire ; il correspond à la loi jointe suivante $(1 \leq j \leq n)$:

$$\begin{aligned} z_j|\theta &\sim \mathscr{M}_p(1; p_1, \ldots, p_k)\,, \\ x_j|z_j, \theta &\sim \mathscr{N}\left(\prod_{i=1}^{k} \mu_i^{z_{ij}}, \prod_{i=1}^{k} \sigma_i^{2z_{ij}}\right). \end{aligned}$$

Une loi a priori commode pour $\theta = (\mu_1, \sigma_1, p_1, \ldots, \mu_k, \sigma_k, p_k)$ est le produit des lois conjuguées $\pi_i(\mu_i, \sigma_i)$, où $\pi_i(\mu_i|\sigma_i)$ est une loi normale $\mathscr{N}(\xi_i, \sigma_i^2/n_i)$, $\pi_i(\sigma_i^2)$ une loi gamma inverse $\mathscr{IG}(\nu_i/2, s_i^2/2)$, et $\pi(p)$ une loi de Dirichlet, $\mathscr{D}(\alpha_1, \ldots, \alpha_k)$, comme dans l'Exemple 6.5.

Notons que, une fois connus les vecteurs d'allocation z_j $(1 \leq j \leq n)$, la structure de mélange disparaît, puisque cette information supplémentaire décompose l'échantillon en sous-échantillons selon les valeurs de z_{ij}. Bien que la loi a posteriori de θ ne puisse pas être utilisée directement, comme le montre l'Exemple 6.5, le conditionnement en $\mathbf{z} = (z_1, \ldots, z_n)$ supprime cette difficulté. En effet, on obtient les lois a posteriori suivantes $(1 \leq j \leq n)$:

$$z_j|x_j, \theta \sim \mathscr{M}_k(1; p_1(x_j, \theta), \ldots, p_k(x_j, \theta)), \tag{6.26}$$

avec $(1 \leq i \leq k)$

$$p_i(x_j, \theta) = \frac{p_i \varphi(x_j; \mu_i, \sigma_i)}{\sum_{t=1}^{k} p_t \varphi(x_j; \mu_t, \sigma_t)},$$

et

$$\begin{aligned} \mu_i|\mathbf{x}, \mathbf{z}, \sigma_i &\sim \mathscr{N}(\xi_i(\mathbf{x}, \mathbf{z}), \sigma_i^2/(n + \sigma_i^2)), \\ \sigma_i^2|\mathbf{x}, \mathbf{z} &\sim \mathscr{IG}\left(\frac{\nu_i + n_i}{2}, \frac{1}{2}\left[s_i^2 + \hat{s}_i^2(\mathbf{x}, \mathbf{z}) + \frac{n_i m_i(\mathbf{z})}{n_i + m_i(\mathbf{z})}(\bar{x}_i(\mathbf{z}) - \xi_i)^2\right]\right), \\ p|\mathbf{x}, \mathbf{z} &\sim \mathscr{D}_k(\alpha_1 + m_1(\mathbf{z}), \ldots, \alpha_k + m_k(\mathbf{z})), \end{aligned} \tag{6.27}$$

où

$$m_i(\mathbf{z}) = \sum_{j=1}^{n} z_{ij}, \qquad \bar{x}_i(j) = \frac{1}{m_i(\mathbf{z})} \sum_{j=1}^{n} z_{ij} x_j,$$

et

$$\xi_i(\mathbf{x},\mathbf{z}) = \frac{n_i\xi_i + m_i(\mathbf{z})\bar{x}_i(\mathbf{z})}{n_i + m_i(\mathbf{z})}, \qquad \hat{s}_i^2(\mathbf{x},\mathbf{z}) = \sum_{j=1}^{n} z_{ij}(x_j - \bar{x}_i(\mathbf{z}))^2.$$

Conditionnellement à $\mathbf{z}$, les lois a posteriori ne prennent en compte que les sous-échantillons correspondant à chaque composante, à l'instar de la décomposition (6.6) de la vraie loi a posteriori. De plus, simuler selon (6.26) et (6.27) est particulièrement simple. Il est donc beaucoup plus facile de produire un échantillon $\theta_1, \ldots, \theta_m$ de $\pi(\theta|\mathbf{x})$ par échantillonnage de Gibbs que d'utiliser la vraie loi a posteriori directement.

La remarque qui suit le Lemme 6.17 implique que l'échantillonnage de Gibbs entraîne une convergence géométrique uniforme de la chaîne $(\theta^{(m)})$, puisque $\mathbf{z}$ a un support fini.

Comme dernière remarque, nous soulignons que l'échantillonnage de Gibbs n'est pas la seule solution pour la simulation de la loi a posteriori $\pi(\theta|\mathbf{x})$. En effet, comme le montre l'Exemple 6.25, une expression analytique de cette loi est disponible : elle peut donc être utilisée dans un algorithme de Metropolis-Hastings (en plus de l'échantillonnage par tranche produit dans l'Exemple 6.25). Par exemple, Celeux *et al.* (2000) démontrent que la stratégie de Metropolis-Hastings par marche aléatoire peut être utilisée de façon efficace dans ce cadre et admet de meilleures propriétés de mélangeance que l'échantillonnage de Gibbs. Dans le cas des modèles *de chaînes de Markov cachées*, qui généralisent les modèles de mélange comme (6.25) en introduisant une dépendance markovienne entre les z_j, il existe aussi dans certains cas des représentations analytiques de la vraisemblance par intégration sur les variables latentes ; voir les Exercices 6.50 et 6.51, et Robert *et al.* (1999a) .

6.5 Exercices

Section 6.1

6.1 Pour un mélange de deux lois normales, comme celui de l'Exemple 6.5 et les données de la Table 6.1, identifier les hyperparamètres des lois conditionnelles par la méthode des moments.

6.2 Dans le cadre de l'Exemple 6.5, montrer que la loi a posteriori peut en fait s'écrire sous la forme (6.6), et développer $\omega(k_t)$ et $\pi(\theta|(k_t))$. Donner les expressions des estimateurs de Bayes de μ_1, σ_1 et p pour les hyperparamètres obtenus dans l'Exemple 6.21.

6.3 Mêmes questions que l'Exercice 6.2 pour

(i) un mélange de deux lois exponentielles ; et

(ii) un mélange de trois lois uniformes.

6.4 Dans l'Exercice 6.2, comment évolue le temps de calcul en fonction de la taille d'échantillon lorsque

(i) seul le poids p est inconnu ? et

(ii) tous les paramètres sont inconnus ?

6.5 *(Smith et Makov, 1978) Soit

$$x \sim f(x|p) = \sum_{i=1}^{k} p_i f_i(x),$$

avec $p_i > 0$, $\sum_i p_i = 1$, les densités f_i étant connues. L'a priori $\pi(p)$ est une loi de Dirichlet $\mathscr{D}(\alpha_1, \ldots, \alpha_k)$.

a. Montrer que le temps de calcul reste prohibitif malgré la simplicité du modèle lorsque la taille d'échantillon augmente.

Une solution alternative séquentielle, permettant une approximation de l'estimateur de Bayes, est de remplacer $\pi(p|x_1, \ldots, x_n)$ par $\mathscr{D}(\alpha_1^{(n)}, \ldots, \alpha_k^{(n)})$, où

$$\alpha_1^{(n)} = \alpha_1^{(n-1)} + P(z_{n1} = 1|x_n), \ldots, \alpha_k^{(n)} = \alpha_k^{(n-1)} + P(z_{nk} = 1|x_n),$$

et z_{ni} $(1 \leq i \leq k)$ est le vecteur d'indicatrices des composantes de x_n, défini en Section 6.4.

b. Justifier cette approximation et la comparer avec la mise à jour $\pi(p|x_1, \ldots, x_{n-1})$ pour x_n observé.

c. Étudier les performances de cette approximation pour un mélange de deux lois normales $\mathscr{N}(0,1)$ et $\mathscr{N}(2,1)$ pour $p = 0.1, 0.25, 0.5$.

d. Si $\pi_i^n = P(z_{ni} = 1|x_n)$, montrer que

$$\hat{p}_i^{(n)}(x_n) = \hat{p}_i^{(n-1)}(x_{n-1}) - a_{n-1}\{\hat{p}_i^{(n-1)} - \pi_i^n\},$$

où $\hat{p}_i^{(n)}$ est l'approximation quasi bayésienne de $\mathbb{E}^\pi(p_i|x_1, \ldots, x_n)$.

6.6 Dans le cadre de l'Exemple 6.4, déterminer la loi a posteriori de $\pi(N|x)$: (a) pour $n = 10$ et des x_i prenant des valeurs similaires ; et (b) pour $n = 30$ et des x_i prenant des valeurs très différentes. Traiter le même problème lorsque $\pi(N)$ est une loi de Poisson $\mathscr{P}(\lambda)$ et λ varie. Faire particulièrement attention aux problèmes potentiels liés à une évaluation directe.

Section 6.2.1

6.7 *(Morris, 1982) Pour les familles exponentielles naturelles à variance quadratique étudiées dans les Exercices 3.24 et 10.33, on pose

$$P_m(x, \mu) = V^m(\mu) \left\{ \frac{d^m}{d\mu^m} f(x|\mu) \right\} \Big/ f(x|\mu).$$

a. Montrer que P_m est un polynôme de degré m en x et μ.

b. Montrer que $(m > 1)$

$$P_{m+1}(x, \mu) = [P_1(x, \mu) - mV'(\mu)] \quad P_m(x, \mu) - m[1 + (m-1)v_2]V(\mu)P_{m-1}(x, \mu),$$

où $V(\mu) = v_0 + v_1\mu + v_2\mu^2$.

c. Montrer que les polynômes P_m sont orthogonaux et que $\mathbb{E}_\mu[P_m^2(x,\mu)] = a_m V^m(\mu)$.

d. Donner les polynômes associés aux lois normale, de Poisson, gamma, binomiale et binomiale négative. [*Note :* Il s'agit respectivement des polynômes d'*Hermite*, de *Poisson-Charlier*, généralisés de *Laguerre*, *de Krawtchouk* et *de Meixner*.]

Section 6.2.2

6.8 Montrer que, si le support de h ne contient pas celui de $f(x|\theta)\pi(\theta)$, l'approximation par échantillonnage d'importance (6.11) ne converge pas.

6.9 La méthode standard de simulation d'*acceptation-rejet* est définie à partir de densités f et g telles que $f(x) \leq Mg(x)$ pour un certain M par l'algorithme :

ALGORITHME 6.5. –**Acceptation-Rejet**–

1. Tirer $y \sim g(y)$ et $u \sim \mathscr{U}_{[0,1]}$;
2. Si $u > f(y)/Mg(y)$, revenir en 1.
3. Prendre $x = y$.

Montrer que cet algorithme fournit bien une observation x de loi $f(x)$.

6.10 Montrer que, si U_1, U_2 sont iid $\mathscr{U}_{[0,1]}$,

1. Les transformations

$$X_1 = \sqrt{-2\log(U_1)}\,\cos(2\pi U_2)\,, \qquad X_2 = \sqrt{-2\log(U_1)}\,\sin(2\pi U_2)\,,$$

sont iid $\mathscr{N}(0,1)$.

2. Les coordonnées polaires sont de lois

$$r^2 = X_1^2 + X_2^2 \sim \chi_2^2\,,$$
$$\theta = \arctan\frac{X_1}{X_2} \sim \mathscr{U}[0,2\pi].$$

3. En déduire l'algorithme de Box-Muller (Box et Muller, 1958) de génération de lois normales :

ALGORITHME 6.6. –**Box-Muller (1)**–

1. Générer
$$U_1, U_2 \sim \mathscr{U}([0,1])$$
2. Prendre
$$X_1 = \sqrt{-2\log(U_1)}\,\cos(2\pi U_2)\,, \qquad X_2 = \sqrt{-2\log(U_1)}\,\sin(2\pi U_2)\,,$$

6.11 (Suite de l'Exercice 6.9)

1. Montrer qu'une version plus rapide de l'Algorithme 6.6 de Box–Muller est :

ALGORITHME 6.7. –**Box-Muller (2)**–

1. Générer
$$U_1, U_2 \sim \mathscr{U}([-1,1])$$
jusqu'à ce que $S = U_1^2 + U_2^2 \leq 1$.
2. Poser $Z = \sqrt{-2\log(S)/S}$ et déduire
$$X_1 = Z\,U_1, \qquad X_2 = Z\,U_2.$$

en montrant que (U_1, U_2) est uniforme sur la boule unité et que X_1 et X_2 sont indépendants.

2. Donner le nombre moyen de générations dans l'étape 1. et comparer avec l'Algorithme 6.6 via une expérience informatique.
3. Que se passe-t-il si l'on ne restreint pas (U_1, U_2) à la boule unité ?

6.12 *(Gilks et Wild, 1992) On considère une méthode générale d'*acceptation-rejet* pour des densités *log-concaves* sur $\mathbb{R}$. Cette méthode est fondée sur des bornes supérieures et inférieures adaptatives de la densité, qui sont mises à jour après chaque simulation.

a. Pour $f(x)$ donné, proportionnel à la densité à simuler, on suppose qu'il existe $u(x)$ et $\ell(x)$, bornes supérieure et inférieure de $f(x)$ telles que u soit une densité. L'*algorithme d'acceptationrejet avec enveloppe* s'écrit :

ALGORITHME 6.8. –**Simulation par enveloppe**–
Répéter
a) Générer $x \sim u(x)$ et $U \sim \mathscr{U}_{[0,1]}$
b) Accepter x si $U \leq \ell(x)/u(x)$
c) Sinon, accepter x si $U \leq f(x)/u(x)$
jusqu'à ce que x soit accepté.

Montrer que cette méthode produit bien une variable aléatoire de loi f.

b. Les deux fonctions encadrantes peuvent être construites automatiquement comme suit, pour f log-concave. Pour la première simulation, prendre trois valeurs arbitraires x_1, $x_2 > x_1$ et $x_3 > x_2$ telles qu'au moins une d'entre elles soit de chaque côté du mode de f. (Expliquer comment cela peut être fait sans calcul explicite du mode.) Montrer que la borne inférieure $\log \ell(x)$ de $\log f(x)$ peut être obtenue en joignant les trois points $(x_i, \log f(x_i))$ et en posant $\ell(x) = 0$ en dehors de l'intervalle $[x_1, x_3]$. La borne supérieure $\log u(x)$ est obtenue en prenant les compléments des segments utilisés pour $\log \ell(x)$ jusqu'à ce qu'ils se croisent : les queues consistent alors en des extensions des arcs (x_1, x_2) et (x_2, x_3) ; $\log u(x)$ est complété par l'ajout de segments verticaux passant par x_1 et x_3 et continuant jusqu'à ce qu'ils rencontrent les deux arcs.

c. Proposer une méthode de mise à jour des bornes supérieure et inférieure après chaque simulation nécessitant le calcul de $f(x)$.

d. Montrer que les deux fonctions $u(x)$ et $\ell(x)$ sont exponentielles par morceaux et indiquer comment simuler des lois de densité proportionnelle à ces fonctions.

e. Illustrer l'algorithme ci-dessus pour la simulation de la loi $\mathcal{N}(0,1)$. À partir de quand devient-il plus coûteux d'évaluer et de simuler une borne supérieure améliorée, plutôt que de conserver la borne courante ?

6.13 *(Rubinstein, 1981) On considère l'intégrale

$$I = \int_a^b f(x)\,dx,$$

approchée par une méthode de Monte Carlo avec fonction d'importance h :

$$\hat{I} = \frac{1}{m}\sum_{i=1}^{m} f(x_i)/h(x_i).$$

a. Montrer que la variance de $\hat{I}$ est

$$\mathrm{var}(\hat{I}) = \frac{1}{n}\int_a^b \left(\frac{f(x)}{h(x)} - I\right)^2 h(x)\,dx$$

et en déduire qu'elle est minimisée par $h \propto |f|$.

b. En décomposant h en $h^+ - h^-$, déduire qu'une variance nulle est toujours atteignable formellement.

c. Soient $0 \le f(x) \le c$, $v_1, \ldots, v_m \sim \mathcal{U}_{[0,c]}$ et $u_1, \ldots, u_m \sim \mathcal{U}_{[a,b]}$. On définit

$$\hat{I} = (b-a)\frac{1}{m}\sum_{i=1}^{m} f(u_i) \quad \text{et} \quad \tilde{I} = c(b-a)\frac{1}{m}\sum_{i=1}^{m} \mathbb{I}_{v_i \le f(u_i)}.$$

Montrer que

$$I = c(b-a)P(V \le f(U))$$

pour $U \sim \mathcal{U}_{[a,b]}$ et $V \sim \mathcal{U}_{[0,c]}$.

d. En déduire que $\mathbb{E}[\tilde{I}] = I$ et $\mathrm{var}(\tilde{I}) \le \mathrm{var}(\hat{I})$.

e. Discuter la pertinence de la notion d'une fonction d'importance "optimale". (*Indication :* Considérer une suite de lois normales centrées en la valeur d'intérêt, c'est-à-dire en x^* tel que $f(x^*) = I$, et de variances décroissant vers 0.)

6.14 Montrer que, pour une fonction $g(\theta)$ donnée et une distribution d'intérêt $\pi(\theta)$, le choix optimal de la densité d'importance h, en termes de variance de l'estimateur

$$\sum_{i=1}^{n} g(\theta_i)\omega_i\,,$$

est

$$h(\theta) \propto |g(\theta)|\pi(\theta)\,.$$

Donner l'expression de l'estimateur correspondant et en déduire que, si g est de signe constant, la variance résultante est 0. (*Indication :* voir Robert et Casella, 2004, Théorème 3.12, pour une démonstration.)

6.15 (Suite de l'Exercice 6.14) Dans le cas de constantes inconnues, c'est-à-dire quand l'estimateur (6.11) est utilisé, montrer que la solution optimale au sens de la variance est telle que

$$h(\theta) \propto |g(\theta) - \mathbb{E}[g]|\pi(\theta)\,.$$

Section 6.2.3

6.16 Justifier l'approximation de Laplace pour $h(\theta) = (\theta - \mu)^2$ et $b(\theta)$ polynôme de degré 2. Que se passe-t-il si le degré de b est plus grand ? Obtenir le développement général de Laplace à partir de développements de Taylor de b et h.

6.17 *(Tierney *et al.*, 1989) Déduire de l'approximation de Laplace que

$$\frac{\int b_N(\theta)e^{-nh_N(\theta)}d\theta}{\int b_D(\theta)e^{-nh_D(\theta)}d\theta} = \frac{A(N)}{A(D)} + \mathrm{O}(\sigma^{-2}),$$

où

$$A(K) = \sigma_K \exp\{-n\hat{h}_k\} \left[\hat{b}_K + \frac{1}{2n}\left\{\sigma_K^2\hat{b}_K'' - \hat{h}_K'''\hat{b}_K'\sigma_K^2 + \frac{5}{12}\hat{b}_K(\hat{h}_K''')^2\sigma_K^6 - \frac{1}{4}\hat{b}_K\sigma_K^2 h_K^{(4)}\right\}\right]$$

et $K = N, D$, si $\hat{h}_K = h(\hat{\theta}_K)$, etc., et $\hat{\theta}_K$ minimise h_K. En déduire le Lemme 6.9 sous l'hypothèse que $\hat{h}_N^{(i)} - \hat{h}_D^{(i)} = \mathrm{O}(n^{-1})$ pour $i = 0, \ldots, 4$ et $\hat{b}_D \neq 0$. Que se passe-t-il si $\hat{b}_D = 0$?

6.18 *(Tierney *et al.*, 1989) Pour $M(s)$ fonction génératrice des moments de $g(\theta)$ et $\hat{M}$ l'approximation de Laplace de M pour (6.16), avec $b_N = b_D = b > 0$ et

$$h_D(\theta) = \{\log[fx|\theta)] + \log[\pi(\theta)] - \log[b(\theta)]\}/n,$$

$h_N(\theta) = h_D(\theta) - sg(\theta)/n$, on définit

$$\hat{\mathbb{E}}(g) = \hat{M}'(0).$$

a. Montrer que $\mathbb{E}^\pi[g(\theta)|x] = \hat{\mathbb{E}}(g) + \mathrm{O}(n^{-2})$.

b. Soit $\hat{\theta}$ le minimum de h_D, $\hat{\theta}_s$ celui de h_N et $\sigma_s^2 = h_N^{(2)}(\theta_s)$. Montrer que

$$\hat{\mathbb{E}}(g) = g(\hat{\theta}) + \frac{d}{ds}\log\sigma_s\Big|_{s=0} + \frac{d}{ds}\log b(\hat{\theta}_s)\Big|_{s=0}\,.$$

c. En déduire que

$$\hat{\mathbb{E}}(g) = \hat{g} + \frac{\sigma_D^2\hat{g}''}{2n} - \frac{\sigma_D^4\hat{h}_D'''\hat{g}'}{2n} + \frac{\sigma_D^2\hat{b}_D'\hat{g}'}{n\hat{b}_D}\,,$$

et donc que cette méthode donne bien l'approximation (6.15) pour la forme standard.

6.19 Dans le cadre de l'Exemple 6.11, choisir les représentations standard et exponentielle complète menant aux approximations proposées.

Section 6.3.2

6.20 *Considérons l'algorithme de *Metropolis-Hastings* de la Section 6.3.2, qui simule une densité $\pi(\theta)$ à partir d'une densité proposée $q(\theta'|\theta)$.

a. Montrer que cet algorithme se simplifie en une simulation standard de π lorsque $q(\theta'|\theta) = \pi(\theta')$.

b. Donner la forme simplifiée de l'algorithme de Metropolis-Hastings lorsque $q(\theta|\theta')$ est symétrique en ses arguments, c'est-à-dire lorsque $q(\theta|\theta') = q(\theta'|\theta)$.

c. Montrer directement, c'est-à-dire sans utiliser la condition d'équilibre (6.17), que $\pi(\theta)$ est une loi stationnaire pour cet algorithme lorsque le support de q contient celui de π. (*Indication :* Calculer la fonction de densité de $\theta^{(m+1)}$ lorsque $\theta^{(m)} \sim \pi(\theta)$ en décomposant l'intégrale en quatre parties et en échangeant les variables muettes θ et ξ dans deux des quatre intégrales.)

d. Dans le cas particulier où π est la loi $\mathcal{N}(0,1)$ et $q(\theta|\theta')$ est $\mathcal{N}(\theta', \sigma^2)$, étudier la probabilité d'acceptation de ξ dans le m-ième pas de simulation, en fonction de σ. Quelle est la loi exacte de $\theta^{(m)}$? En déduire la valeur optimale de σ.

6.21 Prouver la condition d'équilibre ponctuel (6.17) pour l'algorithme de Metropolis-Hastings.

6.22 Déterminer si l'algorithme de Metropolis-Hastings produit une chaîne de Markov réversible, c'est-à-dire telle que la loi de $(x^{(t)}, x^{(t+1)})$ soit la même que celle de $(x^{(t+1)}, x^{(t)})$ en situation de stationnarité.

6.23 (Robert, 1993b) Soient n observations $y_1, \ldots, y_n$ issues d'un modèle *de régression logistique*, où

$$P(y_i = 1) = 1 - P(y_i = 0) = \frac{\exp(\theta^t x_i)}{1 + \exp(\theta^t x_i)},$$

et $x_i, \theta \in \mathbb{R}^p$.

a. Montrer que, conditionnellement aux x_i, cette loi appartient à une famille exponentielle et que $\sum_i y_i x_i$ est une statistique exhaustive.

b. Donner la forme générale de la loi conjuguée pour ce modèle et montrer que le facteur de normalisation ne peut pas être calculé explicitement. Donner une interprétation des hyperparamètres (ξ, λ) de la loi conjuguée en termes d'observations précédentes.

c. Montrer que l'estimateur du maximum de vraisemblance de θ, $\hat{\theta}$, ne peut pas être calculé explicitement, et qu'il satisfait les équations implicites suivantes $(j = 1, \ldots, p)$:

$$\sum_{i=1}^{n} \frac{\exp(\hat{\theta}^t x_i)}{1 + \exp(\hat{\theta}^t x_i)} x_{ij} = \sum_{i=1}^{n} y_i x_{ij}. \tag{6.28}$$

d. Approcher une loi conjuguée par l'algorithme de Metropolis-Hastings. [*Note :* Si une loi conditionnelle gaussienne est utilisée, faire attention au facteur de variance.]

e. Expliquer pourquoi (6.28) peut être utilisé pour contrôler la convergence de l'algorithme pour certaines valeurs particulières du vecteur d'hyperparamètres, (ξ, λ), celles pour lesquelles

$$\mathbb{E}^{\pi}_{\xi,\lambda}\left[\sum_{i=1}^{n}\frac{\exp(\theta^t x_i)}{1+\exp(\theta^t x_i)}x_i\right] = \mathbb{E}^{\pi}_{\xi,\lambda}\left[\sum_{i=1}^{n}\frac{\exp(\theta^t x_i)}{1+\exp(\theta^t x_i)}x_i \,\middle|\, y_1,\ldots,y_n\right] = \sum_{i=1}^{n} y_i x_i.$$

6.24 *Pour une densité d'intérêt donnée, π, et une densité connue f telle que $\pi/f \le M$, des tirages de π peuvent être produits par acceptation-rejet (Exercice 6.9), $\theta_1^{(1)},\ldots,\theta_p^{(1)}$, ou par Metropolis-Hastings, avec f pour densité proposée, $\theta_1^{(2)},\ldots,\theta_n^{(2)}$; alternativement, un échantillon d'importance, $\theta_1^{(3)},\ldots,\theta_n^{(3)}$, peut être généré selon f. Comparer par simulation les variances de

$$\frac{1}{p}\sum_{i=1}^{p}\theta_i^{(1)}, \qquad \frac{1}{n}\sum_{i=1}^{n}\theta_i^{(2)}, \qquad \frac{1}{n}\sum_{i=1}^{n}\frac{\pi(\theta_i^{(3)})}{f(\theta_i^{(3)})}\theta_i^{(3)}.$$

[*Note :* p est le nombre aléatoire d'observations produites après n valeurs proposées dans l'algorithme d'acceptation-rejet.]

6.25 Soient une loi de probabilité P et une fonction ϱ telle que $0 \le \varrho(x) \le 1$ et $\mathbb{E}^P[1/\varrho(x)] < \infty$. Une chaîne de Markov $(x^{(n)})$ est construite de la façon suivante : $x^{(n)}$ est remplacé par $x^{(n+1)}$ en générant $y \sim P$ et en prenant

$$x^{(n+1)} = \begin{cases} y & \text{avec probabilité } \varrho(x^{(n)}), \\ x^{(n)} & \text{avec probabilité } 1-\varrho(x^{(n)}). \end{cases}$$

a. Montrer que cette variation de l'algorithme de Metropolis-Hastings converge vers la loi stationnaire de densité

$$\varrho(x)^{-1}/\mathbb{E}^P[\varrho(x)^{-1}]$$

par rapport à P.

b. Appliquer au cas où P est la loi $\mathscr{B}e(\alpha+1,1)$ et $\varrho(x)=x$.

c. Étudier les performances de cette méthode lorsque $\alpha = 0.2$. [*Note :* Voir Robert et Casella, 1999, Exemple 8.2.8, pour une illustration des mauvaises performances de ce générateur.]

Section 6.3.3

6.26 L'Algorithme 6.2 d'échantillonnage de Gibbs bivarié est fondé sur les lois conditionnelles $\pi(\theta|\lambda)$ et $\pi(\lambda|\theta)$. Comme le décrit la Section 6.3, il consiste à simuler successivement $\pi(\theta|\lambda)$ et $\pi(\lambda|\theta)$. Cet exercice démontre qu'une telle simulation de $\pi(\theta,\lambda)$ se justifie d'un point de vue probabiliste.

a. Exprimer la loi jointe $\pi(\theta,\lambda)$ en fonction de ces lois conditionnelles.

b. Pour deux fonctions $q(\theta|\lambda)$ et $s(\lambda|\theta)$ données, fournir une condition nécessaire et suffisante pour que q et s soient proportionnelles à des lois conditionnelles.

c. Traiter les questions ci-dessus dans le cas de n niveaux pour les modèles complétés, c'est-à-dire lorsque des lois conditionnelles sont disponibles pour $\theta, \lambda_1, \ldots, \lambda_{n-1}$.

6.27 (Suite de l'Exercice 6.26) Le théorème de Hammersley-Clifford établit que la loi jointe $\pi(\vartheta)$ d'un vecteur $\vartheta = (\theta_1, \ldots, \theta_p)$ peut être obtenue à partir des lois conditionnelles complètes, $\pi_j(\theta_j | \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p)$. Montrer que

$$\pi(\vartheta) \propto \prod_{j=1}^{p} \frac{\pi_{\ell_j}(\theta_{\ell_j} | \theta_{\ell_1}, \ldots, \theta_{\ell_{j-1}}, \theta'_{\ell_{j+1}}, \ldots, \theta'_{\ell_p})}{\pi_{\ell_j}(\theta'_{\ell_j} | \theta_{\ell_1}, \ldots, \theta_{\ell_{j-1}}, \theta'_{\ell_{j+1}}, \ldots, \theta'_{\ell_p})}$$

pour toute permutation ℓ de $\{1, 2, \ldots, p\}$ et tout $\theta' \in \Theta$. [*Note :* Clifford et Hammersley n'ont jamais publié ce résultat ; voir Hammersley, 1974, et Robert et Casella, 2004, Section 9.1.4, pour plus de détails.]

6.28 *(Diebolt et Robert, 1994) Soient deux chaînes de Markov $(\theta^{(m)})$ et $(\lambda^{(m)})$ utilisées en échantillonnage de Gibbs bivarié, pour les lois conditionnelles $\pi_1(\theta|x, \lambda)$ et $\pi_2(\lambda|x, \theta)$.

a. Montrer que les noyaux de transition de ces chaînes sont respectivement

$$K(\theta'|\theta) = \int_{\Lambda} \pi_1(\theta'|x, \lambda)\pi_2(\lambda|x, \theta)\, d\lambda,$$
$$\text{et } H(\lambda'|\lambda) = \int_{\Theta} \pi_2(\lambda'|x, \theta)\pi_1(\theta|x, \lambda)\, d\theta.$$

b. Montrer que $\pi_1(\theta|x)$ et $\pi_2(\lambda|x)$ sont bien stationnaires pour ces noyaux.

c. Établir que, si $\theta^{(m)} \sim \pi_1^m(\theta|x, \lambda^{(0)})$ et $\lambda^{(m)} \sim \pi_2^m(\lambda|x, \lambda^{(0)})$,

$$||\pi_1^m(\cdot|x, \lambda^{(0)}) - \pi_1(\cdot|x)||_1 \leq ||\pi_2^m(\cdot|x, \lambda^{(0)}) - \pi_2(\cdot|x)||_1.$$

d. Déduire le Lemme 6.17 à partir de la question c. et du fait qu'une chaîne de Markov irréductible admettant une distribution stationnaire est ergodique. Montrer que, si $(\lambda^{(m)})$ est géométriquement ergodique de taux ϱ, $(\theta^{(m)})$ converge aussi au taux ϱ, soit

$$||\pi_1^m(\cdot|x, \lambda^{(0)}) - \pi_1(\cdot|x)||_1 \leq C\varrho^m.$$

e. La chaîne $(\lambda^{(m)})$ est φ-*mélangeante* s'il existe φ, décroissant géométriquement, et une mesure finie μ telle que

$$\left|\pi_2^m(\lambda|x, \lambda^{(0)}) - \pi_2(\lambda|x)\right| \leq \varphi(m)\mu(\lambda).$$

Montrer que, lorsque $(\lambda^{(m)})$ est φ-mélangeante,

$$|\pi_1^m(\theta|x, \lambda^{(0)}) - \pi_1(\theta|x)| \leq \varphi(m) \int_{\Lambda} \pi_1(\theta|x, \lambda)\mu(d\lambda)$$

et en déduire que, si Λ est compact, $(\theta^{(m)})$ est aussi φ-mélangeante.

f. De même, montrer que la convergence géométrique de $(\lambda^{(m)})$ et le fait que Λ soit compact sont des conditions suffisantes pour que, pour toute fonction h satisfaisant

$$\mathbb{E}^{\pi}[||h(\theta)||^2|x,\lambda] < \infty\,,$$

il existe C_h tel que

$$||\,\mathbb{E}^{\pi_1^m}[h(\theta)|x,\lambda^{(0)}] - \mathbb{E}^{\pi_1}[h(\theta)|x]\,||^2 \leq C_h\varrho^m.$$

g. Tirer profit du fait que, *lorsque Λ est fini, la chaîne $(\lambda^{(m)})$ est nécessairement géométriquement convergente et φ-mélangeante* (Billingsley, 1985). Déterminer l'importance des résultats ci-dessus dans le cadre de l'estimation d'un mélange.

h. Étendre le principe de dualité au cas d'un modèle hiérarchique à niveaux multiples, en utilisant le fait que les lois conditionnelles ne dépendent que des niveaux voisins.

6.29 Deux machines sont utilisées en parallèle ; les temps jusqu'à la première panne sont respectivement $x \sim f(x|\theta)$ et $y \sim g(y|\eta)$. On sait quelle machine est en panne lorsqu'une panne a lieu.

a. Donner la loi de z, temps jusqu'à la première panne du système, et construire un algorithme d'échantillonnage de Gibbs afin d'obtenir des estimateurs de Bayes de θ et η lorsqu'un échantillon $z_1,\ldots,z_n$ est disponible et lorsque des lois a priori conjuguées sont utilisées à la fois pour θ et pour η.

b. Mettre en œuvre cet algorithme dans les cas particuliers (a) f et g sont des densités normales de moyennes θ et η, et de variance 1 ; (b) f et g sont des lois exponentielles de paramètres θ et η.

Section 6.3.4

6.30 Pour une chaîne $(\theta^{(t)},\lambda^{(t)})$ produite par échantillonnage de Gibbs bivarié

a. Montrer que, pour toute fonction h,

$$\text{cov}(h(\theta^{(1)}),h(\theta^{(2)})) = \text{var}\left\{\mathbb{E}[h(\theta)|\lambda]\right\}\,.$$

b. Donner une représentation correspondante pour $\text{cov}(h(\theta^{(1)}),h(\theta^{(t)}))$.

c. En déduire que la covariance $\text{cov}(h(\theta^{(1)}),h(\theta^{(t)}))$ est toujours positive et décroissante en t.

d. Conclure sur la domination de la moyenne usuelle par sa version Rao-Blackwellisée.

6.31 Montrer que, dans le cadre de l'Exemple 6.15, les lois marginales de θ et λ ne peuvent pas être calculées explicitement et que, de plus, il faut que $B < +\infty$ pour que les lois marginales soient définies.

Section 6.3.5

6.32 Pour un modèle hiérarchique comme (6.21), montrer que la loi d'un λ_i donné, conditionnellement à tous les autres paramètres du modèle $\pi(\lambda_i|x,\theta,(\lambda_j)_{j\neq i})$ $(1 \leq i \leq p)$ ne dépend que de ses deux voisins les plus proches dans le vecteur $(x,\theta,\lambda_1,\ldots,\lambda_p)$. (*Indication :* Faire une représentation graphique du modèle.)

6.33 Montrer que, si l'échantillonneur de Gibbs est mis en œuvre avec plus de deux niveaux conditionnels, comme pour, par exemple, (6.23), les sous-chaînes résultantes correspondant aux différents niveaux ne sont pas des chaînes de Markov.

6.34 Pour le modèle multinomial de l'Exemple 6.21, expliquer pourquoi simuler $\pi((\mu,\eta)|x)$ plutôt que $\pi(\mu|x,\eta)$ et $\pi(\eta|x,\nu)$ devrait accélérer la convergence. (*Indication :* Étudier la corrélation entre $\mu^{(t)}$ et $\mu^{(t+1)}$ dans les deux cas.)

6.35 Montrer que, pour un algorithme d'échantillonnage de Gibbs, si une étape de simulation arbitraire, telle que, disons, la simulation de $\pi(\theta_1|\theta_2,\ldots,\theta_k)$, est remplacée par une étape *unique* de Metropolis-Hastings, la validité de l'algorithme est préservée. Commenter l'intérêt capital de cette propriété dans la pratique.

6.36 Soit une loi $\pi(\theta_1,\theta_2)$ non disponible analytiquement, mais telle que les deux lois conditionnelles $\pi(\theta_1|\theta_2)$ et $\pi(\theta_2|\theta_1)$ soient connues et puissent être simulées.

a. Montrer qu'il est possible de mettre en œuvre l'algorithme de Metropolis-Hastings. (*Indication :* Montrer que la seule difficulté est de simuler $\pi(\theta_1)$ ou $\pi(\theta_2)$, et utiliser l'Exercice 6.26.)

b. En déduire que l'échantillonnage de Gibbs peut s'appliquer dans tous les cas, tout comme la forme générale de l'algorithme de Metropolis-Hastings.

6.37 Montrer qu'une étape d'échantillonnage de Gibbs est un cas particulier de l'algorithme de Metropolis tel que la probabilité d'acceptation soit toujours égale à 1.

Section 6.3.6

6.38 On cherche à simuler une loi normale tronquée $\mathcal{N}_p(0,I_p)$ restreinte au polygone $\theta_i^t x_i \le z_i$ $(1\le i\le n)$.

a. Donner la loi de θ_j conditionnelle à θ_k $(k\neq j)$ et construire un échantillonneur de Gibbs pour la simulation de cette loi normale tronquée. (*Indication :* Voir Geweke, 1991, ou Robert, 1995, pour des algorithmes d'acceptation-rejet de simulation d'une loi normale tronquée unidimensionnelle.)

b. Proposer un algorithme alternatif de Metropolis-Hastings fondé sur la simulation d'une loi $\mathcal{N}_p(\mu,\Sigma)$, où μ et Σ sont calculés à partir des frontières du polygone.

c. Proposer un échantillonneur par tranche faisant intervenir une seule variable auxiliaire et un autre en faisant intervenir p.

d. Comparer ces différents algorithmes.

Section 6.3.7

6.39 (Rubin *et al.*, 1992) Une étude a été menée sur le campus de l'Université Cornell afin de modéliser le comportement sexuel des étudiants de premier et second cycles. Sur une population de R_m (R_f) étudiants masculins (féminins), r_m (r_f) ont répondu à l'enquête et t_m (t_f) ont déclaré être actifs sexuellement (durant les deux derniers mois).

a. Les premières quantités d'intérêt sont T_f et T_m, nombres d'étudiants féminins et masculins sexuellement actifs. En utilisant un modèle hypergéométrique sur t_m, et en supposant t_f, r_m et r_f fixés, calculer un estimateur de Bayes de T_f et T_m pour

$$T_i \sim \mathscr{B}(R_i, p_i), \qquad p_i \sim \mathscr{B}e(\alpha, \beta), \qquad \pi(\alpha, \beta) = 1/\alpha\beta \qquad (i = f, m).$$

(*Application numérique :* $R_f = 5\,211$, $r_f = 253$, $t_f = 111$, $R_m = 6\,539$, $r_m = 249$ et $t_m = 22$.)

b. Durant cette enquête, les répondants sexuellement actifs étaient interrogés sur le nombre de partenaires qu'ils ont eu pendant les deux derniers mois, y_f et y_m, ainsi que le nombre de partenaires étudiants de Cornell, x_m et x_f.

Considérant une loi de Poisson $\mathscr{P}(\lambda_i)$ pour le nombre de partenaires supplémentaires $y_i - 1$ et une loi binomiale $\mathscr{B}(y_i, \varrho_i)$ pour le nombre de partenaires étudiants de Cornell ($i = f, m$), avec $\varrho_f = T_m/N_m$ et $\varrho_m = T_f/N_f$, calculer l'estimateur de Bayes de la population en contact sexuel avec les étudiants de Cornell, N_m et N_f, pour les lois a priori

$$\lambda_i \sim \mathscr{E}xp(\lambda_0), \qquad \varrho_i \sim \mathscr{B}e(\gamma, \delta), \qquad \pi(\gamma, \delta) = 1/\gamma\delta.$$

(*Application numérique :* $y_m = 54$, $x_m = 31$, $y_f = 135$, $x_f = 67$.)

c. Comparer vos résultats avec l'estimateur du maximum de vraisemblance obtenu dans cette étude : $\hat{N}_f = 4\,186$, $\hat{N}_m = 1\,473$, $\hat{T}_f = 2\,323$ et $\hat{T}_m = 615$.

d. Reprendre l'estimation pour les lois a priori sur les hyperparamètres

$$\pi(\alpha, \beta) = e^{-(\alpha+\beta)}, \qquad \pi(\gamma, \delta) = e^{-(\gamma+\delta)},$$

et

$$\pi(\alpha, \beta) = 1/(\alpha + \beta)^2, \qquad \pi(\gamma, \delta) = 1/(\gamma + \delta)^2.$$

6.40 Dans le cas de la régression logistique (voir l'Exercice 6.23), une structure de données manquantes peut être mise en évidence et utilisée dans un algorithme de Gibbs.

a. Calculer la loi de z_i telle que l'observation y_i est $\mathbb{I}_{z_i \le x_i^t \theta}$.

b. Donner la vraisemblance du modèle complété et déterminer si un algorithme de Gibbs similaire à ceux de la Section 6.4 peut être construit dans le cas particulier $\theta \sim \mathscr{N}_p(\mu, \Sigma)$.

c. Comparer la performance de cet algorithme avec celle d'un algorithme de Metropolis-Hastings plus simple de votre choix.

6.41 Un modèle *probit* est un modèle de régression qualitative où la dépendance sur les variables auxiliaires est donnée par

$$P_\theta(y_i = 1) = 1 - P_\theta(y_i = 0) = \Phi(\theta^t x_i).$$

a. Montrer que, comme dans l'Exercice 6.40, il est possible de compléter le modèle en exhibant une variable latente continue z_i.

b. Proposer un algorithme d'échantillonnage de Gibbs fondé sur les données complétées lorsque $\theta \sim \mathscr{N}_p(\mu, \Sigma)$.

Section 6.4

6.42 (Casella *et al.*, 2000) L'échantillonnage de Gibbs et les autres méthodes MCMC ont résolu les difficultés de l'inférence bayésienne sur des modèles de mélange. Il est cependant possible de produire des estimateurs d'importance dans ce cadre. Nous supposons qu'un échantillon $(x_1, \ldots, x_n)$ de

$$\sum_{j=1}^{k} p_j f(x|\theta_j)$$

est disponible.

a. Considérant les variables d'allocation $z_1, \ldots, z_n$, où $x_i|z_i \sim f(x|\theta_{z_i})$, montrer que la loi a posteriori de $\mathbf{z} = (z_1, \ldots, z_n)$ est donnée par

$$P(\mathbf{z}|\mathbf{x}) = \frac{\displaystyle\prod_{j=1}^{k} \int_\Theta \prod_{\{i:z_i=j\}} f(x_i|\theta_j)\pi_j(\theta_j)d\theta_j}{\displaystyle\sum_{\mathbf{z}\in\mathscr{Z}} \prod_{j=1}^{k} \int_\Theta \prod_{\{i:z_i=j\}} f(x_i|\theta_j)\pi_j(\theta_j)d\theta_j\,,} \tag{6.29}$$

où $\mathscr{Z}$ est l'ensemble des k^n vecteurs d'allocation $\mathbf{z}$.

b. Montrer que

$$P(Z_i = j|x_i) = \frac{p_j m_j(x_i)}{\sum_{j=1}^{k} p_j m_j(x_i)}\,, \tag{6.30}$$

où $m_j(x) = \int f(x|\theta_j)\pi(\theta_j)d\theta_j$, $(j = 1, \ldots, m)$ est la loi marginale univariée de x_i.

c. En déduire que, si les expressions de (6.29) et (6.30) sont toutes les deux disponibles, à une constante de normalisation près, l'estimateur de Bayes $\mathbb{E}[h(\theta)|\mathbf{x}]$ peut être approché par échantillonnage d'importance, les z_i $(i = 1, \ldots, n)$ étant générés à partir des lois marginales de b., si l'expression de $\mathbb{E}[h(\theta)|(x_1, z_1), \ldots, (x_n, z_n)]$ est elle aussi connue.

d. Appliquer au cas d'un mélange de lois exponentielles,

$$\sum_{j=1}^{k} p_j \lambda_j \exp(-\lambda_j x), \qquad x > 0,$$

pour la loi a priori

$$\lambda_j \sim \mathscr{G}(\alpha_j, \beta_j), \qquad j = 1, \ldots, k\,,$$

lorsque les poids p_j et les hyperparamètres α_j, β_j sont connus. En particulier, déterminer des transformations $h(\lambda_1, \ldots, \lambda_k)$ telles que les espérances conditionnelles $\mathbb{E}[h(\theta)|(x_1, z_1), \ldots, (x_n, z_n)]$ soient connues.

6.43 Pour un mélange gaussien, détailler le raisonnement menant aux lois conditionnelles (6.26) et (6.27) et donner une expression explicite de $\mathbb{E}^\pi[\mu_i|x, z]$.

6.44 (Suite de l'Exercice 6.5) Une approche simplifiée des mélanges est de considérer qu'un mélange à k composantes n'est qu'une perturbation d'un mélange à $(k-1)$ composantes (Mengersen et Robert, 1996, Robert et Mengersen, 1999) et d'estimer un mélange à k composantes séquentiellement en k.

a. Écrire un programme MCMC à cet effet, qui estime uniquement la *nouvelle* composante dans le mélange à k composantes.

b. Comparer par des simulations les performances de cette version approchée avec une estimation directe du mélange à k composantes.

6.45 Pour une petite taille d'échantillon, effectuer plusieurs simulations pour comparer l'échantillonnage de Gibbs avec un calcul direct de l'estimateur de Bayes pour un mélange de deux lois normales.

6.46 Montrer que les lois a priori conjuguées ne peuvent pas donner une réponse non informative dans le cas d'un mélange gaussien à deux composantes lorsque les variances des lois a priori tendent vers $+\infty$.

6.47 (Robert et Soubiran, 1993) Obtenir les formules équivalentes à (6.26) et (6.27) pour un mélange de lois normales multidimensionnelles. (*Indication :* Utiliser la Section 4.4.1 pour le choix d'une loi a priori conjuguée et détailler la simulation de la loi de Wishart.)

6.48 (Binder, 1978) Soit un échantillon $x_1, \ldots, x_n$ tiré d'un mélange

$$x \sim f(x|\theta) = \sum_{i=1}^{k} p_i f_i(x),$$

tel que les densités f_i et les poids p_i soient connus. Le problème est d'identifier l'origine des observations, $g = (g_1, \ldots, g_n)$, avec

$$g_j = \sum_{i=1}^{k} i \mathbb{I}_{z_{ij}=1} \qquad (1 \leq j \leq n).$$

a. Montrer que des difficultés de calcul ont aussi lieu dans ce cadre, pour l'obtention des estimateurs de Bayes.

b. Donner l'estimateur de Bayes de g lorsque $p \sim \mathscr{D}(1/2, \ldots, 1/2)$ et $f_i(x) = \varphi(x; \mu_i, 1)$ avec $\mu_i \sim \mathscr{N}(\xi_i, 1)$.

c. Comment mettre en œuvre l'échantillonnage de Gibbs pour ce problème ?

6.49 Adapter les méthodes d'échantillonnage de Gibbs développées dans la Section 6.4 pour un mélange de lois au cas d'un *modèle censuré*, c'est-à-dire pour des observations y_i^* telles que

$$y_i^* = \begin{cases} y_i & \text{si } y_i \leq c, \\ c & \text{sinon,} \end{cases}$$

si $y_i \sim f(y|\theta)$, où $f(\cdot|\theta)$ appartient à une famille exponentielle.

6.50 (Robert *et al.*, 1993a) Un *modèle de chaîne de Markov cachée* généralise le modèle de mélange étudié dans l'Exemple 6.5 et dans la Section 6.4 en introduisant une certaine dépendance entre les observations $x_1, \ldots, x_t$. Si on complète ces observations par les variables indicatrices (inconnues) des états z_i, le modèle devient hiérarchique $(1 \leq i \leq t)$:

$$x_i|z_i, \theta \sim f(x|\theta_{z_i})$$

et (z_i) constitue une chaîne de Markov sur $\{1, \ldots, K\}$ de matrice de transition $\mathbb{P} = (p_{jk})$, où

$$p_{jk} = P(z_i = k | z_{i-1} = j) \qquad (2 \le i \le t)$$

(on pose $z_1 = 1$ pour des raisons d'identifiabilité). On suppose de plus que $f(\cdot|\theta)$ appartient à une famille exponentielle.

a. Donner la vraisemblance de ce modèle et en déduire que ni le maximum de vraisemblance ni l'estimation bayésienne sous des lois conjuguées sur θ et $\mathbb{P}$ ne donnent des expressions explicites dans ce cas.

b. Considérant le cas particulier où $f(\cdot|\theta)$ est $\mathcal{N}(\xi, \sigma^2)$ avec $\theta = (\xi, \sigma^2)$, montrer qu'un échantillonneur de Gibbs comprenant des simulations itératives de $\pi(\theta|\mathbf{x}, \mathbf{z})$ et $\pi(\mathbf{z}|\mathbf{x}, \theta)$ est relativement coûteux en temps de calcul, à cause de $\pi(\mathbf{z}|\mathbf{x}, \theta)$.

c. Montrer que les lois conditionnelles complètes $\pi(z_i|\mathbf{x}, \theta, z_{j\neq i})$ ne dépendent que de z_{i-1} et z_{i+1} et sont beaucoup plus faciles à simuler.

d. Proposer un algorithme d'échantillonnage de Gibbs pour ce modèle. Montrer que la condition $p_{kj} > 0$ pour tout $1 \le j, k \le K$ est suffisante pour assurer la convergence géométrique des chaînes $(\theta^{(m)})$ et $(\mathbf{P}^{(m)})$ vers les vraies lois a posteriori. (*Indication :* Des arguments similaires à ceux de l'Exercice 6.28 peuvent être utilisés.)

6.51 (Robert *et al.*, 1999a) Dans le cadre de l'Exercice 6.50, il existe une façon de simuler la chaîne complète $\mathbf{z} = (z_2, \ldots, z_n)$ conditionnellement aux paramètres θ, et donc de mettre en œuvre une technique d'augmentation de données. La représentation de la loi conditionnelle de $\mathbf{z}$ est appelée *récurrences avant-arrière* (ou *forward-backward*) et est connue depuis longtemps en traitement du signal (Baum et Petrie, 1966).

a. Établir la *relation dite de récurrence arrière* $(1 \le i \le n-1)$

$$f(x_i, \ldots, x_n|\theta, z_i = j) = \sum_{k=1}^{K} p_{jk} f(x_i|\theta_j) f(x_{i+1}, \ldots, x_n|\theta, z_{i+1} = k)\,, \qquad (6.31)$$

avec $f(x_n|z_n = j) = f(x_n|\theta_j)$.

b. Calculer à partir de la formule de récurrence arrière la probabilité $P(z_1 = j|x_1, \ldots, x_n, \theta)$ sous l'hypothèse que z_1 est distribuée marginalement selon la loi stationnaire associée à la matrice de transition $\mathbb{P}$.

c. Calculer les probabilités $P(z_i = j|x_1, \ldots, x_n, \theta, z_1, \ldots, z_{i-1})$ $(i = 2, \ldots, n)$.

d. En conclure que le vecteur $(z_1, \ldots, z_n)$ peut être simulé conditionnellement aux observations et θ en un temps $\mathrm{O}(nK^2)$ et donc que la technique d'augmentation de données peut être mise en œuvre dans certains modèles de chaînes de Markov cachées.

6.52 Dans un cadre de mélange, comparer les performances (en termes de temps de calcul) de l'échantillonnage de Gibbs avec celui d'un algorithme de Metropolis-Hastings par marche aléatoire.

Note 6.6.3

6.53 La décomposition d'une loi du khi deux décentré proposée dans l'Exemple 6.26 permet-elle une mise en œuvre de l'échantillonnage de Gibbs ? Donner une approximation par l'algorithme de Metropolis-Hastings.

6.54 (Heitjan et Rubin, 1991) *Des données grossières* sont définies comme une agrégation d'observations en classes. Pour une variable aléatoire "complète" $y_i \sim f(y|\theta)$, prenant ses valeurs dans $\mathscr{Y}$, et une partition A_j $(j \in I)$ de $\mathscr{Y}$, les observations sont $x_i = j$ si $y_i \in A_j$.

a. Donner une illustration concrète de ce modèle.

b. Proposer un algorithme d'échantillonnage de Gibbs dans le cas où $f(\cdot|\theta)$ est une loi normale $\mathscr{N}(\xi, \sigma^2)$ avec $\theta = (\xi, \sigma^2)$ et $A_j = [j, j+1)$ $(j \in \mathbb{Z})$.

Le nombre de passages de voitures durant une période d'une minute a été observé pendant trois cent soixante minutes consécutives ; les observations résultantes sont données dans la Table 6.2.

c. En posant une loi de Poisson $\mathscr{P}(\theta)$ sur le nombre de passages, appliquer l'échantillonnage de Gibbs afin d'estimer le paramètre θ pour ce jeu de données et la loi a priori $\pi(\theta) = 1/\theta$.

Tab. 6.2. Nombre de passages de voitures pour une suite d'intervalles d'une minute.

Nombre de voitures	0	1	2	3	4 ou plus
Nombre de passages	139	128	55	25	13

Note 6.6.4

6.55 Dans le cadre de l'Exemple 6.19,

a. Montrer que la loi marginale associée aux lois conditionnelles complètes $\pi(\theta|\lambda)$ et $\pi(\lambda|\theta)$ satisfait

$$\frac{\pi(\theta)}{\pi(\lambda)} = \frac{\theta}{\lambda}, \qquad \theta, \lambda < B\,.$$

b. En déduire que la loi jointe correspondant à ces deux lois conditionnelles n'est pas définie lorsque B tend vers l'infini.

Note 6.6.6

6.56 Pour la suite $(\hat{\theta}_{(j)})_j$ produite par l'algorithme EM,

a. Montrer que

$$Q(\hat{\theta}_{(j+1)}|\hat{\theta}_{(j)}, \mathbf{x}) \geq Q(\hat{\theta}_{(j)}|\hat{\theta}_{(j)}, \mathbf{x}).$$

b. On note $k(\mathbf{z}|\theta, \mathbf{x})$ la loi conditionnelle de $\mathbf{z}$ sachant $\mathbf{x}$. Montrer que

$$\mathbb{E}_{\hat{\theta}_{(j)}}\left[\log\left(\frac{k(\mathbf{z}|\hat{\theta}_{(j+1)}, \mathbf{x})}{k(\mathbf{z}|\hat{\theta}_j, \mathbf{x})}\right)|\hat{\theta}_{(j)}, \mathbf{x}\right] \leq 0\,.$$

(*Indication :* Utiliser l'inégalité de Jensen.)

c. Conclure que

$$L(\hat{\theta}_{(j+1)}|\mathbf{x}) \geq L(\hat{\theta}_{(j)}|\mathbf{x}),$$

l'égalité étant vérifiée si et seulement si $Q(\hat{\theta}_{(j+1)}|\hat{\theta}_{(j)}, \mathbf{x}) = Q(\hat{\theta}_{(j)}|\hat{\theta}_{(j)}, \mathbf{x})$.

6.6 Notes

6.6.1 Générateurs uniformes pseudo-aléatoires.

Tout algorithme de génération d'une variable aléatoire de loi quelconque repose sur la génération de *variables aléatoires uniformes* sur $[0,1]$. Puisque la production exacte d'une suite iid de variables uniformes $\mathscr{U}([0,1])$ est impossible, il existe des méthodes reposant sur un mécanisme purement déterministe produisant des suites imitant le comportement d'une suite de variables iid $\mathscr{U}([0,1])$, au sens où cette suite déterministe est acceptée comme une suite iid $\mathscr{U}([0,1])$ par tout test statistique. Par exemple, le générateur proposé par Ripley (1987) est de type *congruentiel*, et est défini comme suit.

ALGORITHME 6.9. –**Générateur congruentiel**–

1. Initialiser avec une racine initiale arbitraire x_0
2. Itérer

$$x_i = (69069x_{i-1} + 1) \bmod 2^{32},$$
$$u_i = 2^{-32}x_i.$$

La suite correspondante des u_i peut être considérée comme une suite iid $\mathscr{U}_{[0,1]}$, bien que son support soit en réalité fini.

Des générateurs uniformes pseudo-aléatoires sont disponibles sur la plupart des ordinateurs et dans la plupart des langages informatiques, et peuvent être utilisés en tant que tels, même si certains de ces générateurs ne sont pas testés exhaustivement et peuvent avoir des propriétés indésirables (voir Robert et Casella, 1999, Exercice 2.5).

Marsaglia et Zaman (1993) ont développé un générateur uniforme simple à racines multiples dont la période est supérieure à 2^{95} ; voir Robert et Casella (2004, Note 2.6.1) pour plus de détails.

6.6.2 Les logiciels `BUGS` *et* `CODA`

Spiegelhalter *et al.* (1995a,b,c) de la MRC Biostatistics Unit de Cambridge, en Angleterre, ont développé un logiciel MCMC. Ce logiciel offre différentes possibilités pour programmer un échantillonneur de Gibbs partiellement automatique (`BUGS` signifie *Bayesian inference Using Gibbs Sampling*). Il s'agit d'un langage informatique, ressemblant au `C` ou à `R`, et fondé sur des déclarations sur le modèle, les données et les spécifications a priori, éventuellement hiérarchiques ; ce langage autorise une grande variété de transformations de la plupart des distributions standard. `BUGS` produit un échantillon de Gibbs, fait de valeurs simulées des paramètres, après un nombre arbitraire d'itérations d'échauffement, et pour un intervalle entre valeurs retenues lui aussi arbitraire.

Une restriction importante sur la modélisation a priori est que des lois a priori conjuguées ou des densités log-concaves doivent être utilisées pour permettre soit une simulation standard, soit l'utilisation de l'algorithme ARMS de Gilks *et al.* (1995), mais des lois plus complexes peuvent être prises en compte par

discrétisation de leur support. L'autre restriction est que des lois a priori impropres ne peuvent pas être utilisées et doivent être remplacées par des lois a priori vagues, c'est-à-dire de grande variance a priori.

Le logiciel `BUGS` se complète d'un logiciel de diagnostic de convergence[56], `CODA`, qui comporte les méthodes d'évaluation de convergence MCMC les plus courantes. Ce "package" `S-Plus` a été développé par Best *et al.* (1995) et peut être utilisé indépendamment de `BUGS`. Les méthodes mises en œuvre dans `CODA` sont décrites dans Robert et Casella (2004, Chapitre 12) : elles incluent les diagnostics de convergence de Gelman et Rubin (1992), Geweke (1992), Heidelberger et Welch (1983), Raftery et Lewis (1992a), ainsi que les tracés d'autocorrélation pour chaque variable et les corrélations croisées entre variables.

6.6.3 Mélanges cachés

La décomposition hiérarchique (6.19) sur laquelle repose l'échantillonnage de Gibbs est aussi utile pour la sélection de la loi a priori, lorsque la distribution d'échantillonnage n'appartient pas à une famille exponentielle et qu'il n'existe pas de loi a priori conjuguée. C'est le cas par exemple pour les lois de Student et du khi deux décentré. Une décomposition de $f(x|\theta)$ de la forme

$$f(x|\theta) = \int f(x|\theta, z)g(z|\theta)\, dz$$

peut alors permettre une modélisation a priori de θ via des lois a priori conjuguées (pour $f(x|\theta, z)$ ou $g(z|\theta)$). Comme dans la Section 3.3.3, nous appelons cette représentation *mélange caché*, pour marquer la différence avec les problèmes de mélanges standard pour lesquels la structure de mélange elle-même est d'intérêt ; voir aussi la Note 3.8.3.

Exemple 6.26. Soit $x \sim \chi^2_p(\theta)$, une observation tirée d'une loi du khi deux décentré.Cette loi peut s'écrire comme le mélange

$$\begin{aligned} x|\theta, z &\sim \chi^2_{p+2z}, \\ z|\theta &\sim \mathcal{P}(\theta/2). \end{aligned}$$

Donc seul $g(z|\theta)$ dépend de θ et une loi a priori possible pour θ est $\mathcal{G}(\alpha, \beta)$, puisqu'il s'agit de la loi conjuguée pour la loi de Poisson. ‖

Exemple 6.27. Soit $x|\mu, \sigma \sim \mathcal{T}(m, \mu, \sigma^2)$, avec $\theta = (\mu, \sigma)$ inconnu. En se fondant sur la représentation de Dickey (1968),

$$x|\theta, z \sim \mathcal{N}(\mu, z), \qquad z|\sigma^2 \sim \mathcal{IG}(m/2, m\sigma^2/2),$$

on peut proposer

$$\mu \sim \mathcal{N}(\xi, \tau^2), \qquad \sigma^2 \sim \mathcal{G}(\alpha, \beta),$$

comme loi a priori et on obtient

[56]Ces deux logiciels sont actuellement disponibles sur le site de la MRC Biostatistics Unit, à l'adresse `www.mrc-bsu.cam.ac.uk`.

$$\begin{aligned} z|x,\theta &\sim \mathscr{IG}\left(\frac{m+1}{2}, \frac{m\sigma^2+(x-\mu)^2}{2}\right), \\ \sigma^2|x,z &\sim \mathscr{G}(\alpha+(m/2), \beta+(m/2z)), \\ \mu|x,z &\sim \mathscr{N}\left(\frac{z\mu+\tau^2 x}{z+\tau^2}, \frac{z\tau^2}{z+\tau^2}\right). \end{aligned} \tag{6.32}$$

Les lois conditionnelles (6.32) permettent directement une simulation par échantillonnage de Gibbs. Notons la différence avec l'exemple normal classique (voir la Section 4.4). Dans ce cas, σ^2 suit une loi a priori gamma plutôt qu'inverse gamma et, fait plus important, μ et σ sont a priori indépendants. La décomposition conditionnelle mène donc à une modélisation plus satisfaisante que dans le cas normal. ‖

Recourir à une structure de mélange caché pour $f(x|\theta)$ ou pour $\pi(\theta)$ simplifie bien entendu la simulation de $\pi(\theta|x)$ par échantillonnage de Gibbs lorsque la loi a posteriori n'est pas disponible.

Exemple 6.28. (Suite de l'Exemple 6.27) Si, dans un but de robustesse, la loi a priori est en fait

$$\mu \sim \mathscr{T}(\nu, \xi, \tau^2), \qquad \sigma^2 \sim \mathscr{G}(\alpha, \beta),$$

la représentation en mélange caché correspondante est

$$\mu|\delta \sim \mathscr{N}(\xi, \delta), \qquad \delta \sim \mathscr{IG}(\nu/2, \nu\tau^2/2),$$

et la simulation de $\pi(\mu, \sigma|x)$ peut être obtenue par échantillonnage de Gibbs, via les lois conditionnelles suivantes :

$$\begin{aligned} z|x,\theta &\sim \mathscr{IG}\left(\frac{m+\nu}{2}, \frac{m\sigma^2+(x-\mu)^2}{2}\right), \\ \sigma^2|x,z &\sim \mathscr{G}(\alpha+(m/2), \beta+(m/2z)), \\ \mu|x,z,\delta &\sim \mathscr{N}\left(\frac{\delta\mu+\tau^2 x}{\delta+\tau^2}, \frac{\delta\tau^2}{\delta+\tau^2}\right), \\ \delta|\theta &\sim \mathscr{IG}\left(\frac{\nu+1}{2}, \frac{\nu\tau^2+(x-\mu)^2}{2}\right). \end{aligned}$$

‖

6.6.4 Lois a posteriori impropres

Comme l'a souligné la Note 1.8.3, des lois a priori π qui satisfont

$$\int_\Theta \pi(\theta) f(x|\theta) d\theta = \infty$$

ne peuvent pas être utilisées. Cette condition est difficile à vérifier pour des modèles complexes et il existe de nombreuses situations où (a) une vérification analytique est impossible ; et (b) les lois conditionnelles obtenues à partir de $\pi(\theta) f(x|\theta)$ sont propres. Considérons, par exemple, le cas de l'Exemple 6.19 :

lorsque B tend vers l'infini, la loi jointe sur (θ, λ) n'est pas définie ; les lois conditionnelles sont cependant des lois exponentielles standard $\mathcal{E}xp(\lambda)$ et $\mathcal{E}xp(\theta)$ (Exercice 6.55). Une difficulté supplémentaire est qu'un échantillonneur de Gibbs fondé sur ces lois conditionnelles peut très bien ne pas mettre en évidence le caractère impropre de la loi a posteriori (voir Hobert et Casella, 1996).

Exemple 6.29. Soit le modèle à effets aléatoires usuel $(1 \le i \le I,\ 1 \le j \le J)$

$$y_{ij} = \theta + u_i + \epsilon_{ij}, \qquad u_i \sim \mathcal{N}(0, \sigma^2),\ \epsilon_{ij} \sim \mathcal{N}(0, \tau^2)\,.$$

La loi a priori de Jeffreys correspondante est $\pi(\theta, \tau^2, \sigma^2) = 1/\sigma^2\tau^2$. Alors (voir Robert et Casella, 2004, Exemple 10.31 et Problème 10.25), la loi a posteriori jointe de $(\theta, \tau^2, \sigma^2)$ n'est pas définie, tandis que les lois conditionnelles le sont et peuvent (hélas !) être utilisées dans un échantillonneur de Gibbs. ‖

Malgré l'impossibilité fondamentale d'utiliser pour une inférence bayésienne des loi a posteriori impropres, qui sont effectivement des mesures $f(x|\theta)\pi(\theta)$ de masse infinie, il existe des cas où de telles mesures peuvent être utiles. En particulier, il est possible d'augmenter artificiellement le paramètre θ par un paramètre auxiliaire α et d'introduire une loi a priori impropre $\pi(\alpha)$ telle que la loi a posteriori jointe $\pi(\alpha, \theta|x) = \pi(\alpha)\pi(\theta)f(x|\theta)$ soit aussi impropre, tout en préservant le caractère propre de la densité correctement définie $\pi(\theta|x)$ à l'intérieur de la chaîne de Markov.

Exemple 6.30. (Meng et Van Dyk, 1999) Une loi de Student t de paramètre $\theta = (\mu, \sigma)$, $\mathcal{T}(\nu, \mu, \sigma^2)$, peut s'écrire

$$x = \mu + \sigma y_1/(\nu y_2)^{1/2}, \quad \text{avec} \quad y_1 \sim \mathcal{N}(0,1),\ y_2 \sim \chi^2_\nu\,.$$

(voir l'Exercice 1.1 et l'Exemple 3.17). Si on introduit $\alpha > 0$ tel que

$$x|y_2 \sim \mathcal{N}(\mu, \alpha\sigma^2/(\nu y_2)), \qquad y_2 \sim \alpha\chi^2_\nu\,,$$

cela ne change pas le modèle étudié puisque la quantité α/y_2 ne dépend pas de α. Le paramètre α n'est donc pas identifiable et, pour une loi a priori sur α impropre, disons $\pi(\alpha) = \alpha^{-1}\exp(-\beta/\alpha)$, la loi a posteriori marginale de α est égale à sa loi a priori : la loi a posteriori jointe de (θ, α) n'est pas définie.

Il est cependant possible de créer une chaîne de Markov $(y_2^{(t)}, \theta^{(t)}, \alpha^{(t)})$ par une méthode simple d'augmentation de données, appliquée aux lois conditionnelles complètes obtenues à partir de

$$\pi(\alpha)\pi(\mu, \sigma)f(x|\mu, \alpha, \sigma, y_2)f(y_2|\alpha)$$

et telles que (a) cette mesure σ-finie soit stationnaire pour cette chaîne ; et (b) la sous-chaîne $(\theta^{(t)})$ converge vers la loi a posteriori bien définie $\pi(\theta|x)$. ‖

Les lois a posteriori impropres apparaissent alors comme des outils permettant d'accélérer l'exploration de l'espace des paramètres Θ par des chaînes de Markov nulles récurrentes ou même transientes, dans des espaces plus grands ; voir Casella (1996), Meng et Van Dyk (1999), Hobert (2000a,b), et Liu et Wu (1999) pour plus de détails.

6.6.5 Algorithmes MCMC dans des modèles dynamiques

Nous avons introduit dans la Section 4.5 divers modèles dynamiques et souligné le fait que la complexité de l'espace des paramètres induite par les contraintes de stationnarité ainsi que l'absence d'expression explicite pour la vraisemblance imposent le recours à des algorithmes MCMC. Les représentations à espace d'état des Sections 4.5.3 et 4.5.4 et la reparamétrisation du Lemme 4.24 jouent un rôle clé dans l'obtention d'échantillonneurs de Gibbs pour ces modèles.

Par exemple, dans le modèle AR(p), les ϱ_j ($1 \leq j \leq p$) sont des fonctions linéaires des autocorrélations partielles ψ_k ($1 \leq k \leq p$), lorsque les ψ_ℓ ($\ell \neq k$) sont fixés :

$$\varrho_j = a_{kj} + b_{kj}\psi_k \ ,$$

avec $(1 \leq \ell \leq i-1)$

$$\begin{aligned}
&a^{ii} = \psi_i,\ b^{ii} = 0,\ a^{i\ell} = a^{(i-1)\ell} - \psi_i a^{(i-1)(i-\ell)},\ b^{i\ell} = 0, \qquad \text{si} \quad i < k\\
&a^{ii} = 0,\ b^{ii} = 1,\ a^{i\ell} = a^{(i-1)\ell},\ b^{i\ell} = -a^{(i-1)(i-\ell)} \qquad \text{si} \quad i = k\\
&a^{ii} = \psi_i,\ b^{ii} = 0,\ a^{i\ell} = a^{(i-1)\ell} - \psi_i a^{(i-1)(i-\ell)},\ b^{i\ell} = b^{(i-1)\ell} - \psi_i b^{(i-1)(i-\ell)}\\
&\qquad\qquad \text{si} \quad i > k
\end{aligned}$$

et

$$a_{ik} = a^{pi} \ , \quad b_{ik} = b^{pi} \quad (1 \leq i \leq p)\,.$$

Donc, si les ψ_i sont simulés un par un, la vraisemblance (4.24) a une structure normale

$$\prod_{t=1}^{T} \exp\left\{-\frac{1}{2\sigma^2}\left(x_t - \mu - \sum_{j=1}^{p}(a_{ij} + b_{ij}\psi_i)(x_{t-j} - \mu)\right)^2\right\}\,.$$

Une décomposition conditionnelle similaire peut être utilisée pour les modèles MA(q) et ARMA(p,q) des Sections 4.5.3 et 4.5.4, en tirant profit de la structure linéaire de la représentation à espace d'état qui préserve la structure normale. Des solutions alternatives fondées sur la représentation récursive (4.28) et sur des étapes de Metropolis-Hastings ont été étudiées dans Billio *et al.* (1998).

6.6.6 Retour à l'estimation de mélange

L'importance des mélanges de distributions standard comme outils de modélisation ne peut pas être minimisée : ces modèles se situent à la frontière des modélisations paramétrique et non paramétrique et permettent la description de phénomènes plus complexes (relativement aux lois standard), tout en respectant le *principe de parcimonie* (c'est-à-dire permettant le recours à un nombre raisonnable de paramètres pour décrire un phénomène). Ce point est illustré par la construction de lois a priori dans les Notes 3.8.3 et 6.6.3. Les modèles de mélange apparaissent en analyse bayésienne non paramétrique, comme, par exemple, avec les processus de Dirichlet (voir les Notes 1.8.2 et 6.6.7). Ils jouent également un rôle important dans les problèmes de classification (voir Bensmail *et al.*, 1997) et en détection de valeurs aberrantes (Verdinelli et Wasserman, 1992).

Le traitement classique de l'estimation de mélanges finis de lois est présenté dans Titterington *et al.* (1985) et MacLachlan et Basford (1987). Il remonte à

Pearson (1894), qui proposa une méthode d'estimation fondée sur les moments et sur la résolution d'une équation polynomiale de degré 9.

Pour une estimation par maximum de vraisemblance, Dempster *et al.* (1977) et Redner et Walker (1984) ont développé un algorithme dit *algorithme EM* (pour *Expectation-Maximisation*) qui est extraordinairement populaire (voir Meng et Van Dyk, 1997 et MacLachlan et Krishnan, 1997). Cet algorithme est fondé sur la même augmentation de données que l'échantillonnage de Gibbs. Pour une vraisemblance complétée donnée $L^c(\theta|\mathbf{x},\mathbf{z})$, l'*algorithme EM* fonctionne comme suit.

ALGORITHME 6.10. –**Espérance-Maximisation (EM)**–
À l'itération m,

1. Calculer
$$Q(\theta|\hat{\theta}_{(m)},\mathbf{x}) = \mathbb{E}_{\hat{\theta}_{(m)}}[\log L^c(\theta|\mathbf{x},\mathbf{z})|\mathbf{x}]\,,$$
où l'espérance est par rapport à $k(\mathbf{z}|\hat{\theta}_m,\mathbf{x})$ (*étape E*) .
2. Maximiser $Q(\theta|\hat{\theta}_{(m)},\mathbf{x})$ en θ et prendre (*étape M*)
$$\theta_{(m+1)} = \arg\max_{\theta}\ Q(\theta|\hat{\theta}_{(m)},\mathbf{x}).$$

La validité de cet algorithme tient au fait que la vraisemblance *observée* augmente à chaque itération (Exercice 6.56). La suite $(\hat{\theta}_{(m)})_m$ converge donc vers un point stationnaire de la vraisemblance observée (qui peut être un maximum local ou un point-selle) ; voir Robert et Casella (1999, Section 5.3.3) pour plus de détails.

Puisque la convergence de l'algorithme EM dépend du point initial $\hat{\theta}_{(0)}$ et que cet algorithme requiert le calcul de l'espérance dans l'*étape E*, certains auteurs, notamment Broniatowski *et al.* (1983), Celeux et Diebolt (1990), Qian et Titterington (1991) et Lavielle et Moulines (1997), ont proposé des extensions stochastiques de l'algorithme EM.

D'un point de vue bayésien, une étude plus détaillée des méthodes MCMC pour les mélanges est proposée dans Robert (1996a), Roeder et Wasserman (1997), Robert et Mengersen (1999), Celeux *et al.* (2000), Stephens (2000) et Marin *et al.* (2004). En particulier, Celeux *et al.* (2000) montrent que l'ordre des paramètres utilisé pour assurer l'identifiabilité peut avoir des effets désastreux sur l'inférence résultante ; ces auteurs construisent des fonctions spécifiques de coût pour venir à bout du problème de non-identifiabilité.

L'échantillonnage de Gibbs et d'autres méthodes MCMC ont donc permis des améliorations considérables de l'approche bayésienne des modèles de mélange, non seulement pour leur estimation, comme nous l'avons expliqué ci-dessus, mais aussi pour les procédures de tests et la modélisation, puisque des tests bayésiens sur le nombre de composantes d'un mélange ont été proposés (Mengersen et Robert, 1996, Richardson et Green, 1997). De plus, ces études ont aussi mis en lumière des extensions non informatives intéressantes. Comme il est mentionné dans l'Exercice 1.56, les propriétés particulières des modèles de mélange empêchent l'utilisation de lois a priori impropres de la forme

$$\prod_{i=1}^{k} \pi_1(\mu_i, \sigma_i)\,.$$

En fait, dans la décomposition (6.6) de la loi a posteriori comme une somme sur toutes les partitions possibles, certaines de ces partitions n'attribuent aucune observation à une composante donnée i^* du mélange. La loi a priori sur les paramètres correspondants $(\mu_{i^*}, \sigma_{i^*})$ doit donc être *propre*.

Cependant, comme Mengersen et Robert (1996) l'ont montré, une loi a priori impropre peut malgré tout être utilisée si les paramètres de composante sont a priori dépendants. Par exemple, le modèle de mélange peut être reparamétrisé en termes d'un paramètre global de position-échelle (μ, τ), de loi a priori $\pi(\mu, \tau) = 1/\tau$. Dans ce cas, l'information a priori à fournir peut se réduire au choix d'un hyperparamètre unique $\xi > 0$. En effet, si (6.25) s'écrit

$$p_1 \mathcal{N}(\mu, \tau^2) + (1-p_1)\left\{p_2 \mathcal{N}(\mu + \tau\theta_1, \tau^2\sigma_1^2) + (1-p_2)\left\{p_3 \mathcal{N}(\mu + \tau\theta_1 + \tau\sigma_1\theta_2, \tau^2\sigma_1^2\sigma_2^2) + \ldots\right\}\right\},$$

une loi a priori acceptable est $p_i \sim \mathcal{B}e(1/2, 1/2)$, $\theta_i \sim \mathcal{N}(0, \xi^2)$ et $\sigma_i \sim (1/2)\mathcal{U}_{[0,1]} + (1/2)\mathcal{P}a(2, 1)$, cette dernière loi étant justifiée en tant que loi uniforme soit pour σ_i, soit pour $1/\sigma_i$; voir Roeder et Wasserman (1997) et Robert et Titterington (1998) pour des propositions similaires.

6.6.7 Échantillonnage de Gibbs pour les processus de Dirichlet

Nous avons mentionné dans la Note 1.8.2 l'intérêt de l'utilisation de processus de Dirichlet pour l'estimation bayésienne non paramétrique. Nous indiquons ici comment l'échantillonnage de Gibbs peut être mis en œuvre dans le cas gaussien. Soient $x_i \sim \mathcal{N}(\theta_i, \sigma_i^2)$ $(1 \leq i \leq n)$ avec $(\theta_i, \sigma_i^2) \sim \pi$ et π distribué comme un processus de Dirichlet $\mathcal{D}(\alpha, \pi_0)$. Comme nous l'avons déjà mentionné dans la Note 1.8.2, π_0 est l'espérance a priori de π et α est un degré de concentration autour de π_0. La loi marginale correspondante est un mélange de lois normales, dont le nombre de composantes est aléatoire et compris entre 1 et n. Le fait que le nombre de composantes puisse être aussi élevé que la taille d'échantillon reflète le caractère non contraignant de cette modélisation et peut être relié au fait que l'estimateur usuel *à noyau* recourt toujours à n composantes. Une autre conséquence importante de cette modélisation est que les lois a priori conditionnelles des (θ_i, σ_i^2) peuvent s'écrire

$$\pi[(\theta_i, \sigma_i^2)|(\theta_j, \sigma_j^2)_{j\neq i}] = \alpha(\alpha+n-1)^{-1}\pi_0(\theta_i, \sigma_i^2) + (\alpha+n-1)^{-1}\sum_{j\neq i}\mathbb{I}((\theta_i, \sigma_i^2) = (\theta_j, \sigma_j^2)). \tag{6.33}$$

La décomposition (6.33) met en évidence l'effet modérateur de l'a priori de Dirichlet : de nouvelles valeurs de (θ, σ^2) n'apparaissent qu'avec une probabilité $\alpha/(\alpha+n-1)$.

Une loi conditionnelle similaire peut être obtenue a posteriori, à savoir pour les observations $x_1, \ldots, x_n$,

$$\pi[(\theta_i, \sigma_i^2)|(\theta_j, \sigma_j^2)_{j\neq i}, x_i] = q_{i0}\pi_0(\theta_i, \sigma_i^2|x_i) + \sum_{j\neq i} q_{ij}\mathbb{I}((\theta_i, \sigma_i^2) = (\theta_j, \sigma_j^2)), \tag{6.34}$$

où $q_{i0} + \sum_{j \neq i} q_{ij} = 1$ et $(i \neq j)$

$$q_{i0} \propto \alpha \int e^{-(x_i-\theta_i)^2/2\sigma_i^2} \sigma_i^{-1} \pi_0(\theta_i, \sigma_i^2) d\theta_i d\sigma_i^2, \quad q_{ij} \propto e^{-(x_i-\theta_j)^2/2\sigma_j^2} \sigma_j^{-1}\,.$$

Pour les lois conditionnelles (6.34), (θ_i, σ_i^2) est un nouveau paramètre avec probabilité q_{i0} et est égal à un autre paramètre avec probabilité $1 - q_{i0}$. Donc, l'échantillonnage de Gibbs peut être mis en œuvre en simulant successivement ces lois conditionnelles pour chaque i et en proposant comme loi marginale pour $(x_1, \ldots, x_n)$ un mélange de k lois normales, où k est le nombre de valeurs différentes parmi les simulations (θ_i, σ_i^2). Notons que ce nombre k varie à chaque itération.

Une autre conséquence de cette représentation est que, si on s'intéresse à la densité prédictive f, il est possible de simuler un échantillon de taille T de la loi $\pi(\theta, \sigma^2 | x_1, \ldots, x_n)$, $(\theta^{(t)}, {\sigma^{(t)}}^2)$ $(t = 1, \ldots, T)$, en simulant successivement (θ_i, σ_i^2) $(1 \leq i \leq n)$ selon (6.34) et $(\theta_{n+1}, \sigma_{n+1}^2)$ selon

$$\begin{aligned}\pi(\theta_{n+1}, \sigma_{n+1}^2) &= \pi[(\theta_{n+1}, \sigma_{n+1}^2) | (\theta_{i \neq n+1}, \sigma_{i \neq n+1}^2)] \\ &= \alpha(\alpha + n)^{-1} \pi_0(\theta_{n+1}, \sigma_{n+1}^2) \\ &\qquad + (\alpha + n)^{-1} \sum_{j=1}^{n} \mathbb{I}((\theta_{n+1}, \sigma_{n+1}^2) = (\theta_j, \sigma_j^2)).\end{aligned}$$

La densité prédictive peut être alors estimée par

$$\frac{1}{T} \sum_{t=1}^{T} f(x | \theta^{(t)}, {\sigma^{(t)}}^2)\,, \tag{6.35}$$

et est donc du même ordre de complexité qu'un estimateur de la densité à noyau, puisqu'elle fait formellement intervenir T termes. En fait, l'échantillon des $(\theta^{(t)}, {\sigma^{(t)}}^2)$ comporte un petit nombre de valeurs simulées selon $\pi_0(\theta_{n+1}, \sigma_{n+1}^2)$ et principalement des valeurs (θ_i, σ_i^2) $(1 \leq i \leq n)$ elles aussi simulées selon π_0, mais avec des répliques. Des améliorations de cette méthode directe de simulation de processus de Dirichlet a priori sont proposées dans Escobar et West (1995), comme le calcul du nombre de composantes dans la loi des (θ_i, σ_i^2) $(1 \leq i \leq n)$. Cependant, le choix des hyperparamètres est relativement important pour de bonnes performances de l'estimateur résultant.

7

Choix et comparaison de modèles

> "Right this minute, wherever he is, Galad is puzzling over something he may never have faced before. Two things that are right, but opposite."
>
> **Robert Jordan**, *The Fires of Heaven.*

7.1 Motivations

Nous l'avons vu dans le Chapitre 5 : le *choix de modèle* peut être considéré comme un cas particulier de la théorie des tests. Les raisons pour lesquelles nous avons traité ce problème à part sont présentées ci-dessous. Ce chapitre devrait être accessible sans autre pré-requis sur le choix de modèle que l'idée simple que c'est un outil pour comparer des modèles et éventuellement en choisir un parmi ceux-ci.

Du point de vue conceptuel, la procédure inférentielle dépasse le cadre du Chapitre 5 : nous travaillons maintenant sur des modèles et non plus sur des paramètres. Ainsi, pour un problème donné, le choix entre un modèle exponentiel et un modèle de Weibull sera certainement plus lourd de conséquences que de décider si un paramètre θ vaut 1 ou 1.2, par exemple. En d'autres termes, l'incertitude sur la distribution d'échantillonnage $f(x)$ est ici très grande et dépasse largement le cadre des chapitres précédents, où elle portait seulement sur la valeur d'un paramètre inconnu (de dimension finie).

Du point de vue de la modélisation, le choix de modèle relève plus de l'estimation que des tests classiques. Par rapport au Chapitre 5, où nous avons

vu que tester l'hypothèse $H_0 : \theta \in \Theta_0$ est équivalent à estimer la fonction indicatrice $\mathbb{I}_{\Theta_0}$, le choix de modèle peut consister à choisir entre plusieurs possibilités, disons les modèles $\mathscr{M}_1, \ldots, \mathscr{M}_p$, et la décision sur "le" modèle revient à estimer l'indice $\mu \in \{1, \ldots, p\}$ associé à ce modèle (ou, plus exactement, à trouver la distribution a posteriori de cet indice). Naturellement, il existe de nombreux cas où il faut choisir de façon ferme et définitive le meilleur modèle (c'est-à-dire le modèle le plus approprié compte tenu des données), mais cela semble moins catégorique que de décider si l'hypothèse H_0 est vraie.

Du point de vue numérique, le choix de modèle met en jeu des structures plus complexes qui nécessitent presque systématiquement le recours à des techniques numériques avancées comme celles du Chapitre 6. D'où la séparation entre le Chapitre 5 et le présent chapitre, qui nous permet également de revenir au calcul des facteurs de Bayes et pseudo-facteurs de Bayes à l'aide de méthodes de Monte Carlo et MCMC (Section 7.3). En réalité, la comparaison de modèles implique l'emploi d'outils encore plus évolués que ceux du Chapitre 6. C'est pourquoi nous présenterons dans la Section 7.3.4 des méthodes de simulation permettant de manipuler des collections d'espaces de paramètres (aussi appelés *espaces de dimension variable*) et conçues spécialement pour le choix de modèle.

Enfin, comme nous le sous-entendions ci-dessus en parlant d'élargissement du cadre d'inférence, nous allons laisser un moment le domaine bien balisé des modèles paramétriques : à plusieurs reprises dans ce chapitre, nous nous retrouverons dans des cas pour lesquels la "vraie" distribution f est inconnue et où nous essayons de déterminer la distance entre f et une (ou plusieurs) familles de distributions $\{f_\theta; \theta \in \Theta\}$. Pour les tests de validité d'ajustement de la Section 7.6 par exemple, nous avons besoin d'un estimateur non paramétrique de f. Nous rencontrerons des problèmes analogues pour la sélection de variables (Section 7.5), où une solution est d'introduire un modèle *imbriquant*, différent du vrai modèle.

Il reste que beaucoup des idées exposées dans ce chapitre le sont aussi dans le Chapitre 5, étant donné que les techniques employées sont similaires, principalement les probabilités a posteriori et les facteurs de Bayes. De nombreux auteurs utilisent cet argument pour minimiser les différences entre les tests classiques et le choix de modèle. Voir par exemple Berger et Pericchi (2001), dont l'étude sur le choix de modèle comprend surtout des exemples de tests d'hypothèses nulles comme $H_0 : \theta = 0$.

Le choix de modèle, ainsi que les sujets connexes de sélection de variables et de tests de validité d'ajustement ont été l'objet d'une attention considérable ces dernières années, en partie grâce au développement de nouvelles méthodes numériques, et nous n'en présentons ici qu'une vision très partielle. Les lecteurs désireux d'approfondir le sujet pourront consulter par exemple le recueil d'articles édité par Racugno (1999).

7.1.1 Choix entre plusieurs modèles

Le choix de modèle semble s'affranchir du paradigme bayésien dans le sens où la distribution d'échantillonnage f elle-même n'est pas connue précisément. Cette incertitude rend difficile le conditionnement par rapport à l'observation x. Ce changement de paradigme apparaîtra encore plus nettement dans la Section 7.6 où nous chercherons à répondre à la question : *f appartient-elle à la famille $\{f_\theta;\, \theta \in \Theta\}$?*, l'hypothèse alternative étant complètement ouverte. Considérons d'abord le cadre plus restrictif dans lequel plusieurs modèles (paramétriques) sont en concurrence,

$$\mathscr{M}_i : x \sim f_i(x|\theta_i), \qquad \theta_i \in \Theta_i, \quad i \in I\,,$$

l'ensemble I des indices pouvant être éventuellement infini. Dans ce cas, le point de vue bayésien est plus facile à appliquer : on peut envisager de construire une distribution a priori pour chaque modèle $\mathscr{M}_i$ comme s'il s'agissait du seul vrai modèle considéré.

Le cadre minimal consiste à choisir parmi un nombre réduit de modèles. Ces modèles ont pu être sélectionnés pour des raisons très variées, des plus simples, comme l'historique de la discipline ou la commodité de calcul, aux plus compliquées et mieux justifiées.

Exemple 7.1. Dans l'Exemple 1.5, nous avons vu un jeu de données analysé par Lenk (1999) et étudiant la corrélation entre le taux de chômage et le nombre mensuel d'accidents dans le Michigan entre 1978 et 1987. En fait, avant de s'intéresser au lien entre les deux variables, on pourrait proposer deux modèles différents pour le nombre d'accidents N dans un mois :

$$\mathscr{M}_1 :\, N \sim \mathscr{P}oi(\lambda), \qquad \lambda > 0$$

et

$$\mathscr{M}_2 :\, N \sim \mathscr{N}eg(m,p), \qquad m \in \mathbb{N}^*\,, p \in [0,1]\,.$$

‖

Dans des cas plus compliqués, il y a trop peu d'information disponible pour éliminer un nombre substantiel de modèles et, par conséquent, l'ensemble de ceux qui restent à considérer est grand. Nous sommes alors plus proches d'une perspective non paramétrique.

Exemple 7.2. Un exemple cité dans la plupart des ouvrages portant sur l'estimation de mélanges est celui des *données galactiques*. D'abord abordé par Roeder (1992), il a ensuite été analysé par, entre autres, Chib (1995), Escobar et West (1995), Phillips et Smith (1996), Richardson et Green (1997) Roeder et Wasserman (1997) et Robert et Mengersen (1999). Il consiste en l'observation de quatre-vingt-deux vitesses de galaxies, représentées sur la Figure

7.1. Pour des raisons liées à l'Astrophysique, cet ensemble peut être modélisé par un mélange de distributions normales dont le nombre de composantes k est *inconnu*. (Une composante du mélange est associée à un *groupement de galaxies*.) Les modèles en concurrence sont donc

$$\mathscr{M}_i : n_j \sim \sum_{\ell=1}^{i} p_{\ell i}\, \mathscr{N}(\mu_{\ell i}, \sigma_{\ell i}^2)\,, \tag{7.1}$$

pour i allant de 1 à une borne supérieure arbitraire. ‖

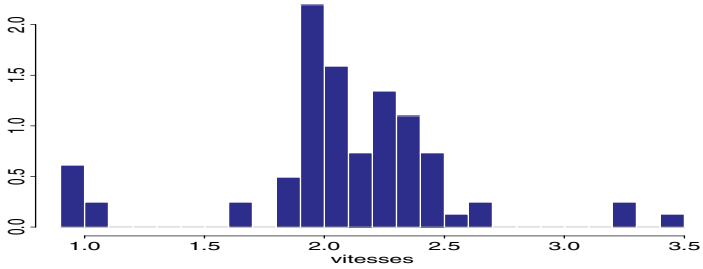

Fig. 7.1. Histogramme des données galactiques de Roeder (1992).

Dans d'autres contextes, comme celui de la sélection de *covariables* (ou *variables explicatives*) (Section 7.5), le nombre de modèles à considérer augmente de façon très importante avec l'inclusion de diverses combinaisons possibles de covariables.

Exemple 7.3. (Gelfand, 1996) Pour évaluer la vitesse de croissance de cinq orangers, on mesure leurs circonférences (y_{it} pour l'arbre i) à différents âges T_t. Les résultats sont présentés en Table 7.1. Les modèles étudiés sont ($i = 1, \cdots, 5, t = 1, \ldots, 7$)

$$\begin{aligned}
\mathscr{M}_1 &: y_{it} \sim \mathscr{N}(\beta_{10} + b_{1i}, \sigma_1^2) \\
\mathscr{M}_2 &: y_{it} \sim \mathscr{N}(\beta_{20} + \beta_{21} T_t + b_{2i}, \sigma_2^2) \\
\mathscr{M}_3 &: y_{it} \sim \mathscr{N}\left(\frac{\beta_{30}}{1 + \beta_{31}\exp(\beta_{32} T_t)}, \sigma_3^2\right) \\
\mathscr{M}_4 &: y_{it} \sim \mathscr{N}\left(\frac{\beta_{40} + b_{4i}}{1 + \beta_{41}\exp(\beta_{42} T_t)}, \sigma_4^2\right)\,,
\end{aligned}$$

où les b_{ji} sont des effets aléatoires, distribués selon une loi $\mathscr{N}(0, \tau^2)$. Ces modèles sont construits selon la graduation suivante : $\mathscr{M}_1$ est un modèle à effet individuel simple–sans effet temporel ; dans $\mathscr{M}_2$, l'effet temporel est linéaire ; la dépendance temporelle devient non linéaire dans $\mathscr{M}_3$ et on ajoute en sus des effets individuels pour obtenir le modèle $\mathscr{M}_4$. ‖

L'Exemple 7.3 montre bien qu'il y a souvent beaucoup d'arbitraire lors de la création de familles de modèles pour la sélection. De même, dans l'Exemple 7.2, l'hypothèse de normalité a été retenue pour son côté pratique et non pour des motivations concrètes issues de l'Astrophysique.

Tab. 7.1. Circonférences de cinq orangers (en millimètres) pour différents âges (en jours). (*Source :* Gelfand, 1996.)

	arbre				
jours	1	2	3	4	5
118	30	33	30	32	30
484	58	69	51	62	49
664	87	111	75	112	81
1004	115	156	108	167	125
1231	120	172	115	179	142
1372	142	203	139	209	174
1582	145	203	140	214	177

On perçoit bien dans les Exemples 7.1 à 7.3 une difficulté fondamentale liée au choix de modèle : alors qu'aucun modèle n'est rigoureusement exact, *plusieurs* modèles peuvent convenir dans une situation donnée. Se forcer à choisir un et un seul modèle reproduit donc le problème rencontré dans le Chapitre 5, où les procédures de test dont les valeurs sont restreintes à $\{0, 1\}$ semblaient inadaptées. En particulier, l'incertitude quant au modèle retenu n'est pas prise en compte. (Ce problème trouvera une solution radicale dans la Section 7.4 où on évite complètement le choix d'un modèle particulier.)

Dans l'Exemple 7.2 comme dans l'Exemple 7.3, certains modèles sont des sous-modèles d'autres modèles. Cela crée un problème *imbriqué* supplémentaire. Ainsi, dans l'Exemple 7.2, un mélange à k composantes est un sous-modèle d'un mélange à $(k+p)$ composantes, avec p composantes de poids nuls. Alors que, du point de vue de la modélisation, on a toujours intérêt à prendre le modèle le plus complet, la décision est moins évidente d'un point de vue statistique, puisque ce choix nécessitera d'estimer un plus grand nombre de paramètres à partir du même échantillon ! Un critère de choix de modèle doit donc non seulement mesurer l'ajustement[57] aux données, mais aussi prendre en compte les erreurs d'estimation.

[57] On parle de *surajustement* lorsqu'on bâtit un modèle qui s'accorde exceptionnellement bien aux données courantes mais dont les performances de prévision sont très médiocres. Cette opposition entre ajustement et erreur d'estimation étend l'opposition entre *biais* et *variance* rencontrée en Statistique classique et, en particulier, dans l'approche bayésienne empirique (Chapitre 10).

7.1.2 Champs d'application

Les exemples précédents le montrent bien, le choix de modèle n'est pas une procédure (d'estimation) monolithique, mais peut être employé pour de nombreuses raisons qui ne sont pas toujours évidentes pour (ou qui ne sont pas toujours énoncées explicitement par) l'expérimentateur (ou le "client"). Par conséquent, il semble impossible de se placer dans un cadre strict de Théorie de la Décision (ou tout du moins de préserver le même cadre pour toutes les utilisations envisagées). Parmi ces applications possibles, le choix de modèle peut être utile comme

(i) une première étape dans la *construction d'un modèle*, comme dans l'Exemple 7.1, lorsque l'intuition suggère quelques modèles et que l'expérimentateur veut déterminer lequel réalise le "meilleur" ajustement des données disponibles. Il ne s'agit là que d'un premier pas vers la Statistique non paramétrique, dans la mesure où il n'y a aucune raison de penser que l'un des modèles considérés est correct.

(ii) inversement, une dernière étape de la *vérification de modèles*, comme dans l'Exemple 7.3. Un modèle ou une famille de modèles ont été choisis pour diverses raisons théoriques ou pratiques et on cherche à savoir s'ils correspondent aux données. De même, dans le domaine des tests d'adéquation, le modèle n'est pas clairement défini en dehors de l'hypothèse nulle (comme nous l'expliquerons dans la Section 7.6).

(iii) une aide à l'*amélioration de modèles*, comme pour passer de $\mathscr{M}_1$ à $\mathscr{M}_2$ ou de $\mathscr{M}_3$ à $\mathscr{M}_4$ dans l'Exemple 7.3. Étant donné un modèle, éventuellement validé par un test d'adéquation, le but est d'étudier des modifications pour améliorer l'ajustement, ou, en d'autres termes, d'*imbriquer* le modèle existant dans une classe de modèles pour vérifier que le choix initial est suffisamment bon.

(iv) au contraire, un outil pour l'*élagage de modèles*[58], lorsque le modèle considéré est jugé trop compliqué pour être d'une quelconque utilité pratique, comme dans l'Exemple 7.2 avec $k = 50$, ou lorsque, en vertu du principe de parcimonie (Note 6.6.6), on souhaite examiner des sous-modèles plus simples pour voir s'ils s'ajustent assez bien aux données. C'est le cas en particulier dans le cadre de la sélection de variables, où on a à sa disposition un grand ensemble de covariables et on souhaite ne conserver que les plus importantes.

(v) plus simplement, une *comparaison entre modèles*, lorsqu'on hésite entre quelques modèles qui convenaient bien lorsqu'ils étaient utilisés sur d'autres échantillons et qu'on cherche un moyen de trouver celui qui a le meilleur ajustement sur l'échantillon courant, comme dans l'Exemple 7.1.

[58] Cette expression prend son sens littéral lorsqu'il s'agit d'élaguer de la plupart de ses branches un arbre de modèles possibles en sélection de variables.

(vi) de façon plus ambitieuse, une manière de faire du *test d'hypothèses*, suivant un protocole scientifique classique selon lequel on échafaude plusieurs hypothèses à l'aide de considérations théoriques et où on les vérifie par des expériences dédiées. (On pense notamment à la naissance de la théorie de la gravitation, puis au passage à la théorie de la gravitation d'Einstein opposée à celle de Newton, ou encore aux théories cosmologiques d'expansion ou de contraction de l'Univers, voir par exemple Feyerabend, 1975.)

(vii) dans un cadre plus limité, une façon de tester *l'efficacité de prévision*, comme, par exemple, dans le domaine de la finance. Contrairement à l'application (vi), les modèles en eux-mêmes n'intéressent pas l'expérimentateur qui se pose simplement la question de les évaluer en termes de leurs performances de prévision. Dans le cadre de l'Exemple 7.2, on pourrait ainsi chercher à évaluer la capacité pour chaque modèle d'allouer une nouvelle galaxie au groupe de galaxies le plus adéquat.

Les applications du choix de modèle sont manifestement aussi variées que celles de la Statistique puisqu'il existe bien peu de cas où un modèle ou une famille paramétrique donnés sont unanimement acceptés ! Citons tout de même quelques domaines où le choix de modèle s'est révélé particulièrement utile : en analyse d'images, lorsqu'on compare différentes structures de voisinage (Cressie, 1993) ; pour les modèles graphiques et systèmes experts lorsqu'on cherche à supprimer des liens entre variables (Cowell *et al.*, 1999) ; dans les modèles à dimensions variables, comme les modèles ARMA(p, q) avec p et q inconnus ; pour l'*inférence causale*, où il s'agit de décider si A a un effet sur B, connaissant un ensemble de variables $C_1, \ldots, C_p$ (Shafer, 1996, Robins et Wasserman, 2000).

7.2 Comparaison bayésienne de modèles

7.2.1 Modélisation spécifique de l'a priori

Comme pour d'autres problèmes, la réponse bayésienne standard consiste à placer une distribution a priori sur les éléments inconnus, ce qui, dans le cas présent, revient à proposer une modélisation a priori non plus seulement sur les paramètres, mais aussi sur les modèles eux-mêmes. L'espace des paramètres associé à l'ensemble des modèles (7.1) peut s'écrire

$$\boldsymbol{\Theta} = \bigcup_{i \in I} \{i\} \times \Theta_i \,, \tag{7.2}$$

l'indice de modèle $\mu \in I$ étant maintenant intégré à l'espace des paramètres. Par conséquent, il suffit de savoir attribuer des probabilités p_i aux différentes valeurs d'indice, c'est-à-dire en fait aux différents modèles $\mathscr{M}_i$ $(i \in I)$, puis de définir des lois a priori $\pi_i(\theta_i)$ sur les sous-espaces des paramètres Θ_i pour appliquer, comme d'habitude, le théorème de Bayes :

$$p(\mathscr{M}_i|x) = P(\mu = i|x) = \frac{p_i \int_{\Theta_i} f_i(x|\theta_i)\pi_i(\theta_i)d\theta_i}{\sum_j p_j \int_{\Theta_j} f_j(x|\theta_j)\pi_j(\theta_j)d\theta_j}\,. \tag{7.3}$$

Une première solution simple est d'utiliser la modélisation a priori pour obtenir un estimateur MAP (marginal) de μ, ce qui est équivalent à déterminer le modèle de plus grande probabilité a posteriori $p(\mathscr{M}_i|x)$. On peut également calculer directement une densité prédictive en y avec la moyenne :

$$\sum_j p_j \int_{\Theta_j} f_j(y|\theta_j)\pi_j(\theta_j|x)d\theta_j = \sum_j p(\mathscr{M}_j|x)\, m_j(y) \tag{7.4}$$

Néanmoins, il est souvent nécessaire de faire appel de façon plus marquée à la Théorie de la Décision.

Le formalisme bayésien usuel ou tout du moins la modélisation a priori se heurte ici à des difficultés nouvelles : la solution consistant à représenter la collection de modèles par (7.2) suppose la construction d'une distribution a priori (π_i, p_i) pour chaque $i \in I$, ce qui est délicat lorsque I est infini. De plus, toutes les lois a priori π_i doivent être des lois propres puisqu'il n'y a pas unicité des facteurs d'échelles pour les lois a priori impropres, comme nous l'avons vu dans le Chapitre 5. En outre, si certains modèles sont imbriqués dans d'autres, c'est-à-dire si

$$\mathscr{M}_{i_0} \subset \mathscr{M}_{i_1}\,,$$

le choix de π_{i_0} devrait être lié à celui de π_{i_1} et peut-être aussi celui de p_{i_0} à celui de p_{i_1}. Par exemple, si $\mathscr{M}_1 = \mathscr{M}_2 \cup \mathscr{M}_3$, il n'est pas absurde d'exiger que

$$p(\mathscr{M}_1) = p(\mathscr{M}_2) + p(\mathscr{M}_3)\,,$$

ou au moins que $p(\mathscr{M}_1) \geq p(\mathscr{M}_2) + p(\mathscr{M}_3)$. De façon analogue, si deux modèles $\mathscr{M}_{i_0}$ et $\mathscr{M}_{i_1}$ ne sont pas imbriqués l'un dans l'autre, la modélisation a priori devrait pouvoir s'adapter à un troisième modèle $\mathscr{M}_{i_2}$ imbriquant $\mathscr{M}_{i_0}$ et $\mathscr{M}_{i_1}$. (En Économétrie, on appelle *imbrication* (traduction libre d'*encompassing*) cette technique de création d'un supermodèle.)

Formulons une dernière remarque importante et spécifique au problème du choix de modèle : *les paramètres communs à plusieurs modèles doivent être considérés comme des entités différentes*. Ce problème est souvent négligé dans la littérature, y compris dans Jeffreys (1961), parce que les paramètres communs peuvent être formellement intégrés en utilisant la *même* loi de distribution a priori, même (surtout !) quand celle-ci est impropre. Une autre façon, moins extrême, de contourner le principe ci-dessus est de suggérer, comme dans Berger et Pericchi (1998), que l'utilisation du *même* a priori impropre pour les paramètres communs permet de régler le problème de la constante de normalisation (Exercice 7.4), mais nous ne saurions recommander de façon systématique cette solution spécifique.

Exemple 7.4. (Suite de l'Exemple 7.3) Regardons de plus près les modèles $\mathscr{M}_1$ et $\mathscr{M}_2$: bien que β_{10} et β_{20} aient en commun le fait d'être des intercepts, comme σ_1^2 et σ_2^2 celui d'être des variances, ils sont bel et bien des quantités différentes, à cause de la présence du terme $\beta_{21}T_t$ dans le modèle $\mathscr{M}_2$. En particulier, dans le cas où $\mathscr{M}_2$ est le vrai modèle, β_{10} correspond à β_{20} décalé de la moyenne des $\beta_{21}T_t$ et σ_1^2 est plus grand que σ_2^2 à cause d'une adéquation moins fidèle (voir l'Exercice 7.5). ‖

Le problème d'inférence n'est pas plus facile à formaliser dans le cadre de la Théorie de la Décision, à cause de toutes les applications potentielles du choix de modèle, décrites dans la Section 7.1.2, et qui ne sont pas nécessairement compatibles entre elles. Le choix de modèle est en général une partie d'un *processus de décision* global : le modèle est d'abord construit, puis amélioré par extension ou réduction (comme nous l'avons expliqué dans les points (iii) et (iv) ci-dessus). Ce n'est qu'ensuite qu'on décide de sélectionner ce modèle comme le "vrai" modèle en vue d'applications futures. Trouver une fonction de coût tenant compte de toutes ces étapes est clairement impossible, mais c'est envisageable si on s'intéresse plus spécifiquement à l'étape de sélection. Par exemple, les *moyennes de modèles* comme celle décrite en (7.4) ne sont pas acceptables de ce point de vue parce que la procédure d'estimation, en incluant tous les modèles compatibles avec les données, pèche par excès d'indécision ! Si on ne dispose d'aucune (ou de trop peu) d'information sur les conséquences d'un mauvais choix de modèle et qu'on est par conséquent incapable de construire une fonction de coût, $\mathrm{L}(\mu, d)$ ou $\mathrm{L}((\mu, \theta_\mu), (d, \vartheta))$, d'aide à la décision, une solution, défendue à la fin de la Section 7.1.1, est de prévenir le surapprentissage en introduisant dans la fonction de coût des *termes de pénalisation* portant sur le nombre de paramètres du modèle (c'est-à-dire sa taille). Ce point est détaillé dans la Section 7.2.3. Voir aussi Carota *et al.* (1996) pour une façon de juger les modèles à l'aide de la Théorie de la Décision, relevant plus du point (iii) ci-dessus et faisant usage des divergences de Kullback-Leibler comme dans la Section 7.5.

Une autre difficulté réside dans le calcul de densités prédictives et marginales et d'autres quantités à évaluer dans le cadre du choix de modèle. Il ne s'agit bien sûr pas d'un problème spécifique au choix de modèle (voir le Chapitre 6), mais un certain nombre de particularités plaident pour la recherche de solutions sur mesure :

(i) Les espaces de paramètres sont souvent de dimension infinie, comme dans (7.1), ce qui oblige à faire appel à des notions plus compliquées de théorie de la mesure.

(ii) Le fait de devoir intégrer sur *plusieurs* espaces de paramètres pour évaluer des quantités a posteriori ou prédictives augmente d'autant le temps de calcul nécessaire, sans possibilité, en général, d'exporter les résultats des calculs d'un sous-espace à un autre.

(iii) L'implémentation d'algorithmes MCMC (Chapitre 6) dans un espace de paramètres vu comme somme directe de différents sous-espaces nécessite des techniques markoviennes plus élaborées.

(iv) Dans certains contextes, comme celui de la sélection de variables, la collection de modèles est finie mais exponentiellement grande et elle ne peut donc pas être explorée intégralement.

Dans tous les cas, sauf peut-être pour les modèles les plus simples, on a donc recours à des techniques numériques (approximatives) avancées, car il est impossible d'obtenir une représentation analytique et exacte de la loi a posteriori. Nous détaillons ces techniques en Section 7.3.

7.2.2 Facteurs de Bayes

Une fois définies la modélisation (7.1) et les distributions a priori correspondantes, la procédure inférentielle est assimilable à un problème de test générique. La solution proposée par Kass et Raftery (1995) et soutenue également par Berger et Pericchi (2001) est de faire appel aux facteurs de Bayes. Par exemple, dans le cas de la comparaison des modèles $\mathscr{M}_1$ et $\mathscr{M}_2$:

$$B_{12} = \frac{P(\mathscr{M}_1|x)}{P(\mathscr{M}_2|x)} \Big/ \frac{P(\mathscr{M}_1)}{P(\mathscr{M}_2)} = \frac{\displaystyle\int_{\Theta_1} f_1(x|\theta_1)\pi_1(\theta_1)d\theta_1}{\displaystyle\int_{\Theta_2} f_2(x|\theta_2)\pi_2(\theta_2)d\theta_2}\,.$$

Le cadre est donc analogue à celui de la Section 5.2 et, par conséquent, les difficultés sont également similaires, bien qu'accrues par un plus grand nombre de modèles à considérer (peut-être même une infinité !) et par la nécessité d'utiliser beaucoup plus fréquemment des lois a priori non informatives. Remarquons ici qu'on peut comparer les modèles sur la base des facteurs de Bayes, couple $(\mathscr{M}_i, \mathscr{M}_j)$ par couple, grâce à la cohérence des facteurs de Bayes, qui vérifient $B^\pi_{ij} = B^\pi_{ik} B^\pi_{kj}$, ce qui assure la transitivité de l'ordonnancement de modèles. (Mais rappelons que cette propriété n'est pas vérifiée par les pseudo-facteurs de Bayes définis dans la Section 5.2.6.)

Pour exactement les mêmes raisons que dans la Section 5.2.5, les *lois a priori impropres* sont à proscrire (à moins qu'elles ne portent sur des paramètres communs à tous les modèles, comme nous l'avons décrit précédemment). En outre, les lois a priori vagues, c'est-à-dire les lois a priori propres ayant une très grande variance–utilisées notamment dans `BUGS`, voir Note 6.6.2–ne résolvent pas le problème, comme le montre le *paradoxe de Jeffreys-Lindley* (Section 5.2.5).

Exemple 7.5. (Suite de l'Exemple 7.1) Soient les distributions a priori

$$\pi_1(\lambda) = \mathscr{G}a(\alpha, \beta)\,, \qquad \pi_2(m, p) = \frac{1}{M}\, \mathbb{I}_{\{1,\cdots,M\}}(m)\mathbb{I}_{[0,1]}(p)\,.$$

La seconde loi a priori est uniforme sur l'espace des paramètres Θ_2. Le facteur de Bayes est alors :

$$B_{12}^{\pi} = \frac{\dfrac{\beta^{\alpha}}{\Gamma(\alpha)} \displaystyle\int_0^{\infty} \frac{\lambda^{\alpha+x-1}}{x!} e^{-\lambda\beta} d\lambda}{\dfrac{1}{M} \displaystyle\sum_{m=1}^{M} \int_0^1 \binom{m}{x-1} p^x (1-p)^{m-x} dp}$$

$$= \frac{\Gamma(\alpha+x)}{x!\,\Gamma(\alpha)} \beta^{-x} \Big/ \frac{1}{M} \sum_{m=1}^{M} \frac{x}{(m-x+1)(m+1)}$$

$$= M(m+1) \frac{(x+\alpha-1)\cdots\alpha}{x(x-1)\cdots 1} \beta^{-x} \Big/ \sum_{m=1}^{M} \frac{x}{m-x+1}$$

Les choix de α et β ont une grande influence sur la valeur de B_{12}^{π}, en particulier lorsque tous deux tendent vers 0 (Exercice 7.10). ‖

Cette difficulté fondamentale avec les lois a priori impropres peut se résoudre par des solutions pseudo-bayésiennes, en ayant recours à un échantillon minimal d'apprentissage ou à des observations virtuelles, comme dans la Section 5.2.6. (C'est d'ailleurs dans le contexte du choix de modèle linéaire ou log-linéaire que Spiegelhalter et Smith, 1980, furent parmi les premiers à suggérer l'utilisation de pseudo-facteurs bayésiens.) L'évaluation de modèles sous des lois a priori impropres peut ensuite être conduite avec des facteurs de Bayes intrinsèque ou fractionnel (avec les mêmes réserves que dans la Section 5.2.6).

Exemple 7.6. (Suite de l'Exemple 7.2) Comme nous l'avons dit dans l'Exemple 5.19, il n'existe pas d'échantillon minimal d'apprentissage pour les modèles de mélange, quel que soit le nombre d'observations. Par conséquent, les facteurs de Bayes intrinsèques et fractionnels ne sont pas applicables ici.

Une première solution, suggérée dans Diebolt et Robert (1994) à des fins de simulation et validée ensuite par Wasserman (1999), est d'imposer que l'échantillon $(x_1, \ldots, x_n)$ contienne suffisamment d'observations (au sens des échantillons d'apprentissage) issues de chaque composante (voir également Richardson et Green, 1997). Bien que raisonnable lorsque toutes les composantes sont clairement identifiées, cette méthode a le désavantage de créer une dépendance entre les observations (qui restent tout de même échangeables) et le calcul des pseudo-facteurs de Bayes devient dans ce cas très lourd.

Une alternative, adoptée dans Mengersen et Robert (1996) pour tester $k = 1$ contre $k = 2$, est d'affecter une distribution a priori non informative $\pi(\mu, \tau)$ au paramètre global de position-échelle du modèle (ou de l'échantillon) et d'exprimer les paramètres de chaque composante en tant que perturbations de ce paramètre de position-échelle, avec des lois a priori propres. Étant donné

que (μ, τ) est commun à toutes les composantes, le problème de normalisation lié à l'a priori impropre est moins pénalisant[59]. ‖

7.2.3 Le critère de Schwarz

Avant d'aborder les questions liées aux termes de pénalisation et aux solutions bayésiennes approchées (grossièrement), nous devons présenter brièvement quelques notions d'approximations asymptotiques des facteurs de Bayes[60].

Pour les modèles réguliers, lorsque $\mathscr{M}_1 \subset \mathscr{M}_2$, le rapport de vraisemblance entre $\mathscr{M}_2$ et $\mathscr{M}_1$ est approximativement distribué selon une loi du $\chi^2_{p_2-p_1}$,

$$-2\log\lambda_n \;\approx\; \chi^2_{p_2-p_1}$$

en supposant que $\mathscr{M}_1$ est le vrai modèle (Gouriéroux et Monfort, 1996, et Lehmann et Casella, 1998). On a

$$\begin{aligned} P(\mathscr{M}_2 \text{ choisi}\,|\mathscr{M}_1) &= P(\lambda_n < c|\mathscr{M}_1) \\ &\simeq P(\chi^2_{p_2-p_1} > -2\log(c)) > 0\,. \end{aligned}$$

Donc, d'un point de vue fréquentiste, un critère dépendant seulement du rapport de vraisemblance ne converge pas vers une réponse certaine sous $\mathscr{M}_1$ (mais il converge sous $\mathscr{M}_2$). C'est la raison pour laquelle on ajoute des facteurs de pénalisation au rapport de vraisemblance pour compenser ce biais, comme dans le cas du critère d'Akaike (1983),

$$-2\log\lambda_n - \alpha(p_2 - p_1)\,. \tag{7.5}$$

Pour $\alpha = \log 2$, on retrouve l'approximation obtenue par une procédure d'Aitkin (1991) dans laquelle l'auteur utilise les données deux fois, une première fois pour construire un (pseudo-) a priori propre en utilisant la distribution a posteriori, puis une seconde fois pour calculer le facteur de Bayes comme si la distribution a priori était exacte (Exercice 5.16).

Le *développement de Laplace*, explicité dans la Section 6.2.3, donne une approximation d'intégrale,

$$\int_\Theta \exp\{n\,h(\theta)\}d\theta = \exp\{n\,h(\hat\theta)\}(2\pi)^{p/2}n^{-p/2}|H^{-1}(\hat\theta)| + O(n^{-1})\,,$$

[59] Évidemment, cet appel à un paramètre commun à tous les modèles contredit notre recommandation ci-dessus sur les paramètres différents dans chaque modèle.

[60] Cette section a pour but d'illustrer le lien entre approximation bayésienne et critères de pénalisation usuels, pas de présenter ces critères. Elle peut donc être laissée de côté en première lecture, surtout si les lecteurs ne sont pas familiers avec ces critères.

où p est la dimension de Θ, $\hat{\theta}$ le point où h atteint son maximum et H la matrice hessienne de h. En développant à la fois le numérateur et le dénominateur du facteur de Bayes grâce à cette approximation, on obtient :

$$B_{12}^{\pi} \simeq \frac{L_{1,n}(\hat{\theta}_{1,n})}{L_{2,n}(\hat{\theta}_{2,n})} \left| \frac{H_1^{-1}(\hat{\theta}_{1,n})}{H_2^{-1}(\hat{\theta}_{2,n})} \right|^{1/2} \left(\frac{n}{2\pi}\right)^{(p_2-p_1)/2},$$

avec p_1 et p_2 dimensions de Θ_1 et Θ_2, $L_{1,n}$ et $L_{2,n}$ fonctions de vraisemblance calculées sur n observations, et $\hat{\theta}_{1,n}$ et $\hat{\theta}_{2,n}$ maximums respectifs de L_1 et L_2. D'où :

$$\log(B_{12}^{\pi}) \simeq \log \lambda_n + \frac{p_2 - p_1}{2} \log(n) + K(\hat{\theta}_{1,n}, \hat{\theta}_{2,n}), \tag{7.6}$$

en notant λ_n le rapport de vraisemblance usuel pour la comparaison de $\mathscr{M}_1$ et $\mathscr{M}_2$,

$$\lambda_n = L_{1,n}(\hat{\theta}_{1,n})/L_{2,n}(\hat{\theta}_{2,n}),$$

et $K(\hat{\theta}_{1,n}, \hat{\theta}_{2,n})$ le terme restant.

Cette approximation est à l'origine du *critère de Schwarz* (Schwarz, 1978) : pour $\mathscr{M}_1 \subset \mathscr{M}_2$, le facteur de Bayes est approché par

$$S = -\log \lambda_n - \frac{p_2 - p_1}{2} \log(n)$$

si le terme de reste $K(\hat{\theta}_{1,n}, \hat{\theta}_{2,n})$ dans (7.6) est négligeable devant les deux autres, c'est-à-dire est en $O(1)$. (Voir Gelfand et Dey, 1994, Section 8, pour un exemple où ce terme n'est pas négligeable.)

Le critère de Schwarz, également appelé *BIC* (pour *Bayes Information Criterion*), est donc une première approximation à l'ordre 1 du facteur de Bayes, comme le décrivent Kass et Raftery (1995). Néanmoins, la pertinence de ce critère dans un contexte bayésien est contestable pour deux raisons : (i) l'influence de l'hypothèse a priori disparaît ; (ii) cette approximation n'est acceptable que pour les modèles réguliers. Ainsi, dans l'Exemple 7.2, le comportement asymptotique (du logarithme) du rapport de vraisemblance $-2\log\lambda_n$ est beaucoup plus complexe que celui de l'approximation $\chi^2_{p_2-p_1}$ (voir, par exemple, Dacunha-Castelle et Gassiat, 1999) et le critère de Schwarz est inefficace. Berger et Pericchi (2001) recensent d'autres exemples de vraisemblances irrégulières. En outre, dans des situations non iid, les définitions de n et p peuvent être ambiguës, comme le soulignent Spiegelhalter *et al.* (1998). Du point de vue de la complexité de calcul, remarquons que pour déterminer le critère de Schwarz, il faut disposer des estimateurs du maximum de vraisemblance pour *tous* les modèles.

Exemple 7.7. (Suite de l'Exemple 7.2) On décompose le critère de Schwarz en

$$\begin{aligned} S &= \log\left\{L_{2,n}(\hat{\theta}_{2,n})/L_{1,n}(\hat{\theta}_{1,n})\right\} - \frac{p_2 - p_1}{2}\log(n) \\ &= \log L_{2,n}(\hat{\theta}_{2,n}) - \frac{p_2}{2}\log(n) - \log L_{1,n}(\hat{\theta}_{1,n}) + \frac{p_1}{2}\log(n)\,. \end{aligned}$$

La partie relative au modèle $\mathscr{M}_i$ est donc

$$S_i = \log L_{i,n}(\hat{\theta}_{i,n}) - \frac{p_i}{2}\log(n)\,.$$

Si $\mathscr{M}_k$ est associé à la composante k du modèle, $p_k = 3k-1$. Pour les données de vitesses de galaxies, Raftery (1996) obtient

$$S_1 = -271.8\,, \quad S_2 = -249.7\,, \quad S_3 = -256.7\,, \quad S_4 = -263.6\,,$$

en utilisant l'algorithme EM (voir Note 6.6.6) pour obtenir des approximations des estimateurs du maximum de vraisemblance $\hat{\theta}_{i,n}$ pour $k > 1$. On en déduit que, selon le critère de Schwarz, il faut préférer le modèle à deux composantes aux autres. (Mais insistons de nouveau sur l'absence de validité asymptotique de l'approximation par une loi du χ^2 de la distribution du rapport de vraisemblance dans ce cas.) ‖

7.2.4 Déviance bayésienne

Spiegelhalter *et al.* (1998) et Spiegelhalter *et al.* (2002) proposent une alternative bayésienne aux critères AIC (critère d'information d'Akaike) et BIC, utilisant la *déviance* et donc appelé DIC (pour *Deviance Information Criterion*). Ce critère est plus satisfaisant que les précédents parce qu'il prend en compte l'information a priori et intègre un facteur de pénalisation naturel à la log-vraisemblance. De plus, il permet d'utiliser des lois a priori impropres, puisque chaque modèle est considéré séparément. En revanche, il ne rentre pas naturellement dans un schéma décisionnel bayésien et certains, comme Dawid (2002), critiquent sa pertinence dans une perspective bayésienne. Sans vouloir entamer une discussion de cet ordre, on peut effectivement remarquer que la définition même du critère DIC est entachée d'imprécision et que sa généralisation en dehors des familles exponentielles et des modèles linéaires généralisés n'est pas naturelle (voir Celeux *et al.*, 2005).

Comme nous l'avons souligné en Section 7.2.3, étant donné un modèle $f(x|\theta)$ avec une distribution a priori associée $\pi(\theta)$, la déviance[61] $D(\theta) = -2\log(f(x|\theta))$ n'est pas une bonne mesure discriminante, puisqu'elle est biaisée en faveur des modèles à plus grande dimension. Bien sûr, cela reste vrai pour sa distribution a posteriori. Spiegelhalter *et al.* (2002) introduisent une déviance pénalisée,

[61] Dans les modèles linéaires généralisés (McCullagh et Nelder, 1989), la *déviance* est en général ajustée avec un terme supplémentaire en y comme $f(y|\hat{\theta}(y))$ avec $\hat{\theta}(y)$ un estimateur arbitraire de θ. Lorsque ce terme ne dépend pas du modèle ou est choisi une fois pour toutes pour un modèle particulier comme le modèle complet ou imbriquant, il n'y a évidemment aucune différence entre le choix de modèle fondé sur $D(\theta)$ et celui fondé sur $D(\theta) + 2\log f(y|\hat{\theta}(y))$.

$$\begin{aligned}\text{DIC} &= \mathbb{E}[D(\theta)|x] + p_D \\ &= \mathbb{E}[D(\theta)|x] + \{\mathbb{E}[D(\theta)|x] - D(\mathbb{E}[\theta|x])\}\ ,\end{aligned} \tag{7.7}$$

associée à une pseudo-dimension p_D. L'évaluation de modèles selon ce critère suit alors le principe que *plus le critère DIC est faible, meilleur est le modèle.*

Le facteur $\mathbb{E}[D(\theta)|x]$ dans (7.7) peut être vu comme une mesure d'*ajustement* aux données, alors que p_D est un terme évaluant la *complexité*, appelé *nombre effectif de paramètres*. L'analogie avec le critère d'information d'Akaike (7.5) découle naturellement de $\text{DIC} = D(\mathbb{E}[\theta|x]) + 2p_D$. (Spiegelhalter *et al.*, 2002) montrent que, dans un contexte non hiérarchique où la distribution a posteriori de θ est approximativement normale, DIC et AIC sont en fait équivalents. Remarquons également que DIC suit la décomposition classique de l'erreur quadratique en carré du biais et variance,

$$\mathbb{E}_\theta[(\delta-\theta)^2] = (\mathbb{E}_\theta[\delta]-\theta)^2 + \mathbb{E}_\theta[(\delta - \mathbb{E}_\theta[\delta])^2]\,,$$

mais dans un cadre non paramétrique (à l'exception de $\mathbb{E}[\theta|x]$, qui dépend de la paramétrisation).

Exemple 7.8. (Spiegelhalter *et al.*, 1998) Pour une analyse de variance simple ($i = 1, \ldots, p$)

$$y_i = \theta_i + \sigma_i \epsilon_i\,, \qquad \epsilon_i \sim \mathcal{N}(0,1)\,,$$

la divergence s'écrit $D(\theta) = \sum_i \sigma_i^{-1}(\theta_i - y_i)^2$. Par conséquent, si $\theta_i = \theta$ ($i = 1, \ldots, p$) et $\pi(\theta) = 1$,

$$\mathbb{E}[D(\theta)|y_1, \ldots, y_p] = \sum_{i=1}^{k} \sigma_i^{-1}(y_i - \mathbb{E}[\theta|y_1, \ldots, y_p])^2 + 1 \tag{7.8}$$

avec $\mathbb{E}[\theta|y_1, \ldots, y_p] = \sum_i \sigma_i^{-1} y_i / \sum_i \sigma_i^{-1}$. Dans ce cas, on a $p_D = 1$.

En revanche, si on considère le modèle $\theta_i \sim \mathcal{N}(\mu, \tau^2)$ en supposant les hyperparamètres μ et τ connus, il vient

$$\mathbb{E}[D(\theta)|y_1, \ldots, y_p] = \sum_{i=1}^{k} \sigma_i^{-1}(1-\varrho_i)^2(y_i-\mu)^2 + \sum_{i=1}^{k} \varrho_i\,, \tag{7.9}$$

avec $\varrho_i = \sigma_i^2\tau^2/(\sigma_i^2+\tau^2)$. ‖

Le calcul pratique de la déviance bayésienne nécessite le plus souvent le recours à des algorithmes MCMC, les cas comme celui de l'Exemple 7.8 ou ceux présentés dans Spiegelhalter *et al.* (2002) étant particulièrement rares. L'implémentation de ces algorithmes est toutefois relativement aisée, une fois programmée la simulation d'un échantillon MCMC $(\theta^{(1)}, \ldots, \theta^{(T)})$, puisque $\mathbb{E}[D(\theta)|y_1, \ldots, y_p]$ est une simple espérance a posteriori d'une fonction explicite de θ.

Exemple 7.9. (Spiegelhalter *et al.*, 1998) Une étude sur le cancer de la lèvre dans cinquante-six régions d'Écosse met en relation le nombre de cas recensés y_i et les nombres attendus au niveau national E_i de la façon suivante :

$$y_i \sim \mathscr{P}(\lambda_i E_i)\,,$$

$\lambda_i = \exp(\theta_i)$ étant le risque de cancer des lèvres spécifique à la zone. Des covariables possibles sont x_i, le pourcentage de la population travaillant en extérieur, et la localisation géographique de la région, représentée par une liste $\mathscr{A}_i$ de régions adjacentes. On peut envisager les modèles suivants :

$$\begin{aligned}\mathscr{M}_1 &: \theta_i = \alpha + \beta x_i\,,\\ \mathscr{M}_2 &: \theta_i = \varphi_i\,,\\ \mathscr{M}_3 &: \theta_i = \varphi_i + \beta x_i\,,\end{aligned}$$

les φ_i étant spatialement corrélés, c'est-à-dire

$$\varphi_i|\varphi_{j,\,j\neq i} \sim \mathscr{N}\left(\sum_{j\in\mathscr{A}_i} \varphi_j/n_i, \tau^2/n_i\right)\,,$$

avec n_i nombre de régions adjacentes. (Ce modèle spatial, appelé *modèle spatial autorégressif*, est souvent utilisé en Statistique spatiale. Voir Besag, 1974, ou Cressie, 1993, et l'Exercice 7.18.)

Avec des lois a priori non informatives pour les hyperparamètres (sauf pour τ^2 qui suit une loi $\mathscr{IG}(1,1)$), l'algorithme MCMC donne des valeurs approchées de DIC de 242.8, 88.5 et 89.0 pour les trois modèles, avec des nombres de paramètres p_D correspondants de 2.1, 31.6 et 29.4, respectivement. Les modèles $\mathscr{M}_2$ et $\mathscr{M}_3$ ont donc des performances équivalentes, nettement meilleures que celles du modèle $\mathscr{M}_1$. Soulignons toutefois que, alors que le nombre réel de paramètres dans le modèle $\mathscr{M}_1$ est de 2, il est respectivement de 57 et 58 pour $\mathscr{M}_2$ et $\mathscr{M}_3$. ‖

Spiegelhalter *et al.* (2002) suggèrent d'autres applications de la déviance bayésienne comme par exemple le calcul des *résidus de déviance*. Ils mettent également en garde contre l'absence d'invariance par reparamétrisation de $D(\mathbb{E}[\theta|x])$ et conseillent d'utiliser la paramétrisation canonique pour les modèles linéaires généralisés[62].

7.3 Aspects numériques

Comme dans d'autres contextes, l'approche bayésienne du choix de modèle se heurte souvent à la difficulté numérique d'évaluer des intégrales du type

[62] Une solution est de remplacer $\mathbb{E}[\theta|x]$ par son estimateur MAP. On obtient alors un critère avec une vraie invariance dans la paramétrisation, avec la contrepartie que cet estimateur est plus difficile à évaluer que la moyenne a posteriori.

$$m_i(x) = \int f_i(x|\theta_i)\pi_i(\theta_i)d\theta_i \tag{7.10}$$

et, corrélativement, des rapports d'intégrales

$$\int f_1(x|\theta_1)\pi_1(\theta_1)d\theta_1 \Big/ \int f_2(x|\theta_2)\pi_2(\theta_2)d\theta_2\,,$$

sans compter toutes les complications supplémentaires liées à la dérivation des facteurs de Bayes intrinsèques et fractionnels. On peut bien sûr faire appel aux techniques présentées dans le Chapitre 6, qui sont principalement des approximations asymptotiques et des méthodes de simulation de Monte Carlo ou par MCMC. D'autres idées, plus spécifiques, ont cependant été développées pour le calcul des facteurs de Bayes et de quantités associées, comme le détaillent Chen *et al.* (2000).

7.3.1 Échantillonnage d'importance pour facteurs de Bayes

Cette technique, introduite dans la Section 6.2.2, convient particulièrement au calcul de distributions prédictives comme (7.10). Étant donné une distribution d'*importance*, de densité proportionnelle à g, et un échantillon $\theta^{(1)}, \ldots, \theta^{(T)}$, on obtient une approximation de la densité marginale du modèle $\mathscr{M}_i$, $m_i(x)$, en écrivant :

$$m_i^{IS}(x) = \sum_{t=1}^{T} f_i(x|\theta^{(t)})\frac{\pi_i(\theta^{(t)})}{g(\theta^{(t)})} \Big/ \sum_{t=1}^{T} \frac{\pi_i(\theta^{(t)})}{g(\theta^{(t)})}\,,$$

le dénominateur prenant la place de la constante de normalisation manquante. (On remarque que, si g est une densité de probabilité, l'espérance de $\pi(\theta^{(t)})/g(\theta^{(t)})$ est égale à 1.)

Une bonne raison, parmi d'autres, d'employer l'échantillonnage d'importance dans le cadre du choix de modèle est qu'on peut réutiliser l'échantillon $(\theta^{(1)}, \ldots, \theta^{(T)})$ pour plusieurs modèles $\mathscr{M}_i$ du moment qu'ils mettent en jeu les mêmes (types de) paramètres. (Alors que ce n'est en revanche pas possible dans les Exemples 7.1 et 7.2 puisque les différents modèles correspondent à des espaces de dimensions différentes.) On pourra consulter Chen et Shao (1997) pour un exemple utilisant des facteurs de Bayes.

La variance de $m^{IS}(x)$ peut cependant être infinie, comme cela a été évoqué en Section 6.2.2. Raftery (1996) s'intéresse au problème du choix de la fonction d'importance dans ce contexte des lois marginales, pour un modèle donné de densité d'échantillonnage $f(x|\theta)$ et de distribution a priori $\pi(\theta)$. L'idée la plus immédiate est de prendre $g(\theta) = \pi(\theta)$. On obtient alors l'estimateur suivant de la densité marginale :

$$m^{IS}(x) = \frac{1}{T}\sum_t f(x|\theta^{(t)})\,.$$

Ce choix est malheureusement mauvais lorsque les données sont informatives, car la plupart des valeurs simulées $\theta^{(t)}$ tombent en dehors de la région modale de la vraisemblance et de la loi a posteriori. (Dans le cas limite où π est impropre, cette option est évidemment impossible.) Naturellement, dans la mesure où les queues de la distribution π sont en général plus larges que celles de $\pi(\theta|x)$, les problèmes liés à une variance infinie sont rares avec une telle fonction d'importance.

Un autre choix possible est $g(\theta) = f(x|\theta)\pi(\theta)$, c'est-à-dire de simuler suivant la loi a posteriori sans connaître la constante de normalisation. L'estimateur associé est alors

$$m^{IS}(x) = 1 \Big/ \frac{1}{T} \sum_{t=1}^{T} \frac{1}{f(x|\theta^{(t)})}\,, \tag{7.11}$$

qui est, en fait, la *moyenne harmonique* des vraisemblances. Par conséquent, $m^{IS}(x)$ est une approximation de la constante de normalisation de g. Bien que cette solution soit compatible avec des lois a priori impropres, tant que les distributions a posteriori sont définies, la variance correspondante est souvent infinie. Une technique pour régler ce problème est appelée l'*échantillonnage d'importance défensif*. Elle consiste à choisir un mélange de g (ou plutôt de $\pi(\theta|x)$) et d'une distribution à queues lourdes, $\varpi(\theta)$:

$$(1-\varrho)\pi(\theta|x) + \varrho\varpi(\theta)\,, \qquad \varrho > 0\,.$$

avec ϱ petit. Le rôle du second terme n'est pas de fournir une approximation intéressante de la loi a posteriori mais simplement de stabiliser l'estimateur pour assurer une variance finie. (Voir Hesterberg, 1998, et Owen et Zhou, 2000 pour plus de détails sur cette méthode.) Newton et Raftery (1994) proposent par exemple $\varpi(\theta) = \pi(\theta)$.

Une solution proche de la précédente, suggérée par Gelfand et Dey (1994), est de générer un échantillon de $\theta^{(t)}$ suivant la loi a posteriori et d'utiliser

$$m^{IS}(x) = 1 \Big/ \frac{1}{T} \sum_{t=1}^{T} \frac{h(\theta^{(t)})}{f(x|\theta^{(t)})\pi(\theta^{(t)})}\,, \tag{7.12}$$

plutôt que (7.11), où h est une densité *quelconque* (Exercice 7.19). L'estimateur (7.12) a de plus une variance finie si

$$\int \frac{h^2(\theta)}{f(x|\theta)\pi(\theta)} d\theta < \infty\,.$$

h étant un paramètre (fonctionnel) libre, on peut (en principe) le choisir tel que cette condition soit satisfaite. Bien évidemment, le choix pratique de h n'est pas si aisé, surtout en grande dimension.

7.3.2 Échantillonnage par passerelle

Les méthodes de Monte Carlo dédiées à l'estimation de rapports de constantes de normalisation, ou, de façon équivalente, de facteurs de Bayes, se sont multipliées depuis 1995. Les lecteurs intéressés pourront trouver une présentation complète de ces méthodes dans le livre de Chen *et al.* (2000). Nous nous contentons ici de présenter une solution liée à l'échantillonnage d'importance.

L'*échantillonnage par passerelle* (traduction libre de *bridge sampling*) a été proposé par Meng et Wong (1996) à partir de principes déjà utilisés en Physique des particules : si deux modèles partagent le même espace des paramètres Θ, si $\pi_1(\theta|x) = c_1\tilde{\pi}_1(\theta|x)$ et $\pi_2(\theta|x) = c_2\tilde{\pi}_2(\theta|x)$, alors l'égalité

$$\frac{c_2}{c_1} = \frac{\mathbb{E}^{\pi_2}[\tilde{\pi}_1(\theta|x)\,h(\theta)]}{\mathbb{E}^{\pi_1}[\tilde{\pi}_2(\theta|x)\,h(\theta)]} \tag{7.13}$$

est vraie pour toute *fonction passerelle* $h(\theta)$ telle que les deux espérances soient finies (Exercice 7.21). L'estimateur par échantillonnage de passerelle est alors

$$B_{12}^S = \frac{\dfrac{1}{n_1}\displaystyle\sum_{i=1}^{n_1} \tilde{\pi}_2(\theta_{1i}|x)\,h(\theta_{1i})}{\dfrac{1}{n_2}\displaystyle\sum_{i=1}^{n_2} \tilde{\pi}_1(\theta_{2i}|x)\,h(\theta_{2i})}\,, \tag{7.14}$$

où les θ_{ji} sont simulés selon les lois $\pi_j(\theta|x)$ $(j = 1, 2,\ i = 1, \ldots, n_j)$.

Par exemple, si

$$h(\theta) = 1/\left[\tilde{\pi}_1(\theta|x)\tilde{\pi}_2(\theta_{1i}|x)\right]\,,$$

B_{12}^S est un rapport de moyennes harmoniques, généralisant (7.11). Meng et Wong (1996) calculent une fonction passerelle (asymptotiquement) optimale

$$h^*(\theta) = \frac{n_1 + n_2}{n_1\pi_1(\theta|x) + n_2\pi_2(\theta|x)}\,.$$

Cette expression n'est pas directement exploitable, puisque les constantes de normalisation de $\pi_1(\theta|x)$ et $\pi_2(\theta|x)$ sont inconnues. (Il s'agit précisément de la raison pour laquelle nous avons recours à ces techniques !) Néanmoins, elle montre qu'une bonne fonction passerelle doit couvrir les supports des deux distributions a posteriori, dans les mêmes proportions si $n_1 = n_2$.

Exemple 7.10. Dans le cas des *modèles linéaires généralisés*, c'est-à-dire des modèles explicatifs (ou conditionnels) liés aux familles exponentielles,

$$f(y|\theta) = h(y)\,e^{\theta\cdot y - \psi(\theta)}\,,$$

la moyenne $\mathbb{E}[y|\theta] = \nabla\psi(\theta)$ étant une fonction des covariables, x, de la forme $\nabla\psi(\theta) = \Psi(x^t\beta)$, le choix de la *fonction de lien* Ψ n'est jamais évident. Lorsque la variable expliquée y est à valeurs dans $\{0, 1\}$ et

$$\mathbb{E}[y|x] = P(y = 1|x)\,,$$

les choix suivants de Ψ sont par exemple courants (McCullagh et Nelder, 1989)

– la fonction de lien *logit*, $\Psi(t) = \exp(t)/(1 + \exp(t))$;
– la fonction de lien *probit*, $\Psi(t) = \Phi(t)$, fonction de répartition de la distribution $\mathcal{N}(0, 1)$; et
– la fonction de lien *log-log*, $\Psi(t) = 1 - \exp(-\exp(t))$.

Bien que diverses justifications soient avancées pour chacune des fonctions (Gouriéroux et Monfort, 1996), elles sont insuffisantes pour éliminer les deux autres possibilités. Les trois modèles en compétition sont alors

$$\mathcal{M}_1 : y|x \sim \frac{e^{yx^t\beta_1}}{1 + e^{yx^t\beta_1}}$$

$$\mathcal{M}_2 : y|x \sim \Phi(x^t\beta_2)^y\,[1 - \Phi(x^t\beta_2)]^{1-y}$$

$$\mathcal{M}_3 : y|x \sim \exp\{-(1-y)\exp(x^t\beta_3)\}\,[1 - \exp\{-\exp(x^t\beta_3)\}]^y\,.$$

Si la loi a priori π sur les β_i est normale, $\beta \sim \mathcal{N}_p(\xi, \tau^2 I_p)$, et si la fonction passerelle est $h(\beta) = 1/\pi(\beta)$, l'estimateur d'échantillonnage par passerelle est alors $(1 \leq i < j \leq 3)$

$$B_{ij}^S = \frac{1}{n}\sum_{t=1}^{n} L_j(\beta_{it}|x) \Bigg/ \frac{1}{n}\sum_{t=1}^{n} L_i(\beta_{jt}|x)\,,$$

où les β_{it} sont simulés[63] selon $\pi_i(\beta_i|x) \propto L_i(\beta_i|x)\pi(\beta_i)$. ‖

Dans un cas particulier où les deux lois a priori sont égales, à un hyperparamètre près, Gelman et Meng (1998) décrivent une meilleure méthode que l'échantillonnage par passerelle, appelée *échantillonnage par chemin* (traduction de *path sampling*) et présentée dans la Note 7.8.1.

7.3.3 Méthodes MCMC

Bien que l'échantillonnage d'importance semble particulièrement indiqué dans ce contexte, on peut également faire appel à des méthodes MCMC pour simuler des échantillons de distributions complexes. Par exemple, l'estimation par échantillonnage par passerelle peut s'appuyer sur des échantillons MCMC plutôt que sur des échantillons i.i.d. si les lois $\pi_j(\theta|x)$ sont trop compliquées, comme dans l'Exemple 7.10.

[63] Le détail de la simulation par algorithme MCMC de ces lois a priori est abordé dans Robert et Casella (1999, Note 9.7.3) et Gelman *et al.* (2003) par exemple. Une solution repose sur l'utilisation de l'algorithme de Metropolis-Hastings à marche aléatoire (voir Section 6.3.2).

Exemple 7.11. (Suite de l'Exemple 7.10) Dans les modèles $\mathscr{M}_j$ ($j = 1, 2, 3$), la partie de la loi a posteriori liée à la vraisemblance est

$$\prod_{i=1}^{n} \Psi(x_i^t\beta_j)^{y_i}[1-\Psi(x_i^t\beta_j)]^{1-y_i}\,.$$

Dans le cas de la fonction de lien probit ($j = 2$), on a $\Psi(t) = \Phi(t)$, fonction de répartition de la loi normale $\mathscr{N}(0, 1)$. Une solution naturelle fondée sur l'échantillonnage de Gibbs est alors de créer des variables auxiliaires $z_i \sim \mathscr{N}(0, 1)$ telles que $\Psi(x_i^t\beta_2) = \mathbb{E}\left[\mathbb{I}_{z_i \le x_i^t\beta_2}\right]$, ce qui revient à simuler selon la distribution jointe

$$\pi(\beta_2, z_1, \ldots, z_n) \propto \pi(\beta_2) \prod_{i=1}^{n} \mathbb{I}^{y_i}_{z_i \le x_i^t\beta_2} \mathbb{I}^{1-y_i}_{z_i \ge x_i^t\beta_2}\,.$$

Pour les deux autres fonctions de lien, un échantillonneur par tranche standard (voir Section 6.3.6) convient : pour le modèle logit, l'inégalité $u_i \le \Psi(x_i^t\beta_1)$ permet de déduire le résultat

$$x_i^t\beta_1 \ge \log(u_i/(1-u_i))\,,$$

et, pour le modèle log-log, $u_i \le \Psi(x_i^t\beta_3)$ est équivalent à

$$x_i^t\beta_3 \ge \log(-\log(1-u_i))\,.$$

Pour les trois modèles, les composantes des β_j sont donc simulées selon des distributions normales multidimensionnelles tronquées. ‖

Par conséquent, l'approximation (7.12) de la distribution marginale peut être calculée sur un échantillon MCMC $(\theta^{(t)})$ de $\pi(\theta|x)$.

Exemple 7.12. (Suite de l'Exemple 7.4) Si les distributions a priori des quatre modèles sont de la forme ($j = 1, \ldots, 4$)

$$\pi_j(\beta_{\cdot j}, \sigma_j^2, \tau_j^2) \propto \sigma_j^2\tau_j^2 e^{-2(\sigma_j^{-2}+\tau_j^{-2})}\,,$$

en notant $\beta_{\cdot j}$ le vecteur contenant les β_{ij} pour le modèle $\mathscr{M}_j$, Gelman (1996) suggère d'évaluer les quatre modèles en simulant un échantillon de $\theta_j^{(t)}$ selon les lois a posteriori correspondantes, en adoptant les approximations suivantes pour les distributions prédictives

$$\hat{f}_j(y|y_1, \ldots, y_n) = \frac{1}{T}\sum_{t=1}^{T} f_j(y|\theta_j^{(t)})\,,$$

puis de vérifier si des échantillons tirés selon ces lois prédictives correspondent à l'échantillon $y_1, \ldots, y_n$. Les résultats de cette expérience sont rapportés dans

le Tableau 7.2 : on remarque que les modèles $\mathscr{M}_3$ et $\mathscr{M}_4$ s'accordent de façon satisfaisante avec les intervalles prédictifs, contrairement aux modèles $\mathscr{M}_1$ et $\mathscr{M}_2$. Il est évident qu'il ne s'agit là que d'un premier indicateur d'ajustement et qu'il faudrait ensuite calculer les facteurs de Bayes exacts, mais cette évaluation empirique peut être suffisante pour éliminer les modèles les moins adaptés. ‖

Tab. 7.2. Adéquation des quatre modèles de prédiction de croissance d'orangers, en pourcentage des observations à l'intérieur des intervalles prédictifs à 50% et 90%. (*Source :* Gelfand, 1996.)

Modèle	50%	95%
$\mathscr{M}_1$	89	100
$\mathscr{M}_2$	29	51
$\mathscr{M}_3$	46	100
$\mathscr{M}_4$	60	86

Chib (1995) propose d'utiliser l'échantillonneur de Gibbs pour l'approximation de densités marginales, en adoptant la représentation bayésienne suivante. Quelle que soit θ, valeur fixe du paramètre, la formule de Bayes implique que

$$\log m(x) = \log f(x|\theta) + \log \pi(\theta) - \log \pi(\theta|x)\,.$$

Lorsque $\theta = (\theta_1, \theta_2)$ et lorsque $\pi(\theta_1|\theta_2, x)$ et $\pi(\theta_2|\theta_1, x)$ sont tous les deux calculables analytiquement, *constantes de normalisation comprises*, l'argument de Rao-Blackwellisation de la Section 6.3.4 fournit une approximation des lois marginales a posteriori $\pi(\theta_1|x)$

$$\hat{\pi}(\theta_1|x) = \frac{1}{T}\sum_{t=1}^{T} \pi(\theta_1|\theta_2^{(t)}, x)\,,$$

les $\theta_2^{(t)}$ étant simulés par un échantillonneur de Gibbs. (Notons que le choix de partitionner θ en (θ_1, θ_2) est guidé par la possibilité de calculer explicitement $\pi(\theta_1|\theta_2, x)$ et $\pi(\theta_2|\theta_1, x)$.) Chib (1995) établit alors l'approximation suivante de $\log m(x)$:

$$\log f(x|\theta) + \log \pi(\hat{\theta}) - \log \pi(\hat{\theta}_2|\hat{\theta}_1, x) - \log \hat{\pi}(\hat{\theta}_1|x)\,,$$

avec $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$ une approximation de l'estimateur MAP de θ, par exemple. Si les densités conditionnelles ne sont pas toutes les deux calculables analytiquement ou bien si on ne dispose pas d'une des constantes de normalisation, Chib (1995) propose d'introduire plus de partitions, mais le calcul en est d'autant plus compliqué (Exercice 7.24).

L'avantage le plus notable des techniques MCMC pour le choix de modèle est leur capacité à prendre en compte les *modèles à dimensions variables*, c'est-à-dire les modèles $\mathscr{M}_k$ reposant sur différents ensembles de paramètres, sans intersection entre eux et éventuellement de dimensions différentes.

Exemple 7.13. (Suite de l'Exemple 7.2) La dimension de l'espace des paramètres pour un mélange normal à k composantes est $3k-1$, en prenant en compte la contrainte

$$\sum_{\ell=1}^{k} p_{k\ell} = 1\,.$$

Si la loi a priori sur k est une distribution de Poisson $\mathscr{P}(\lambda)$, l'espace des paramètres est de dimension infinie, puisque k n'est pas borné. ‖

Si, pour le choix de modèle, la difficulté essentielle réside dans le calcul de la probabilité a posteriori correspondant au modèle $\mathscr{M}_k$, $\pi(\mu = k|x)$, cette représentation pose également des problèmes plus fondamentaux, le premier étant la notion même de paramètres du modèle, qui peut être décrite soit comme une suite $(\theta_1, \ldots, \theta_k, \ldots)$, soit comme un couple (k, θ_k). Un autre point délicat concerne la difficulté en théorie de la mesure à représenter une densité a priori sur une somme directe d'espaces. La construction des échantillonneurs MCMC correspondants n'en est que plus compliquée.

Une première solution, proposée par Carlin et Chib (1995), consiste à *saturer le modèle*, c'est-à-dire à considérer tous les modèles à la fois : pour un ensemble fini de modèles $\mathscr{M}_k$ $(k = 1, \cdots, K)$ avec des lois a priori associées $\pi_k(\theta_k)$ et des poids a priori ϱ_k, l'espace des paramètres est

$$\Theta = \{1, \ldots, K\} \times \prod_{k=1}^{K} \Theta_k$$

et, si μ représente l'indicateur de modèle, la distribution a posteriori s'écrit

$$\pi(\mu, \theta_1, \ldots, \theta_K|x) \propto \varrho_\mu f_\mu(x|\theta_\mu) \prod_{k=1}^{K} \pi_k(\theta_k)\,.$$

Puisque

$$m(x|\mu = j) = \int f_j(x|\theta_j)\pi(\theta_1, \ldots, \theta_K|\mu = j)\, d\theta = \int f_j(x|\theta_j)\pi_j(\theta_j)\, d\theta_j$$

ne dépend pas des $\pi_k(\theta_k)$ pour $k \neq j$, Carlin et Chib (1995) proposent d'utiliser des *pseudo-lois a priori* $\tilde{\pi}_k(\theta_k|\mu = j)$ pour simuler les paramètres θ_k lorsque $k \neq j$. Ils implémentent cette méthode à l'aide d'un échantillonneur de Gibbs sur $(\mu, (\theta_1, \ldots, \theta_K))$, en simulant μ selon

$$P(\mu = j|x, \theta_1, \ldots, \theta_K) \propto \varrho_j f_j(x|\theta_j)\pi_j(\theta) \prod_{k \neq j} \tilde{\pi}_k(\theta_k|\mu = j)\,.$$

Les auteurs remarquent que, assez naturellement, cette méthode donne de meilleurs résultats lorsque les pseudo-lois a priori sont proches des vraies distributions a posteriori, mais il existe toujours un risque de négliger des régions importantes des espaces des paramètres Θ_k dans la calibration des pseudo-lois a priori. L'inconvénient essentiel de la méthode de Carlin et Chib (1995) est que réaliser une simulation pour *chacun* des modèles à *chaque* étape de l'algorithme est coûteux en termes de temps de calcul lorsque K est grand. De plus, lorsque K est infini, cette technique ne peut pas être utilisée.

Exemple 7.14. (Carlin et Chib, 1995) On considère un jeu de mesures sur quarante-deux pins. On réalise une régression sur la variable grain (force du bois) y_i en fonction soit de la densité du bois x_i, soit d'une densité modifiée (adaptée à la résine) z_i. Les deux modèles en concurrence sont

$$\mathscr{M}_1 : y_i = \alpha + \beta x_i + \sigma\varepsilon_i$$

et

$$\mathscr{M}_2 : y_i = \gamma + \delta z_i + \tau\varepsilon_i\,,$$

avec $(\alpha, \beta, \sigma^2)$ et (γ, δ, τ^2) tous deux associés aux lois a priori (bayésiennes empiriques) conjuguées :

$$\begin{pmatrix}\alpha\\ \beta\end{pmatrix}, \begin{pmatrix}\gamma\\ \delta\end{pmatrix} \sim \mathscr{N}\left(\begin{pmatrix}3000\\ 185\end{pmatrix}, \begin{bmatrix}10^6 & 0\\ 0 & 10^4\end{bmatrix}\right), \quad \sigma^2, \tau^2 \sim \mathscr{IG}(a, b)\,,$$

(a, b) étant choisis tels que la moyenne et l'écart type de σ^2 et τ^2 soient 300^2. (Dans une analyse bayésienne réelle, il faudrait évaluer les conséquences de cette modélisation a priori par une analyse de robustesse comme le décrit la Section 3.6.)

Dans ce cas, les pseudo-distributions a priori sont fixées à partir des lois a priori sur σ^2 et τ^2 et de vagues lois a priori conjuguées sur (α, β) et (γ, δ) :

$$\begin{aligned}\alpha|\mu = 2 &\sim \mathscr{N}(3000, 52^2)\,, \quad & \beta|\mu = 2 &\sim \mathscr{N}(185, 12^2)\,,\\ \gamma|\mu = 1 &\sim \mathscr{N}(3000, 43^2)\,, \quad & \delta|\mu = 1 &\sim \mathscr{N}(185, 9^2)\,.\end{aligned}$$

Afin de forcer la prise en compte du modèle $\mathscr{M}_1$, les auteurs utilisent des poids déséquilibrés, $\varrho_1 = .9995$ et $\varrho_2 = .0005$. (Cette pratique semble être assez courante dans l'approche à base de pseudo-lois a priori et est une façon de compenser un éventuel mauvais choix de pseudo-lois a priori.)

Ils obtiennent une approximation de 4 420 pour B_{21} (après correction des poids), avec un intervalle de confiance simulé de $(4\,353, 4\,487)$. (L'intervalle de confiance est simplement déduit de la variance binomiale sur la probabilité a posteriori $P(\mu = 1|x)$.) Le modèle $\mathscr{M}_2$ peut donc être privilégié sans grand risque. ‖

Exemple 7.15. (Suite de l'Exemple 7.2) Dans le cas des modèles de mélanges de galaxies, en se posant seulement le problème de choisir entre toirs (modèle $\mathscr{M}_1$) et quatre (modèle $\mathscr{M}_2$) composantes, Carlin et Chib (1995) appliquent un modèle de données complétées comme dans la Section 6.4, en considérant des affectations z_i^k $(i = 1, \cdots, n,\ \ k = 1, 2)$. Comme dans l'Exemple 7.14, on utilise les résultats de tests préliminaires sur les deux distributions pour fixer les valeurs des pseudo-lois a priori. Celles qui portent sur les paramètres sont les distributions conjuguées correspondant aux estimateurs a posteriori de chaque modèle, alors que les pseudo-lois a priori des z_i^μ, pour $\mu \neq k$, sont calculées à partir des fréquences observées. Les auteurs évaluent le facteur de Bayes à 0.5153, avec un écart type de 0.0146, ce qui plaide (modérément) pour le modèle à trois composantes. (Mais ils indiquent aussi que ce résultat peut être modifié jusqu'à prendre la décision inverse, *contre* le modèle à trois classes, en choisissant simplement d'autres lois a priori sur les poids.) ‖

7.3.4 MCMC à sauts réversibles

Pour les modèles à dimensions variables, Green (1995) propose un autre type de technique de *saturation*, plus localisée que celle de Carlin et Chib (1995). Étant donné deux modèles $\mathscr{M}_1$ et $\mathscr{M}_2$ de dimensions éventuellement distinctes, l'idée de base est d'éliminer la différence entre les dimensions en complétant les paramètres respectifs θ_1 et θ_2 avec des variables auxiliaires $u_{1\to2}$ et $u_{2\to1}$ telles que

$$(\theta_1, u_{1\to2}) \ \text{ et } \ (\theta_2, u_{2\to1})$$

soient en bijection :

$$(\theta_2, u_{2\to1}) = \Psi_{1\to2}(\theta_1, u_{1\to2})\ . \tag{7.15}$$

Si θ_1 est distribué selon une loi $\pi_1(\theta_1)$ et $u_{1\to2}$ selon $g_{1\to2}(u)$, la distribution de (7.15) s'écrit

$$\pi_1(\theta_1) g_{1\to2}(u_{1\to2}) \left| \frac{\partial \Psi_{1\to2}(\theta_1, u_{1\to2})}{\partial(\theta_1, u_{1\to2})} \right|^{-1}$$

d'après la formule du jacobien. Si nous souhaitons à présent vérifier si (7.15) est distribué selon une loi $\pi_2(\theta_2) g_{2\to1}(u_{2\to1})$, la probabilité d'acceptation de Metropolis-Hastings est

$$\min\left(\frac{\pi_2(\theta_2) g_{2\to1}(u_{2\to1})}{\pi_1(\theta_1) g_{1\to2}(u_{1\to2})} \left| \frac{\partial \Psi_{1\to2}(\theta_1, u_{1\to2})}{\partial(\theta_1, u_{1\to2})} \right|, 1 \right)\ .$$

À l'inverse de l'approche adoptée par Carlin et Chib (1995), cette technique ne considère que des modifications locales d'un modèle à un autre :

un déplacement de $\mathscr{M}_i$ vers $\mathscr{M}_j$ n'utilise explicitement que les θ_j et variables auxiliaires $u_{i\to j}$ associés.

La théorie sous-tendant les méthodes de *MCMC à sauts réversibles* ne se résume naturellement pas à la présentation succincte ci-dessus, ne serait-ce que parce qu'elle est plus exigeante à l'égard de la densité jointe sur $(\theta_2, u_{2\to 1})$ et $(\theta_1, u_{1\to 2})$, qui doit satisfaire une *condition d'équilibre ponctuel* comme en (6.17). Les lecteurs intéressés pourront consulter Green (1995) et Richardson et Green (1997) pour plus de détails. Le point essentiel est que, étant donné la probabilité $\varrho_{i\to j}$ de choisir le modèle $\mathscr{M}_j$ à partir du modèle $\mathscr{M}_i$, la probabilité d'acceptation d'un déplacement s'écrit effectivement

$$\min\left(\frac{\varrho_j\varrho_{j\to i}\pi_j(\theta_j)g_{j\to i}(u_{j\to i})}{\varrho_i\varrho_{i\to j}\pi_i(\theta_i)g_{i\to j}(u_{i\to j})}\left|\frac{\partial\Psi_{i\to j}(\theta_i,u_{i\to j})}{\partial(\theta_i,u_{i\to j})}\right|,1\right), \tag{7.16}$$

avec $(\theta_j, u_{j\to i}) = \Psi_{i\to j}(\theta_i, u_{i\to j})$, sous réserve que le déplacement de $\mathscr{M}_i$ vers $\mathscr{M}_j$ vérifie aussi cette relation. L'algorithme peut alors être complété par des étapes supplémentaires liées à un modèle particulier $\mathscr{M}_i$ ou à des hyperparamètres non dépendants du modèle.

Comme l'indiquent Robert et Casella (2004, Section 6.5.1), l'algorithme à sauts réversibles offre une liberté telle qu'il a trouvé nombre d'applications, bien au-delà du cadre du choix de modèle. Dans la situation de l'Exemple 7.2, Richardson et Green (1997) élaborent un algorithme à sauts réversibles pour les composantes normales, qui conclut que le nombre de composantes pour les données de vitesses de galaxies devrait être de quatre. Nous présentons ci-dessous l'algorithme correspondant pour un mélange de distributions exponentielles, provenant de Gruet *et al.* (1999). (Voir aussi Robert *et al.*, 1999b, pour une généralisation aux modèles de Markov cachés.)

Exemple 7.16. Pour un mélange de distributions exponentielles

$$\sum_{j=1}^{k} p_{jk}\mathscr{E}xp(\lambda_{jk})\,,$$

l'algorithme à sauts réversibles peut être limité à des déplacements entre modèles voisins, c'est-à-dire entre le modèle $\mathscr{M}_k$ et les modèles $\mathscr{M}_{k+1}$ et $\mathscr{M}_{k-1}$. Les mouvements sont assez libres : une composante peut être ajoutée (ou retirée) aléatoirement, tant que la symétrie entre déplacements *montants* et *descendants* est préservée. Par exemple, la naissance de la composante $k+1$ sera proposée en simulant $(p_{(k+1)(k+1)}, \lambda_{(k+1)(k+1)})$ selon la loi a priori $\varpi_{k+1}(p,\lambda)$, en supposant un a priori commun à toutes les composantes. La transformation est alors

$$\begin{aligned}(p_{1(k+1)},\ldots,p_{k(k+1)}) &= ((1-p_{(k+1)(k+1)})p_{1k},\ldots,(1-p_{(k+1)(k+1)})p_{kk})\\ (\lambda_{1(k+1)},\ldots,\lambda_{(k+1)(k+1)}) &= (\lambda_{1k},\ldots,\lambda_{kk},\lambda_{(k+1)(k+1)})\,.\end{aligned}$$

Le jacobien de cette transformation est donc $(1-p_{(k+1)(k+1)})^k$ et la probabilité d'accepter la naissance est

$$\min\left(\frac{\varrho_{k+1}(1-p_{(k+1)(k+1)})^k\,\pi_{k+1}(p_{1(k+1)},\ldots,p_{(k+1)(k+1)},}{\varrho_k\pi_k(p_{1k},\ldots,p_{kk},\lambda_{1k},\ldots,\lambda_{kk})}\right.$$
$$\left.\frac{\lambda_{1(k+1)},\ldots,\lambda_{(k+1)(k+1)})}{\varpi_{k+1}(p_{(k+1)(k+1)},\lambda_{(k+1)(k+1)})},1\right),$$

si les probabilités de choisir une naissance (saut vers $\mathscr{M}_{k+1}$) ou une mort (saut vers $\mathscr{M}_{k-1}$) sont égales.

Le déplacement de $\mathscr{M}_k$ vers $\mathscr{M}_{k+1}$ considéré dans Gruet *et al.* (1999) consiste en la *séparation* d'une composante j choisie aléatoirement de façon à ce que les paramètres $(p_{j(k+1)}, p_{(j+1)(k+1)}, \lambda_{j(k+1)}, \lambda_{(j+1)(k+1)})$ de la nouvelle composante satisfassent la condition des moments

$$\begin{aligned} p_{jk} &= p_{j(k+1)} + p_{(j+1)(k+1)} \\ p_{jk}\lambda_{jk} &= p_{j(k+1)}\lambda_{j(k+1)} + p_{(j+1)(k+1)}\lambda_{(j+1)(k+1)}\,. \end{aligned} \tag{7.17}$$

Le déplacement inverse est la *fusion* de deux composantes j et $j+1$ selon l'équation (7.17). On peut tout aussi bien représenter la séparation en simulant deux variables $u_1, u_2 \sim \mathscr{U}([0,1])$, puis $p_{j(k+1)} = u_1 p_{jk}$ et $\lambda_{j(k+1)} = u_2\lambda_{jk}$. On obtient alors le jacobien

$$\frac{\partial\Psi_{k\to k+1}(p_{jk},\lambda_{jk},u_1,u_2)}{\partial(p_{jk},\lambda_{jk},u_1,u_2)} = p_{jk}/(1-u_1)\,.$$

La Figure 7.2 présente une analyse succincte des performances de l'algorithme à sauts réversibles sur un jeu de données portant sur des séjours hospitaliers avec un mode a posteriori pour k de 4. La carte d'allocation en bas à droite représente les affectations successives des observations en niveaux de gris : on voit que les propriétés de mélange de la chaîne sont bonnes, puisque aucune forme particulière n'émerge. (Voir Gruet *et al.*, 1999, pour plus de détails.) ‖

Notons l'absence de variable auxiliaire $u_{k\to(k-1)}$ pour les mouvements descendants dans les deux situations décrites dans l'Exemple 7.16. Cela se produit souvent lorsqu'un modèle inclut l'autre, mais l'addition de variables auxiliaires est parfois tout de même conseillée dans un souci de gain en temps de calcul.

Des techniques analogues sont décrites dans Ripley (1987), Grenander et Miller (1994), Phillips et Smith (1996) et Stephens (2000), mettant en jeu les naissances et morts de processus à temps continu. (Voir la Note 7.8.2.)

7.4 Moyenne de modèles

Un geste bayésien assez naturel devant l'incertitude sur le choix de modèle est d'inclure *tous* les modèles $\mathscr{M}_k$ envisagés dans la prise de décision, faisant ainsi l'économie de l'étape de choix de modèle. L'idée sous-jacente est qu'on

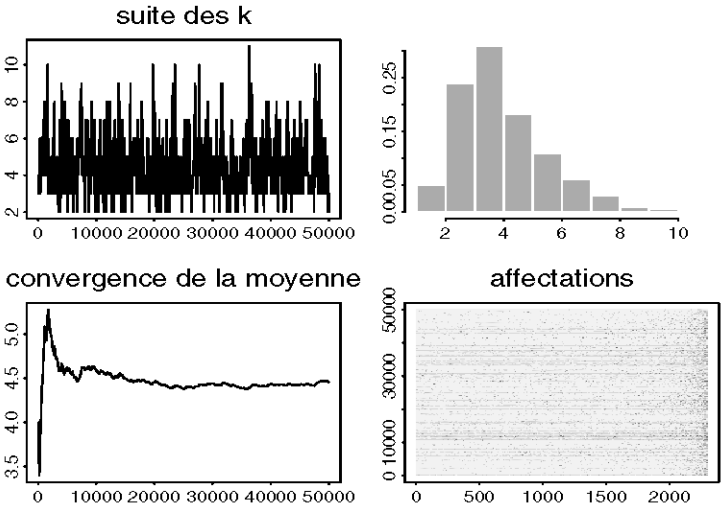

Fig. 7.2. Suite de valeurs $k^{(t)}$ simulées par sauts réversibles, avec l'histogramme correspondant *en haut, à droite*; la convergence de la moyenne empirique *en bas, à gauche*; et la séquence d'affectations aux composantes *en bas, à droite* pour 50 000 itérations. (*Source :* Gruet *et al.*, 1999.)

sous-estime généralement l'incertitude présente à l'étape du choix de modèle en choisissant un modèle, disons $\mathscr{M}_{k_0}$, et en oubliant totalement le caractère aléatoire de ce choix dans les étapes ultérieures. La solution de la moyennisation de tous les modèles, proposée par Raftery *et al.* (1996), permet de remédier à ce problème.

Ce principe n'est évidemment pas applicable dans tous les contextes : le but de la personne qui prend les décisions, ou du statisticien, est justement parfois de choisir un modèle, comme c'est le cas en inférence scientifique, ou d'éliminer les covariables superflues d'un modèle à cause de coûts d'échantillonnage prohibitifs (Section 7.5), dans le domaine de la sélection de variables. De plus, la moyenne de modèles va à l'encontre des efforts de parcimonie (Note 6.6.6) dans la mesure où l'imbrication de tous les modèles dans un seul (super) modèle fait augmenter d'autant le nombre de paramètres et nécessite la simulation et le stockage d'un grande quantité d'échantillons MCMC, puisqu'on fait appel à des algorithmes numériques dans la plupart des cas. C'est en particulier vrai dans l'Exemple 7.2.

Le principe de cette approche est le suivant : pour un échantillon $\mathbf{x} = (x_1, \ldots, x_n)$, la distribution prédictive est obtenue par une moyenne sur tous les modèles possibles,

$$
\begin{aligned}
f(y|x) &= \int_{\boldsymbol{\Theta}} f(y|\theta)\pi(\theta|\mathbf{x})d\theta \\
&= \sum_k \int_{\Theta_k} f_k(y|\theta_k)\pi(k,\theta_k|\mathbf{x})d\theta_k \\
&= \sum_k p(\mathscr{M}_k|\mathbf{x}) \int f_k(y|\theta_k)\pi_k(\theta_k|\mathbf{x})\, d\theta_k \,,
\end{aligned}
$$

en notant $\boldsymbol{\Theta}$ l'espace des paramètres global, tel que défini en (7.2).

Cette idée ne permet malheureusement pas d'échapper à la plupart des problèmes déjà décrits en Section 7.2, comme les nombreux calculs d'intégrales et les simulations sur un espace des paramètres, $\boldsymbol{\Theta}$, qui est une somme d'espaces de différentes dimensions. Néanmoins, le contournement de l'étape de décision sur le label μ du modèle permet d'alléger certaines difficultés. Par exemple, le fait que la collection de modèles soit éventuellement infinie (ou simplement trop grande, comme dans la sélection de variables) n'est pas rédhibitoire dans la mesure où un algorithme MCMC explorant $\boldsymbol{\Theta}$ pourra ignorer les modèles aux probabilités $P(\mathscr{M}_i|\mathbf{x})$ très faibles.

Le problème ici relève davantage de la modélisation, comme nous l'avons vu en Section 7.2.1 : lorsqu'on doit considérer un grand nombre de modèles, le choix des probabilités a priori $\pi(k)$ est fondamental, mais difficile à formaliser et à justifier. Par exemple, dans le cadre de la sélection de variables (Section 7.5), les modèles en concurrence peuvent être représentés par des vecteurs d'indicatrices

$$
\mathscr{M}_k : (\delta_{k1}, \cdots, \delta_{kd}) \,, \qquad \delta_{kj} \in \{0,1\} \,,
$$

avec d nombre de covariables potentielles. Madigan et Raftery (1991) proposent d'utiliser

$$
\pi(k) \propto \prod_{j=1}^{d} \left\{ \varrho_j^{\delta_{kj}} (1-\varrho_j)^{1-\delta_{kj}} \right\} \,,
$$

avec ϱ_j probabilité a priori que *la variable j ait un effet*. Une limitation prévisible de cette distribution est que les covariables sont incluses dans le modèle indépendamment les unes des autres. Cette stratégie n'est justifiée que si elles sont indépendantes, ce qui est une hypothèse hasardeuse dans la plupart des cas. Une autre idée immédiate consistant à mettre des poids égaux à tous les modèles n'est pas moins critiquable : outre le fait que ce soit impossible lorsque le nombre de modèles est infini, cette stratégie semble particulièrement peu pertinente pour des modèles imbriqués, c'est-à-dire lorsque certains modèles sont des cas particuliers d'autres, comme c'est le cas en sélection de variables.

Un avantage des techniques MCMC telles que les sauts réversibles ou les processus de saut (Note 7.8.2), déjà décrit ci-dessus, est leur capacité à explorer un grand nombre de modèles en évitant ceux auxquels sont affectées des probabilités faibles (en admettant que les algorithmes correspondants

convergent correctement). Madigan et Raftery (1991) proposent une autre solution appelée *fenêtre d'Occam*[64]. Ils suggèrent de ne considérer que les modèles tels que

$$\frac{\max_k P(\mathscr{M}_k|x)}{P(\mathscr{M}_\ell|x)} \leq C$$

c'est-à-dire seulement les modèles dont la probabilité n'est pas trop éloignée du modèle le plus probable. Ils conseillent en outre d'exclure les modèles $\mathscr{M}_\ell$, tels qu'il existe un sous-modèle $\mathscr{M}_h \subset \mathscr{M}_\ell$ vérifiant

$$\frac{P(\mathscr{M}_h|x)}{P(\mathscr{M}_\ell|x)} \geq 1\,.$$

Mais une telle réduction dans le nombre de modèles n'est implémentable que si ce nombre est au départ relativement modeste et Clyde (1999) met en garde contre l'apparition possible de biais dans les probabilités résultant de cette simplification.

Exemple 7.17. Dans le cadre de la sélection de variables en régression normale, $y \sim \mathscr{N}(X\beta, \sigma^2 I)$, c'est-à-dire lorsque

$$y_t = \sum_{j=1}^{J} \beta_j x_{jt} + \sigma\varepsilon_t \qquad t = 1, \cdots, T\,,$$

avec des régresseurs orthogonaux

$$X^t X = \text{diag}(x_j' x_j)\,,$$

Clyde (1999) propose des distributions a priori de la forme

$$\beta_j \sim \mathscr{N}(0, c_j^2 \gamma_j)\,, \qquad \gamma_j \sim \mathscr{B}(p_j)\,,$$

les γ_j jouant le rôle d'indicateurs 0-1 pour la présence du j-ième régresseur dans le modèle. Alors, sous l'a priori de Madigan et Raftery (1991),

[64]William d'Occam ou d'Ockham (*circa* 1285–*circa* 1349), théologien anglais (et moine franciscain) d'Oxford, a travaillé sur les bases de l'induction empirique et, en particulier, posé le principe appelé plus tard *"rasoir" d'Occam* (Occam's razor), qui écarte l'admission de causes multiples pour un phénomène si elles ne sont pas justifiées expérimentalement (voir Adams, 1987). Ce principe, *Pluralitas non est ponenda sine necessitate* (traduit généralement par *les entités ne devraient pas être multipliées sans nécessité*) est souvent invoqué en tant que *principe de parcimonie* pour privilégier l'explication la plus simple lorsque deux explications sont également possibles. On le retrouve très fréquemment dans la littérature bayésienne (voir, par exemple, Jeffreys, 1961, Section 6.12, ou Jefferys et Berger, 1992). Nous sommes néanmoins réticents sur l'emploi de cette notion, car elle ne fournit pas un principe de travail et peut donc être utilisée à tort. Il est clair que le seul argument du *rasoir d'Occam* n'est pas suffisant pour justifier pleinement une méthode donnée. (Pour l'anecdote, le personnage de William de Baskerville dans *Le Nom de la Rose* d'Umberto Eco est inspiré d'Occam.)

$$\pi(\gamma_1,\ldots,\gamma_J|y,\sigma) = \prod_{j=1}^{J} \varrho_j^{\gamma_j}(1-\varrho_j)^{\gamma_j}\,, \tag{7.18}$$

avec

$$\varrho_j = \frac{O_j(y,\sigma)}{1+O_j(y,\sigma)}$$

et

$$O_j(y,\sigma) = \frac{p_j}{1-p_j}\left(\frac{x_j'x_j+\sigma^2/c_j^2}{\sigma^2/c_j^2}\right)^{-1/2} \times \exp\left\{\frac{(\hat\beta_j x_j'/\sigma^2)^2}{2(x_j'x_j/\sigma^2+1/c_j^2)}\right\}\,,$$

ce qui signifie que les γ_j sont indépendants a posteriori et que la probabilité d'un sous-modèle donné peut être déduite aisément ainsi que le sous-modèle le plus probable (Exercice 7.26). Ce n'est pas le cas avec la modélisation alternative de George et McCulloch (1997) :

$$\beta_j \sim \mathcal{N}\left(0, c_j^2\gamma_j + [c_j^2/100](1-\gamma_j)\right)\,.$$

Si σ^2 est inconnu, Clyde (1999) utilise le même a priori simple $\sigma^2 \sim \mathcal{IG}(\alpha,\beta)$ pour tous les modèles

$$\pi(\sigma^2|\gamma,y) \sim \mathcal{IG}(\hat\alpha,\hat\beta)$$

et évalue les poids a posteriori des différents modèles soit avec un estimateur intuitif, c'est-à-dire remplaçant σ^2 par un estimé $\hat\sigma^2$ dans l'égalité (7.18), soit avec une moyenne de Rao-Blackwell. ‖

Bien que de tels résultats soient intéressants, ils sont difficiles à transposer à d'autres cadres, comme les modèles linéaires généralisés, sans l'ajout de nouvelles approximations. Par ailleurs, l'hypothèse d'orthogonalité est trop restrictive, car les régresseurs courants ne sont jamais orthogonaux et leur appliquer une transformation orthogonale comme les composantes principales empêche d'obtenir les valeurs des coefficients β_j ce qui est souvent un objectif de l'étude. Enfin, le principe que les paramètres communs doivent être traités comme des entités distinctes dans des modèles différents n'est pas ici respecté, puisque les β_j sont identiques dans tout modèle où ils apparaissent.

7.5 Projections de modèles

Nous présentons dans cette section une approche différente[65] du choix de modèle, développée par Goutis et Robert (1998), puis appliquée à la sélection

[65] Cette section contient des notions moins générales. Elle n'est pas plus difficile que le reste de ce chapitre, mais peut être laissée de côté en première lecture.

de variables par Dupuis et Robert (2001). L'idée sous-tendant cette approche est de *projeter* un modèle complet $f(y|\theta)$ sur des sous-modèles, obtenus par des restrictions sur θ, puis de calculer l'erreur d'approximation commise. Cette approche est en particulier applicable à la *sélection de variables*, c'est-à-dire à la recherche d'un sous-ensemble de covariables, au sein d'un ensemble plus grand (Exemple 7.17).

Exemple 7.18. Dans une étude sur l'influence de facteurs diététiques sur l'apparition de cancer du sein (CS), Raftery et Richardson (1995) considèrent les covariables suivantes :

âge	âge de la première grossesse
âge à la ménopause	âge à la fin des études
âge à la ménarche	indice de masse corporelle
nombre d'enfants	consommation de graisses (totale)
consommation d'alcool	consommation de graisses (saturées)
antécédents familiaux de CS	antécédents de CS bénins

Les observations sont à valeurs dans $\{0,1\}$, correspondant à une dichotomie présence/absence de cancer. On peut donc leur appliquer une modélisation *logistique* impliquant toutes ou partie des covariables ($i = 1, \cdots, 2^{12}$) :

$$\mathscr{M}_i : P(y_j = 1|x_j) = \frac{\exp[\alpha_i + \beta_i^t x_j^{(i)}]}{1 + \exp[\alpha_i + \beta_i^t x_j^{(i)}]}\,,$$

en notant $x^{(i)}$ les coordonnées de x dans la décomposition binaire de i. Par exemple, le modèle $\mathscr{M}_5$ correspond à $i = 5 = \texttt{0}\cdots\texttt{0101}$ et donc $x^{(5)} = (x_{10}, x_{12})$. ‖

Une des principales différences entre l'approche par projections et les axiomes usuels de choix de modèle réside dans les distributions a priori requises. En effet, on ne demande ici la construction d'un a priori $\pi(\theta)$ que pour le modèle complet et on tolère les lois a priori impropres, ce qui n'etait pas le cas dans la Section 7.2, où un a priori propre *par sous-modèle* était nécessaire. En fait, comme nous le verrons ci-dessous, les poids et lois a priori de chaque sous-modèle sont déduits de la distribution a priori originale π, ce qui permet d'éviter les paradoxes de marginalisation et de projection liés à la présence de sous-espaces de dimensions différentes.

Pour une restriction $\theta \in \Theta_0$, Goutis et Robert (1998) proposent le critère d'acceptabilité suivant :

$$d(f(\cdot\,|\theta), \Theta_0) < \epsilon\,, \tag{7.19}$$

où d est une mesure de divergence et

$$\begin{aligned} d(f(\cdot\,|\theta), \Theta_0) &= d(f(\cdot\,|\theta), f(\cdot\,|\theta^\perp)) \\ &= \inf_{\theta_0 \in \Theta_0} d(f(\cdot\,|\theta)\,, f(\cdot\,|\theta_0))\,. \end{aligned}$$

Le paramètre $\theta^\perp$ est alors la projection du paramètre θ sur le sous-modèle. Le choix de modèle peut ainsi être vu comme une évaluation de la différence entre le vrai modèle et un modèle plus parcimonieux. Il s'agit donc d'une modélisation pragmatique tenant compte des réalités expérimentales, dans lesquelles la nullité exacte est rarement vérifiée, et qui règle les problèmes de paramétrisation par l'absence de paramètres dans la représentation (7.19). Par ailleurs, cette méthode ne nécessite que la distribution a priori sur le paramètre complet θ, puisque le paramètre de projection $\theta^\perp$ s'obtient à partir d'une transformation de θ. La probabilité a posteriori dans (7.19) peut donc être calculée en utilisant seulement la distribution a priori. Remarquons que cela n'est pas équivalent à établir la distribution a priori sur $\theta^\perp$ en projetant $\pi(\theta)$ et à utiliser ensuite le facteur de Bayes standard, comme le font McCulloch et Rossi (1992) (Exercice 7.33).

Il y a de nombreuses possibilités pour la mesure de divergence d, mais un choix assez naturel est la pseudo-distance de *Kullback-Leibler*

$$d(f,g) = \int \log\left(\frac{f(z)}{g(z)}\right) f(z)\, dz\,,$$

déjà vue en (2.7). Bernardo et Smith (1994) présentent de nombreux arguments défendant l'utilisation de cette mesure. Ils sont liés à la théorie de l'information, aux règles de pénalisation, aux propriétés de transitivité et d'additivité ou encore aux familles exponentielles et aux modèles linéaires généralisés.

De même, le facteur ϵ dans (7.19) peut être fixé de bien des façons différentes. Par exemple, il peut être calibré sur des distributions simples pour établir un intervalle raisonnable, comme dans la Table 7.3 (Exercice 7.29). Dans le cas d'une restriction simple, ϵ peut être déduit de la distribution (propre) a priori π pour vérifier la condition

$$P^\pi(d(f(\cdot\,|\theta), f(\cdot\,|\theta^\perp)) \leq \epsilon) = 1/2\,.$$

Ce travail a été réalisé dans le cadre des mélanges par Mengersen et Robert (1996), mais la valeur 1/2 est critiquable dans la mesure où elle donne une fausse impression d'objectivité (alors que le résultat dépend en fait de π). Enfin, dans le contexte de sélection de variables et de modèles imbriqués associés, il existe un modèle minimal (ou modèle plus rudimentaire), f_0, obtenu par la régression d'un seul intercept et qui peut donner un ordre de grandeur de ϵ par $\epsilon = \varrho d(f, f_0)$, avec $0 < \varrho < 1$. (Dupuis et Robert, 2001, appellent $d(f, f_0)$ le *coût maximal en potentiel explicatif*.)

Dès lors que d et ϵ sont fixés, la méthode peut être implémentée soit en calculant la probabilité a posteriori $P^\pi(d(f(\cdot\,|\theta), f(\cdot\,|\theta^\perp) \leq \epsilon)$, soit en établissant l'espérance a posteriori de $d(f(\cdot\,|\theta), f(\cdot\,|\theta^\perp))$. Dans le cas de la sélection de variables en régression, quand y est conditionnel à un vecteur x de p covariables, la tâche se complique par la nécessité d'intégrer sur la distribution jointe de (x, y) pour obtenir la distance, soit (Exercice 7.31)

Tab. 7.3. Valeurs des paramètres pour différentes divergences de Kullback-Leibler de ϵ dans le cas des distributions Bernoulli, Poisson et normales. (*Source :* Goutis et Robert, 1998.)

ϵ	0	0.01	0.05	0.1	0.25	0.5	1	2	∞
$\mathscr{B}(p)$	0.5	0.57	0.65	0.71	0.81	0.9	0.96	0.99	1
$\mathscr{P}(\lambda)$	1	1.15	1.35	1.52	1.88	2.36	3.15	4.5	∞
$\mathscr{N}(\mu,1)$	0	0.14	0.32	0.45	0.71	1	1.41	2	∞

$$\mathbb{E}_x[d(f(\cdot|x,\theta), f_{\mathscr{A}}(\cdot|x_{\mathscr{A}},\theta^{\perp}))]\,,$$

avec $\mathscr{A} \subset \{1,\ldots,p\}$ et $x_{\mathscr{A}}$ le sous-ensemble de covariables correspondant. Comme la distribution du vecteur de covariables x est souvent inconnue, on l'estime par la moyenne empirique

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_y\left[\log\left(\frac{f(y|x_i,\theta)}{g(y|x_{i\mathscr{A}},\theta^{\perp})}\right)\middle| x_i\right]\,.$$

Outre les difficultés numériques habituelles pour obtenir des approximations d'espérances ou de probabilités a posteriori, nous sommes confrontés à un nouveau problème, plus spécifique à la sélection de variables. Étant donné p covariables potentielles, il y a 2^p (ou 2^p-1) modèles en concurrence. Lorsque p est grand, une exploration complète de tous les modèles est impossible. Heureusement, comme nous le décrivons dans la Note 7.8.3, certaines propriétés de transitivité et d'additivité de la distance Kullback-Leibler permettent d'élaguer plus rapidement l'arbre des sous-modèles : lorsqu'on cherche parmi tous les sous-ensembles $\mathscr{A}$ de covariables tels que

$$d(\mathrm{M}_g, \mathscr{M}_{\mathscr{A}}) = \mathbb{E}_x[d(f(y|x,\alpha), g(y|x_{\mathscr{A}},\alpha^{\perp}))] < \epsilon\,,$$

le sous-modèle avec le cardinal le plus petit, c'est-à-dire celui qui a le plus faible nombre de covariables, on peut évaluer ce cardinal par *pas descendants*–on part du modèle complet et on descend dans l'arbre des sous-modèles en éliminant une covariable à la fois, celle qui est le plus loin de M_g, jusqu'à ce que la distance devienne trop importante–et par *pas montants*–on part du modèle constant et on ajoute une covariable à la fois, le plus proche de M_g, jusqu'à ce que la distance soit plus petite que ϵ–et vérifier a posteriori qu'aucun autre modèle de même cardinal p_0 ne soit plus proche du modèle complet. Cette dernière étape peut toutefois être particulièrement longue, de l'ordre de $\binom{p}{p_0}$ (Exercice 7.30).

Exemple 7.19. (Suite de l'Exemple 7.18) Pour un a priori constant sur les paramètres de régression (α,β), Dupuis et Robert (2001) obtiennent

les résultats présentés en Table 7.4 via cette procédure de sélection de variables (avec $\varrho = 0.9$ lors de l'étalonnage de ϵ). Les trois étapes de la méthode choisissent le même sous-modèle `100111111001`. D'après la liste des variables explicatives données dans l'Exemple 7.18, cela signifie que le sous-modèle sélectionné n'inclut pas les graisses consommées dans la liste des variables explicatives les plus importantes. L'accord entre l'approche fondée sur l'espérance de la distance a posteriori *(colonne 3)* et celle fondée sur la probabilité a posteriori que la distance soit inférieure à ϵ *(colonne 4)* est remarquable. ‖

Bien que cette approche ait l'avantage de s'appuyer sur une fonction de coût pour sélectionner les sous-modèles et d'éliminer le problème des lois a priori impropres, elle n'est pas exempte de défauts. Le premier d'entre eux est l'énorme quantité de calcul nécessaire lorsque, comme c'est le cas en sélection de variables, le nombre de sous-modèles à étudier est grand. Ensuite, la façon de déterminer la borne ϵ n'est pas irréprochable : par exemple, pourquoi une proportion fixe de la distance serait-elle pertinente pour la prise de décision ? Comment doit-elle dépendre du nombre d'observations ? Un autre inconvénient de cette méthode est qu'elle nécessite un modèle complet (ou de référence) et ne marche donc que pour des modèles imbriqués. S'inspirant d'une idée communément utilisée en Économétrie (voir, par exemple, Gouriéroux et Monfort, 1996), Goutis et Robert (1998) proposent d'étendre la méthode à un cadre plus général en créant un *modèle imbriquant*, mais le problème est difficile puisque le modèle imbriquant n'est pas le vrai modèle et n'a donc qu'un intérêt limité pour la prise de décision. En outre, il existe encore de nombreuses manières de définir le modèle imbriquant, qui conduisent à des résultats différents. On peut par exemple considérer les moyennes arithmétique ou géométrique de modèles (Exercice 7.35).

Exemple 7.20. (Suite de l'Exemple 7.1) Étant donné que les modèles de Poisson $\mathscr{P}(\lambda)$ et binomial négatif $\mathscr{NB}(n,p)$ contiennent des termes de la forme

$$\frac{\lambda^y}{y!}\,,$$

avec $\lambda = p/(1-p)$ dans le cas de la binomiale négative, un modèle imbriquant envisageable est

$$f(y|\lambda,m,\alpha) \propto \frac{1}{y!}\lambda^y\, e^{-\alpha\lambda}\left[\frac{m!}{(m-y)!}\,\frac{1}{(1+e^{\lambda})^m}\right]^{1-\alpha} \qquad 0 \le \alpha \le 1\,.$$

On retrouve le modèle de Poisson pour $\alpha = 1$ et la loi binomiale négative pour $\alpha = 0$. Cette densité est en fait la moyenne géométrique des deux densités mais la constante de normalisation, qui dépend de (λ, m, α), est inconnue. On obtient une solution alternative plus abordable en utilisant la moyenne arithmétique, ce qui donne le mélange

Tab. 7.4. Sous-modèles étudiés par la procédure de sélection de variables pour le jeu de données concernant le cancer du sein. Le résultat de chaque étape est présenté en gras, $d(\mathrm{M}_g, \mathscr{M}_{\mathscr{A}})$ représente l'espérance de la divergence de Kullback-Leibler entre le modèle complet et sa projection sur le sous-ensemble de covariables $\mathscr{A}$ et $P(\mathrm{M}_{\mathscr{A}})$ est la probabilité a posteriori que la distance $d(\mathrm{M}_g, \mathscr{M}_{\mathscr{A}})$ soit inférieure à ϵ. (*Source :* Dupuis et Robert, 2001.)

étape	sous-ensemble $\mathscr{A}$	$d(\mathrm{M}_g, \mathscr{M}_{\mathscr{A}})$ ($\times 740$)	$P(\mathrm{M}_{\mathscr{A}})$
1.	101111111111	0.508	0.98
	101111111011	1.146	0.96
	100111111011	1.800	0.94
	100111111001	**2.726**	**0.91**
2.	000000010000	21.78	0.29
	000010010000	16.97	0.45
	100010010000	13.81	0.55
	100010011000	10.61	0.66
	100010011001	7.601	0.75
	100011011001	5.224	0.83
	100111011001	3.736	0.88
	100111111001	**2.726**	**0.91**
3.	111111110000	8.170	0.73
	111111001010	13.72	0.55
	111100111010	8.349	0.73
	110011111010	5.988	0.81
	001111111010	9.215	0.70
	111110011001	4.542	0.85
	111101011001	4.761	0.85
	111011011001	3.91	0.87
	110111011001	3.265	0.89
	101111011001	3.017	0.90
	011111011001	5.895	0.81
	100111111001	**2.726**	**0.91**
	100111011101	3.109	0.899
	100011111101	3.826	0.88
	111011010011	5.284	0.83
	110110110011	6.04	0.80
	101101110011	5.9	0.81
	101011011011	3.576	0.88
	100111011011	2.77	0.91
	101010111011	5.08	0.84
	011001111011	9.346	0.70
	100110011111	4.151	0.87
	100101011111	4.224	0.86
	100011011111	3.787	0.88

$$p\,\mathscr{P}(\lambda) + (1-p)\,\mathscr{N}eg\left(m, \frac{e^\lambda}{1+e^\lambda}\right) \quad 0 \le p \le 1\,.$$

‖

7.6 Adéquation à une famille de lois

Nous refermons ce chapitre par une courte introduction à l'approche bayésienne du concept d'adéquation (traduction de *goodness of fit*), qui est, d'une certaine façon, le problème de choix de modèle le plus difficile. En effet, dans les questions de type *Le modèle $\mathscr{M}_0$ est-il compatible avec x ?* ou *f appartient-elle à la famille $\{f_\theta;\, \theta \in \Theta\}$?*, il n'y a pas d'hypothèse alternative à $\mathscr{M}_0$. Ainsi, dans l'Exemple 7.1, si nous ne considérons que le modèle de Poisson, juger de sa compatibilité avec les données est d'autant plus difficile que, s'il ne l'est *pas*, il n'y a pas alors de modèle défini[66].

Il semble que la difficulté vienne ici du fait que le paradigme bayésien ne puisse se prononcer sur la validité du modèle qu'en "sortant" du modèle, c'est-à-dire en travaillant dans un cadre élargi (*un métamodèle*) dans lequel le modèle considéré n'est qu'un cas particulier. Mais, en réalité, le problème tient plus à la formulation maladroite de la question qu'au paradigme bayésien lui-même. L'incapacité de ce dernier à répondre à un problème aussi mal posé ne signifie en aucune manière que d'autres méthodes apportant une réponse, comme le test du χ^2, soient plus légitimes ! En fait, le paradigme bayésien clarifie le problème en posant comme condition nécessaire la construction préliminaire d'un modèle alternatif et formalise la définition de métamodèle incluant le modèle d'étude.

Une fois l'ambiguïté levée, il y a de nombreuses façons de définir le modèle alternatif $\mathscr{M}_1$, à moins qu'il ne soit contraint par la disponibilité d'informations a priori précises. Le modèle $\mathscr{M}_1$ peut par exemple être un *modèle imbriquant* $\mathscr{M}_0$. Mais comme nous l'avons vu en Section 7.5, il n'y a pas unicité de choix pour un tel modèle. La notion de modèle imbriquant le plus petit (ou le plus naturel) n'existe pas, en dehors de la réponse triviale de $\mathscr{M}_0$ lui-même ! Neyman (1937) définit une extension de la famille exponentielle

$$f_1(x|\theta,\varphi) \propto f(x|\theta) \exp\left\{-\varphi \log \frac{f(x|\theta)}{f(x|\hat\theta(x))}\right\}, \qquad \varphi \ge 0\,,$$

[66] L'approche fréquentiste contourne cette difficulté en ne travaillant que sous l'hypothèse nulle. Par exemple, le test du χ^2 standard s'appuie sur l'approximation du χ^2 qui n'est valable que lorsque le modèle considéré est le "vrai" modèle. *Dans le cas contraire*, la statistique du χ^2 tend vers l'infini mais on ne sait rien de sa distribution pour une taille d'échantillon donnée.

avec $\hat{\theta}(x)$ estimateur du maximum de vraisemblance (en le supposant défini), mais d'autres extensions hiérarchiques sont envisageables. De plus, la représentation de l'hypothèse alternative par les modèles imbriquants est très limitée, puisque dans un problème d'adéquation, elle doit être "*f n'est pas dans $\mathscr{M}_0$*".

On peut lever ces restrictions en utilisant une représentation *non paramétrique* de l'hypothèse alternative. Des techniques standard de Statistique bayésienne non paramétrique sont présentées dans la Note 1.8.2, comme par exemple les lois a priori par processus de Dirichlet et leurs généralisations, mélanges ou ondelettes. Nous étudions à titre d'exemple la représentation *polynomiale orthogonale* de Verdinelli et Wasserman (1992). Voir Castro *et al.* (1999) pour le cas discret (Exercice 7.38).

On peut exprimer le modèle considéré $\mathscr{M}_0 : x \sim f(x|\theta)$, $\theta \in \Theta$, de la façon suivante :

$$\mathscr{M}_0 : x = F^-(u|\theta), \qquad \theta \in \Theta, \qquad u \sim \mathscr{U}([0,1]),$$

avec $F^-(\cdot|\theta)$ inverse généralisé de la fonction de répartition de $f(\cdot|\theta)$ (Exercice 7.39). On peut donc écrire $\mathscr{M}_0$ comme un cas particulier de

$$\mathscr{M}_1 : x = F^-(u|\theta), \qquad \theta \in \Theta, \qquad u \sim g(u|\psi), \qquad \psi \in \mathscr{S},$$

avec $g(\cdot|\psi)$ distribution sur $[0,1]$, dont un cas particulier est la distribution uniforme $g(u|\psi_0) = 1$, et $\mathscr{S}$ est un espace de dimension infinie. Cette reparamétrisation du modèle nous permet de travailler sur les distributions sur $[0,1]$, plutôt que sur un espace général, et ramène notre tâche à un test d'uniformité (conditionnellement à θ).

Il y a de nombreuses possibilités pour le choix de la famille de distributions $g(\cdot|\psi)$ de dimension infinie. Un choix envisageable est la famille de mélanges de densités bêta,

$$g(u|\psi) = \varrho_0 + (1-\varrho_0) \sum_{j=1}^{+\infty} \varrho_j \frac{u^{\alpha_j}(1-u)^{\beta_j}}{K(\alpha_j,\beta_j)},$$

comme dans Petrone et Wasserman (2002) et l'estimation peut alors être réalisée par des techniques à sauts réversibles. Verdinelli et Wasserman (1998) proposent ici d'utiliser les *polynômes de Legendre* sur $[0,1]$,

$$\phi_j(x) = \frac{1}{2^j j!} \frac{d^j}{dx^j}(x^2-1)^j$$

correspondant aux densités

$$g(u|\psi) \propto \exp\left\{ \sum_{j=1}^{+\infty} \psi_j \phi_j(u) \right\}.$$

(Voir Barron, 1988, 1998, et Lenk, 1999, pour plus de détails.) Le modèle nul $\mathscr{M}_0$ correspond alors à $\psi_1 = \ldots = \psi_p = \ldots = 0$.

La distribution a priori sur (θ, ψ) est choisie de telle sorte que θ et ψ soient indépendants, avec un a priori de référence sur θ. Cette hypothèse d'indépendance n'est pas sans conséquence étant donné que θ a le même a priori sous $\mathscr{M}_0$ et $\mathscr{M}_1$, mais n'est pas identifiable sous $\mathscr{M}_1$ (Exercice 7.40). Les ψ_j sont alors modélisés comme des variables aléatoires indépendantes,

$$\psi_j \sim \mathscr{N}(0, \tau_j^2),$$

avec $\tau_j = \tau/2^j$ pour des raisons de cohérence (Barron, 1988), et τ est associé à un a priori propre vague, $\pi(\tau)$.

La distribution a posteriori est alors donnée par

$$\pi(\theta, \psi, \tau | x_1, \cdots, x_n) \propto \prod_{i=1}^{n} f(x_i|\theta) g(u_i|\psi) \pi(\theta)\, \pi(\psi|\tau)\pi(\tau)\,, \tag{7.20}$$

avec $u_i = F(x_i|\theta)$ (Exercice 7.41). Cette expression n'est manifestement pas calculable, ne serait-ce que parce que les u_i dépendent de θ. On peut néanmoins simuler $\pi(\theta, \psi, \tau | x_1, \cdots, x_n)$ au moyen d'un algorithme MCMC, par exemple avec les étapes de Gibbs :

$$\begin{aligned}
\theta|\psi, x_1, \cdots, x_n &\sim \prod_{i=1}^{n} f(x_i|\theta) g(u_i|\psi)\pi(\theta)\,,\\
\psi|\tau, \theta, x_1, \cdots, x_n &\sim \prod_{i=1}^{n} g(u_i|\psi)\pi(\psi|\tau)\,,\\
\tau|\psi &\sim \pi(\psi|\tau)\pi(\tau)\,.
\end{aligned}$$

Il faut cependant des étapes de Metropolis-Hastings supplémentaires pour simuler θ et ψ.

Une fois une approximation de la distribution a posteriori obtenue, Verdinelli et Wasserman (1998) proposent d'utiliser le facteur de Bayes

$$B_{01} = \frac{\displaystyle\int \prod_{i=1}^{n} f(x_i|\theta)\pi(\theta) d\theta}{\displaystyle\int \prod_{i=1}^{n} f(x_i|\theta) g(F(x_i|\theta)|\psi)\pi(\theta, \psi, \tau) d\theta d\psi d\tau}$$

pour décider si l'adéquation à $\mathscr{M}_0$ est suffisante. (Ils montrent de plus que la procédure est *convergente*, que B_{01} tend vers 0 presque sûrement si $\mathscr{M}_0$ n'est pas le bon modèle et vers l'infini en probabilité dans le cas contraire.) Une autre procédure d'évaluation consiste à remarquer que $\mathscr{M}_0$ correspond à $\tau = 0$ et à utiliser un test d'hypothèse standard sur l'échantillon MCMC.

7.7 Exercices

Section 7.1.1

7.1 La *déviance* d'un modèle est simplement la valeur de la log-vraisemblance pour l'estimateur du maximum de vraisemblance (McCullagh et Nelder, 1989). Calculer $\hat{\lambda}$ et $(\hat{m}, \hat{p})$ pour l'estimateur du maximum de vraisemblance de l'Exemple 7.1 et comparer les déviances.

7.2 Dans le cadre de l'Exemple 7.2, montrer qu'un mélange à k composantes peut être représenté par un mélange à $k+1$ composantes soit en annulant le poids d'une des composantes, soit en fixant la moyenne et la variance de la $(k+1)$-ième composante égale à celles d'une des k premières composantes. Quel est le rapport entre cette multiplicité et la propriété de non-identifiabilité des mélanges vue dans la Note 6.6.6 ?

7.3 Pour l'Exemple 7.3, écrire les distributions marginales des $y_i = (y_{i1}, \ldots, y_{i7})$ en intégrant les effets aléatoires. Est-il possible d'obtenir un résultat explicite avec les lois a priori conjuguées ?

Section 7.2.1

7.4 On considère deux modèles $\mathcal{M}_1 : x \sim f_1(x|\theta_1, \gamma)$ et $\mathcal{M}_2 : x \sim f_2(x|\theta_2, \gamma)$ avec une distribution a priori

$$\pi(\theta_1, \theta_2, \gamma) = \pi_1(\theta_1|\gamma)\pi_2(\theta_2|\gamma)\pi_0(\gamma)\,,$$

π_1 et π_2 étant propres. Montrer que, si π_0 est impropre, le facteur de Bayes B_{12}^{π} ne dépend pas de la constante de normalisation de π_0.

7.5 Dans le cadre de l'Exemple 7.4, on suppose que T_t est distribuée selon une loi uniforme $\mathcal{U}_{[0,\bar{T}]}$ et que $\beta_{21} \sim \mathcal{N}(0, \tau^2)$.

a. En intégrant le terme $\beta_{21}T_t$ dans $\mathcal{M}_2$, calculer le modèle marginal de y_{it}.

b. En déduire la distribution a priori sur les paramètres de $\mathcal{M}_1$ si $\mathcal{M}_2$ est le vrai modèle et $(\beta_{20}, b_{2i}, \sigma_2) \sim \pi(\beta_{20}, b_{2i}, \sigma_2)$.

7.6 *(Barbieri *et al.*, 1999) Soit un modèle $f(x|\varphi, \psi)$, $(\varphi, \psi) \in \Phi \times \Psi$, tel qu'il existe $\psi^* \in \overline{\Psi}$ vérifiant

$$\lim_{\psi \to \psi^*} f(x|\varphi, \psi) = f^*(x|\psi^*)\,,$$

c'est-à-dire tel que la distribution limite ne dépende plus de φ.

a. Montrer que cette condition est vérifiée par le *modèle de calibration linéaire*,

$$z_1 \sim \mathcal{N}(\psi, 1), \qquad z_2 \sim \mathcal{N}(\phi\psi, 1)\,,$$

pour $\psi^* = 0$.

b. Si $\pi(\varphi, \psi)$ est un a priori propre avec une masse en ψ^*, montrer que

$$\pi(\varphi|x) = \pi(\varphi|\psi^*)\pi(\psi^*|x) + \pi(\varphi|\psi \neq \psi^*, x) \int_{\psi \neq \psi^*} \pi(\psi|x)\, d\psi\,.$$

c. Si $H_0 : \varphi = \varphi^0$ doit être testée contre $H_1 : \varphi \neq \varphi^0$, montrer que

$$B_{01} = \pi(\psi^*|x) + \frac{m(x|\psi \neq \psi^*)}{\pi(\varphi^0)} \int_{\psi \neq \psi^*} \pi(\psi|x)\, d\psi$$

en supposant que π a aussi une masse en φ^0.

d. En déduire que le facteur de Bayes est fortement influencé par la modélisation a priori sur ψ^*, quel que soit φ^0.

[*Note :* Gleser et Hwang (1987) étudient ces modèles d'un point de vue fréquentiste et montrent qu'un intervalle de confiance de niveau α sur une fonction non bornée de φ a un volume infini avec une probabilité positive.]

Section 7.2.2

7.7 (Berger et Pericchi, 2001) On considère le modèle linéaire normal $\mathscr{M}_2$

$$y = \alpha\mathbf{1} + z_1\beta_1 + z_2\beta_2 + \epsilon, \qquad \epsilon \sim \mathscr{N}_n(0, \sigma^2\mathbf{I}_n)\,,$$

avec $\beta_1 \in \mathbb{R}^k$, $\beta_2 \in \mathbb{R}^p$ et les z_i centrés et orthogonaux, c'est-à-dire tels que $z_1^t z_2 = 0$. Le sous-modèle $\mathscr{M}_1$ correspond à $\beta_2 = 0$.

a. Montrer que, sous les lois a priori

$$\pi_1(\alpha, \beta_1, \sigma) = 1/\sigma \qquad \text{et} \qquad \pi_2(\alpha, \beta_1, \sigma, \beta_2) = h(\beta_2|\sigma)/\sigma\,,$$

avec $h(\beta_2|\sigma)$ suivant une loi de Cauchy $\mathscr{C}_p(0, z_2^t z_2/n\sigma^2)$, le facteur de Bayes B_{12} ne peut pas être calculé explicitement.

b. Pour le modèle $\mathscr{M}_1$: $y = X\beta + \epsilon$, $\epsilon \sim \mathscr{N}_n(0, \sigma^2\mathbf{I}_n)$, le G-prior de Zellner (1986b) est

$$\pi(\sigma) = 1/\sigma\,, \qquad \pi(\beta|\sigma) \propto \exp\{-\beta^t X^t X\beta/2g\sigma^2\}\,.$$

Montrer que dans ce cas, la densité marginale est exprimable analytiquement.

c. Soit $\mathscr{M}_0$ le modèle associé à $\beta = 0$. On note k la dimension de β dans le modèle $\mathscr{M}_1$. Montrer que la limite du facteur de Bayes B_{01} est $(1+g)^{(k-n)/2}$, quand l'estimateur du maximum de vraisemblance $\hat{\beta}$ tend vers l'infini. Conclure sur les avantages du G-prior pour ce problème.

7.8 ***(Suite de l'Exercice 7.6)** Pour le modèle à calibration linéaire,

a. Montrer que l'a priori de Jeffreys est

$$\pi^J(\varphi, \psi) \propto |\psi|\,.$$

b. Montrer que, pour le test de $H_0 : \varphi = \varphi^0$, avec $\pi_0(\psi) \propto 1$, le *facteur de Bayes fractionnaire* avec la fraction $0 < b < 1$ (voir l'équation (5.10)) est

$$B_{01}^F = b^{-1/2} \exp\left\{-\frac{1-b}{2}\,\frac{(z_1 - z_2\varphi^0)^2}{1+\varphi_0^2}\right\}\,.$$

c. Montrer que le *facteur de Bayes arithmétique intrinsèque* (voir l'équation (5.7)) est

$$B_{01}^A = \sqrt{2} \exp\left\{-\frac{1-0.5}{2}\,\frac{(z_1 - z_2\varphi^0)^2}{1+\varphi_0^2}\right\}\,.$$

d. Étudier l'extension à n observations.

7.9 ***(Suite de l'Exercice 7.8)**

a. Montrer que l'a priori de référence est

$$\pi^R(\varphi,\psi) \propto \frac{1}{\sqrt{1+\varphi^2}}\,.$$

b. Montrer que le facteur de Bayes est

$$B_{01}^F = b^{-1/2}\exp\left\{1 - b2\,\frac{[(z_1^2-z_2^2)(1-\varphi^0)^2 - 4z_1z_2\varphi_0]}{1+\varphi_0^2}\right\}$$
$$\frac{I_0(b(z_1^2+z_2^2)/4)}{I_0((z_1^2+z_2^2)/4)}\,,$$

avec I_0 la fonction de Bessel modifiée (Exercice 4.36).

7.10 *Dans le contexte de l'Exemple 7.5,

a. Effectuer le calcul complet de B_{12}^π pour arriver à l'expression finale de l'exemple.

b. Montrer que la limite de B_{12}^π est différente suivant que α/β tende vers 0 lorsque α et β tendent vers 0, ou que α/β^N tende vers $c>0$, avec $x<N$.

7.11 Calculer la distribution marginale de x_1 si $x_1,\ldots,x_n$ est un échantillon d'un mélange normal à deux composantes tel qu'il y ait au moins deux observations dans chaque composante.

Section 7.2.3

7.12 Montrer que, pour la comparaison de deux modèles linéaires $\mathscr{M}_1$ et $\mathscr{M}_2$, avec respectivement k_1 et k_2 régresseurs, et n observations, sous l'a priori $\pi_j(\beta_j) = \sigma_j^{-1-q_j}$ $(j=1,2)$, le facteur de Bayes associé au critère BIC s'écrit

$$B_{12} = (R_2/R_1)^{n/2}\; n^{(k_2-k_1)/2}\,,$$

en notant R_j les sommes des carrés résiduels.

7.13 (Suite de l'Exercice 7.8) Dans le cas du modèle à calibration linéaire, sous l'a priori de Jeffreys, montrer que le critère de Schwarz donne presque le même résultat que le facteur de Bayes fractionnel avec la fraction $b=0$ dans l'exponentielle. Commenter.

Section 7.2.4

7.14 Si $f(\cdot|\theta)$ appartient à une famille exponentielle, montrer que le nombre effectif de paramètres p_D est toujours positif.

7.15 *(Spiegelhalter *et al.*, 1998) Dans le cadre de l'Exemple 7.8,

a. Montrer que, pour le modèle saturé avec les θ_i indépendants de lois a priori constantes, p_D est égal à p et la déviance bayésienne vaut $2p$.

b. Montrer que la déviance bayésienne associée au modèle agrégé, $\theta_i=\theta$ pour tout i, est donnée par (7.8).

c. Montrer que l'équation (7.9) est vraie.

d. On suppose que $\theta_i \sim \mathscr{N}(\mu,\tau^2)$ avec τ connu et $\pi(\mu)=1$. Montrer que

$$p_D = \sum_{i=1}^p \varrho_i + \sum_{i=1}^p \varrho_i(1-\varrho_i)\Big/\sum_{i=1}^p \varrho_i$$

et que la déviance bayésienne est égale à

$$\text{DIC} = \tau^{-2} \sum_{i=1}^{p} \varrho_i(1-\varrho_i)(y_i - \bar{y})^2 + p_D \,,$$

avec $\varrho_i = \sigma_i^2\tau^2/(\sigma_i^2+\tau^2)$ et $\bar{y} = \sum_i \varrho_i y_i / \sum_i \varrho_i$.

7.16 Donner dans le détail l'implémentation MCMC des trois modèles de l'Exemple 7.9. (*Indication :* La simulation peut être traitée par `BUGS`.)

7.17 *(Spiegelhalter *et al.*, 1998) On considère un modèle linéaire général

$$y \sim \mathcal{N}(A\theta_1, \Sigma_1)\,, \qquad \theta_1 \sim \mathcal{N}(B\theta_2, \Sigma_2)\,.$$

a. Montrer que la distribution a posteriori de θ_1 est de la forme $\mathcal{N}(\bar{\theta}_1, \Psi)$ et calculer $\bar{\theta}_1$ et Ψ.

b. Montrer que $\mathbb{E}[D(\theta)|y] = D(\bar{\theta}_1) + \text{tr}(A'\Sigma_1^{-1}A\Psi)$ et en déduire $p_D = \text{tr}(A'\Sigma_1^{-1}A\Psi)$.

c. Élargir au cas θ_2 aléatoire et $\pi(\theta_2) = 1$.

7.18 Montrer que les lois conditionnelles sur les φ_i définies dans l'Exemple 7.9 sont bien compatibles avec une loi jointe et expliciter cette loi jointe.

Section 7.3.1

7.19 L'espérance de

$$\frac{1}{T}\sum_{t=1}^{T} \frac{h(\theta^{(t)})}{f(x|\theta^{(t)})\pi(\theta^{(t)})}\,,$$

avec les $\theta^{(t)}$ distribués selon $\pi(\theta|x)$, est-elle égale à $m(x)$ quelle que soit la densité de probabilité h ?

7.20 *(Chen et Shao, 1997) Soient deux densités, $\pi_1(\theta) = c_1\tilde{\pi}_1(\theta)$ et $\pi_2(\theta) = c_2\tilde{\pi}_2(\theta)$, sur le même espace de paramètres Θ.

a. Si π est une densité sur Θ, donner des conditions suffisantes sur le support de π pour que

$$\varrho = \frac{c_2}{c_1} = \frac{\mathbb{E}^\pi[\tilde{\pi}_1(\theta)/\pi(\theta)]}{\mathbb{E}^\pi[\tilde{\pi}_2(\theta)/\pi(\theta)]}\,.$$

b. Montrer que la variance asymptotique de l'estimateur de

$$\varrho^{US} = \frac{\sum_{i=1}^n \tilde{\pi}_1(\theta_i)/\pi(\theta_i)}{\sum_{i=1}^n \tilde{\pi}_2(\theta_i)/\pi(\theta_i)}\,,$$

avec les θ_i i.i.d. de loi π, est

$$\varrho^2\mathbb{E}^\pi\left\{\frac{\pi_1(\theta)}{\pi(\theta)} - \frac{\pi_2(\theta)}{\pi(\theta)}\right\}^2\,.$$

c. En supposant que

$$\varrho^{-2}\mathbb{E}^\pi[(\varrho^{US}-\varrho)^2] = \frac{1}{n}\mathbb{E}^\pi\left[\frac{\{\pi_1(\theta)-\pi_2(\theta)\}^2}{\pi^2(\theta)}\right] + \text{o}\left(n^{-1}\right)\,,$$

montrer que la meilleure densité d'importance π est

$$\pi_0(\theta) \propto |\pi_1(\theta) - \pi_2(\theta)|\,,$$

si

$$\int |\pi_1(\theta) - \pi_2(\theta)|\, d\theta < \infty\,.$$

[*Note :* Torrie et Valleau (1977) appellent cette méthode l'*échantillonnage par parapluie*.]

Section 7.3.2

7.21 Étant donné deux densités $\pi_1(\theta) = c_1\tilde{\pi}_1(\theta)$ et $\pi_2(\theta) = c_2\tilde{\pi}_2(\theta)$ sur le même espace de paramètres Θ, et h une fonction arbitraire,

a. Exprimer $\mathbb{E}^{\pi_2}[h(\theta)\tilde{\pi}_1(\theta|x)]$ sous la forme d'une intégrale en fonction de π_1 et π_2.

b. En déduire l'égalité (7.13).

7.22 *Chen *et al.* (2000) définissent l'*erreur quadratique moyenne relative*

$$\mathscr{E}(r, \hat{r}) = \frac{\mathbb{E}[\hat{r} - r]}{r}$$

pour évaluer les performances de l'estimateur $\hat{r}$ du rapport constant r.

a. Montrer que, si $n = n_1 + n_2$ et si n_1/n_2 tend vers ϱ lorsque n tend vers l'infini, alors

$$\mathscr{E}(r, B_{12}^S) \simeq \frac{1}{n\varrho(1-\varrho)} \left[\frac{\int \pi_1(\theta)\pi_2(\theta)\{\varrho\pi_1(\theta) + (1-\varrho)\pi_2(\theta)\}h^2(\theta)\, d\theta}{\left(\int \pi_1(\theta)\pi_2(\theta)\, d\theta\right)^2}\right]$$

pour l'estimateur (7.14), en faisant abstraction de la dépendance en x par souci de simplification. (*Indication :* Utiliser la méthode delta.)

b. En déduire que le choix optimal pour h est

$$h^*(\theta) \propto \frac{1}{\varrho\pi_1(\theta) + (1-\varrho)\pi_2(\theta)}\,.$$

7.23 Pour les trois fonctions de lien de l'Exemple 7.10, proposer une structure à variables latentes z qui permette d'identifier y à l'indicatrice $\mathbb{I}_{z \le x^t\beta}$.

Section 7.3.3

7.24 Soit une distribution a posteriori $\pi(\theta_1, \theta_2, \theta_3|x)$ telle qu'on ait accès aux trois distributions conditionnelles complètes $\pi(\theta_1|\theta_2, \theta_3, x)$, ... et $\pi(\theta_3|\theta_1, \theta_2, x)$.

a. Montrer que

$$\begin{aligned}\log m(x) = \log f(x|\hat{\theta}) + \log \pi(\hat{\theta}) - \log \pi(\hat{\theta}_3|\hat{\theta}_1, \hat{\theta}_2, x)\\ - \log \pi(\hat{\theta}_2|\hat{\theta}_1, x) - \log \pi(\hat{\theta}_1|x)\,.\end{aligned}$$

b. Montrer que $\pi(\theta_1|x)$ peut être estimé par

$$\hat{\pi}(\theta_1|x) = \frac{1}{T}\sum_{t=1}^{T} \pi(\theta_1, \theta_2^{(t)}, \theta_3^{(t)}|x)\,,$$

avec $(\theta_1^{(t)}, \theta_2^{(t)}, \theta_3^{(t)})$ simulé par échantillonnage de Gibbs.

c. Montrer que $\pi(\theta_2|\hat{\theta}_1, x)$ peut être estimé par

$$\hat{\pi}(\theta_2|\hat{\theta}_1, x) = \frac{1}{T}\sum_{t=1}^{T} \pi(\theta_2|\hat{\theta}_1, \theta_3^{(t)}, x)$$

avec $(\theta_2^{(t)}, \theta_3^{(t)})$ simulé par échantillonnage de Gibbs selon les distributions conditionnelles $\pi(\theta_2|\hat{\theta}_1, \theta_3^{(t-1)}, x)$ et $\pi(\theta_3|\hat{\theta}_1, \theta_2^{(t)}, x)$, ce qui revient à remplacer θ_1 par $\hat{\theta}_1$.

d. Étendre au cas où on dispose de p densités conditionnelles complètes et évaluer le coût nécessaire en temps de calcul pour cette méthode d'approximation.

Section 7.3.4

7.25 Dans le cadre de l'Exemple 7.16, montrer que les jacobiens des déplacements de naissance et de séparation sont respectivement donnés par

$$(1 - p_{(k+1)(k+1)})^k \qquad \text{et} \qquad p_{jk}/(1 - u_1)\,.$$

Section 7.4

7.26 On revient sur les distributions a priori proposées par Clyde (1999),

a. Montrer que la distribution a posteriori de $(\gamma_1, \ldots, \gamma_J)$ conditionnellement à σ est donnée par (7.18).

b. En déduire que le sous-modèle le plus probable correspond aux régresseurs X_j avec des poids ϱ_j plus grands que 1/2.

7.27 *(George et Foster, 1999) Dans un modèle de régression normale

$$y = \beta_1 x_1 + \ldots + \beta_p x_p + \sigma\epsilon\,, \qquad \epsilon \sim \mathcal{N}(0, I)\,,$$

si γ est l'indice d'un sous-modèle parmi les 2^p sous-modèles possibles, on note q_γ le nombre de covariables correspondant, X_γ la matrice des régresseurs associée, $\hat{\beta}_\gamma$ l'estimateur des moindres carrés et s_γ^2 la somme des carrés $\hat{\beta}_\gamma' X_\gamma' X_\gamma \hat{\beta}_\gamma$.

a. Soient les distributions a priori

$$\beta_\gamma|\sigma, \gamma, c \sim \mathcal{N}_{q_\gamma}\left(0, c\sigma^2\left(X_\gamma' X_\gamma\right)^{-1}\right)\,, \qquad \pi(\gamma|\omega) = \omega^{q_\gamma}(1-\omega)^{p-q_\gamma}\,.$$

Identifier cet a priori à celui de Madigan et Raftery (1995).

b. Montrer que

$$\pi(\gamma|y, \sigma, c, \omega) \propto \exp\left[\frac{c}{2(1+c)}\{s_\gamma^2/\sigma^2 - F(c,\omega)q_\omega\}\right]$$

avec

$$F(c,\omega) = \frac{1+c}{c}\left(2\log\frac{1+w}{w} + \log(1+c)\right)\,.$$

c. En déduire que la distribution a posteriori intégrée $\pi(\gamma|y, \sigma, c, \omega)$ est une fonction croissante de $s_\gamma^2/\sigma^2 - F(c,\omega)q_\omega$.

d. Conclure que, moyennant un choix adéquat de (c, ω), le log-a posteriori peut être équivalent à n'importe quel critère standard de choix de modèle, de AIC (avec $F(c,\omega) = 2$) à BIC (avec $F(c,\omega) = \log n$), en passant par le RIC de Foster et George (1998) (avec $F(c,\omega) = 2\log p$).

Section 7.5

7.28 Montrer que la divergence de Kullback-Leibler entre deux distributions normales $\mathcal{N}(0,1)$ et $\mathcal{N}(\mu,\sigma^2)$ est

$$\log\sigma + \frac{\mu^2+1}{2\sigma^2} - \frac{1}{2}\,.$$

Adapter la formule à la divergence de Kullback-Leibler entre $\mathcal{N}(\mu_0,\sigma_0^2)$ et $\mathcal{N}(\mu,\sigma^2)$ par un changement d'échelle approprié.

7.29 Pour chacune des distributions suivantes, montrer l'égalité correspondante sur la divergence de Kullback-Leibler :

(i) Bernoulli $B(p)$:

$$d(f(\cdot\,|p_0), f(\cdot\,|p)) = p_0 \log\frac{p_0}{p} + (1-p_0)\log\frac{1-p_0}{1-p}\,;$$

(ii) Poisson $\mathcal{P}(\lambda)$:

$$d(f(\cdot\,|\lambda_0), f(\cdot\,|\lambda)) = \lambda - \lambda_0 + \lambda_0 \log\frac{\lambda_0}{\lambda}\,; \quad \text{et}$$

(iii) Normale $\mathcal{N}(\mu,1)$:

$$d(f(\cdot\,|\mu_0), f(\cdot\,|\mu)) = (\mu-\mu_0)^2/2\,.$$

7.30 On considère un problème de sélection de variables avec p covariables.

a. Montrer que le nombre de sous-modèles est $2^p - 1$ si tous les modèles ont un terme constant et $2^p - 2$ sinon.

b. Montrer que le nombre de modèles avec exactement p_0 covariables est $\binom{p}{p_0}$.

c. En utilisant l'approximation de Stirling, montrer que ce nombre est également d'ordre 2^p pour $p_0 = p/2$.

7.31 On considère un problème de choix de modèle où $(x,y) \sim g(x|\alpha)f(y|x,\theta)$.

a. Montrer que, pour la divergence de Kullback-Leibler,

$$d\left(g(\cdot|\alpha)f(\cdot|\cdot,\theta), g(\cdot|\alpha')f(\cdot|\cdot,\theta')\right) = d(g(\cdot|\alpha), g(\cdot|\alpha')) + \mathbb{E}_\alpha\left[d(f(\cdot|x,\theta), f(\cdot|x,\theta')\right]\,,$$

l'espérance étant prise sous $x \sim g(x|\alpha)$.

b. En déduire que, si le sous-modèle pose des contraintes sur θ uniquement, par exemple $\varphi(\theta) = 0$, la projection de (α,θ) est $(\alpha,\theta^\perp)$ si $\theta^\perp$ est la solution de

$$\arg\min_{\theta';\,\varphi(\theta')=0} \mathbb{E}_\alpha\left[d(f(\cdot|x,\theta), f(\cdot|x,\theta')\right]\,.$$

7.32 Dans le cas d'un modèle de régression linéaire normal, $y \sim \mathcal{N}(x'\beta,\sigma^2)$,

a. Montrer que, si z est un sous-vecteur de x, la divergence de Kullback-Leibler entre $\mathcal{N}(x'\beta, \sigma^2)$ et $\mathcal{N}(z'\gamma, \sigma^2)$ est $||x'\beta - z'\gamma||^2/2\sigma^2$, conditionnellement à x.

b. En déduire que la projection $\beta^\perp$ s'écrit $\beta^\perp = (zz')^{-1}zx'\beta$.

7.33 (Suite de l'Exercice 7.32) On suppose que β est distribué selon une loi a priori conjuguée $\mathcal{N}(\beta_0, \Sigma)$. Calculer la distribution a priori induite de $\beta^\perp$. Que se passe-t-il dans le cas d'un a priori constant sur β ?

7.34 Dans le cadre de l'Exemple 7.20, déterminer si la constante de normalisation de la moyenne géométrique des distributions de Poisson et binomiale négative, $f(y|\lambda, m, \alpha)$, est calculable.

7.35 On compare deux modèles $\mathcal{M}_1$ et $\mathcal{M}_2$, de densités $f_1(\cdot|\theta_1)$ et $f_2(\cdot|\theta_2)$ toutes deux issues d'une famille exponentielle.

a. Montrer que la moyenne géométrique

$$f_1(\cdot|\theta_1)^\alpha \, f_2(\cdot|\theta_2)^{1-\alpha}$$

appartient encore à une famille exponentielle.

b. Montrer que, si, pour $(i = 1, 2)$

$$f_i(y|\theta_i) = h_i(y) \exp\{\theta_i \cdot \varphi_i(y) - \psi_i(\theta_i)\}\,,$$

$(\varphi_1(y), \varphi_2(y))$ est une statistique exhaustive pour la moyenne géométrique.

c. En déduire que, si $(\varphi_1(y), \varphi_2(y))$ est de plein rang, la dimension de cette famille (Définition 3.8) est la somme des dimensions de f_1 et f_2.

d. Dans le cas particulier où $\mathcal{M}_1$ est exponentielle $\mathcal{E}xp(\theta_1)$ et $\mathcal{M}_2$ semi-normale $\mathcal{N}^+(0, 1/\theta_2)$, montrer que le modèle de la moyenne géométrique est la distribution normale tronquée

$$\mathcal{N}^+\left(-\frac{\alpha\theta_1}{(1-\alpha)\theta_2}, \frac{1}{(1-\alpha)\theta_2}\right)\,,$$

et calculer sa constante de normalisation.

Section 7.6

7.36 Considérer l'extension d'une famille exponentielle de Neyman (1937) lorsque $f(x|\theta)$ est la densité d'une loi (i) de Poisson $\mathcal{P}(\theta)$, (ii) exponentielle $\mathcal{E}xp(\theta)$ et (iii) normale $\mathcal{N}(\theta, 1)$. Dans les trois cas, déterminer si la constante de normalisation est calculable.

7.37 Étant donné une densité

$$f(y|\theta) = h(y) \exp\{\theta \cdot \varphi(y) - \psi(\theta)\}$$

d'une famille exponentielle de dimension d (Définition 3.8), montrer que son extension de Neyman appartient encore à une famille exponentielle de dimension $d + 1$.

7.38 *(Castro *et al.*, 1999) On considère un modèle multinomial

$$\mathbf{r} = (r_0, \ldots, r_k) \sim \mathcal{M}_{k+1}(n; \alpha_0, \ldots, \alpha_k)\,,$$

avec $\alpha = (\alpha_1, \ldots, \alpha_k)$.

a. En notant $(0 \leq b \leq 1)$

$$q_2(\mathbf{r};b) = \frac{\int f(\mathbf{r}|\alpha)\pi_2(\alpha)\,d\alpha}{\int f^b(\mathbf{r}|\alpha)\pi_2(\alpha)\,d\alpha}\,,$$

montrer que, sous l'a priori impropre $\pi_2(\alpha) = 1/\alpha_1 \ldots \alpha_k$,

$$q_2(\mathbf{r};b) = \frac{\Gamma(bn)}{\Gamma(n)} \prod_{j=0}^{k} \frac{\Gamma(r_j)}{\Gamma(br_j)}\,,$$

si tous les r_j sont positifs. (Si l'un des r_j est nul, l'a posteriori n'est pas défini.)

b. Si la contrainte sur les α_j est

$$\alpha_j = \binom{k}{j} \mu^j (1-\mu)^{k-j}\,, \qquad 0 < \mu < 1\,,$$

c'est-à-dire si on veut tester que le modèle sous-jacent est vraiment binomial, montrer que, sous l'a priori $\pi_1(\mu) = 1/\mu(1-\mu)$,

$$\begin{aligned} q_1(\mathbf{r};b) &= \frac{\int f(\mathbf{r}|\alpha(\mu))\pi_1(\mu)\,d\mu}{\int f^b(\mathbf{r}|\alpha(\mu))\pi_1(\mu)\,d\mu} \\ &= \frac{B(r, kn - s_r)}{B(br, b(kn - s_r))} \left[\prod_{j=0}^{k} \binom{k}{j}^{r_j}\right]^{1-b}\,, \end{aligned}$$

avec $s_r = r_1 + \ldots + kr_k$ et $B(a,b)$ constante de normalisation de la loi $\mathscr{B}e(a,b)$ (voir Annexe A).

c. Montrer que le facteur de Bayes fractionnel associé à la contrainte en b. est $B_{12}^F = q_1(\mathbf{r};b)/q_2(\mathbf{r};b)$.

d. Appliquer b. aux données du Tableau 7.5.

Tab. 7.5. Nombre de femmes dans une file d'attente de dix personnes dans le métro de Londres (*Source :* Hoaglin *et al.*, 1996.)

Nombre de femmes	0	1	2	3	4	5	6	7	8	9	10
Occurrences	1	3	4	23	25	19	18	5	1	1	0

e. Si la contrainte sur les α_j prend la forme d'un modèle de Poisson, $\alpha_j = e^{-\lambda}\lambda^j/j!$ $(j = 0, \ldots, k)$, montrer que, sous l'a priori $\pi_1(\lambda) = \lambda^{-t}$,

$$\begin{aligned} q_1(\mathbf{r};b) &= \frac{\int f(\mathbf{r}|\alpha(\lambda))\pi_1(\lambda)\,d\lambda}{\int f^b(\mathbf{r}|\alpha(\lambda))\pi_1(\lambda)\,d\lambda} \\ &= \frac{\Gamma(s_r - t + 1)b^{bs_r - t + 1} n^{s_r(b-1)}}{\Gamma(bs_r - t + 1)} \prod_{j=0}^{k} [j!]^{(b-1)r_j}\,, \end{aligned}$$

en définissant s_r comme en b.

f. Montrer que, sous les mêmes hypothèses que dans e., les facteurs de Bayes intrinsèques ne sont pas constructibles, à moins que les cellules ne soient groupées pour former des r_j positifs.

g. Montrer que, pour un modèle continu, cette stratégie est l'équivalent bayésien du test du χ^2 et qu'elle souffre par conséquent du même problème, à savoir le côté arbitraire du regroupement des observations en k cellules.

7.39 Soit F une fonction de répartition dans $\mathbb{R}$. L'*inverse généralisée* de F est définie par

$$F^-(u) = \inf\{x; F(x) \geq u\}$$

a. Montrer que dans le cas $u \sim \mathscr{U}([0,1])$, $F^-(u) \sim F$.

b. En déduire une technique de simulation pour les distributions de Cauchy et exponentielle.

c. Comment généraliser ce résultat pour une distribution multidimensionnelle ?

7.40 Dans le contexte de l'article de Verdinelli et Wasserman (1998), montrer que le paramètre θ n'est pas identifiable sous le modèle alternatif $\mathscr{M}_1$. (*Indication :* Montrer que, pour toute fonction de répartition $F(x)$ et pour tout θ, il existe ψ tel que $F_\theta^- \circ G_\psi = F^-$.)

7.41 (Verdinelli et Wasserman, 1998) Démontrer l'égalité (7.20) en établissant que, sous le modèle $\mathscr{M}_1$,

$$\begin{aligned} x \sim h(x|\theta,\psi) &= g(F(x|\theta)|\psi)\frac{dF(x|\theta)}{dx} \\ &= g(F(x|\theta)|\psi) f(x|\theta)\,. \end{aligned}$$

Note 7.8.1

7.42 Prouver l'égalité (7.21) en montrant que

$$\int\int \frac{d}{d\lambda} \log \tilde{\pi}(\theta|\lambda)\pi(\theta|\lambda)\,d\lambda\,d\theta = -\int_{\lambda_1}^{\lambda_2} \frac{d}{d\lambda} c(\lambda)\,d\lambda\,.$$

7.43 ***(Suite de l'Exercice 7.42)** Montrer que la généralisation de (7.21) au cas multidimensionnel s'écrit

$$\log(c(\lambda_2)/c(\lambda_1)) = \int_0^1 \mathbb{E}_{\lambda(t)}\left[\sum_{j=1}^k \frac{d\lambda_j(t)}{dt}\frac{\partial}{\partial\lambda_j}\log\tilde{\pi}(\theta|\lambda)\right] dt\,,$$

avec $\lambda(t)$ fonction continue de $[0,1]$ dans Λ telle que $\lambda(0) = \lambda_1$ et $\lambda(1) = \lambda_2$. En déduire l'échantillonneur par chemin correspondant. (*Indication :* Voir Gelman et Meng, 1998, pour une solution détaillée.)

Note 7.8.2

7.44 Dans le cadre de l'Exemple 7.21, donner les étapes de sauts réversibles qui correspondent aux déplacements de naissance et de mort.

7.8 Notes

7.8.1 Échantillonnage par chemin

Gelman et Meng (1998) généralisent l'échantillonnage par passerelle à l'*échantillonnage par chemin* en considérant le cas particulier où les deux lois a posteriori dépendent de la même manière d'hyperparamètres, λ_1 et λ_2,

$$\pi_1(\theta|x) = \pi(\theta|\lambda_1) = \tilde{\pi}(\theta|\lambda_1)/c(\lambda_1)\,,$$
$$\pi_2(\theta|x) = \pi(\theta|\lambda_2) = \tilde{\pi}(\theta|\lambda_2)/c(\lambda_2)\,.$$

Si les hyperparamètres sont des réels tels que $\lambda_1 < \lambda_2$, on a, pour toute densité π_0 de support $[\lambda_1, \lambda_2]$,

$$\log(c(\lambda_2)/c(\lambda_1)) = \mathbb{E}\left[\frac{1}{\pi_0(\lambda)}\frac{d}{d\lambda}\log\tilde{\pi}(\theta|\lambda)\right]\,, \tag{7.21}$$

en intégrant sur la densité $\pi(\theta|\lambda)\pi_0(\lambda)$ (Exercice 7.42).

L'estimateur correspondant du logarithme du facteur de Bayes en échantillonnage par chemin est alors

$$B_{12}^{PS} = \frac{1}{n}\sum_{i=1}^{n}\frac{\dfrac{d}{d\lambda}\log\tilde{\pi}(\theta_i|\lambda_i)}{\pi_0(\lambda_i)}\,,$$

avec le choix formellement optimal pour π_0,

$$\pi_0(\lambda) \propto \sqrt{\mathbb{E}\left[\left(\frac{d}{d\lambda}\log\tilde{\pi}(\theta|\lambda)\right)^2\Big|\lambda\right]}\,.$$

(Voir l'Exercice 7.43 pour une extension au cas multidimensionnel.)

7.8.2 Processus de saut

On considère ici une technique analogue à celle par sauts réversibles largement abordée dans la littérature (voir, par exemple, Ripley, 1987, Grenander et Miller, 1994, ou Phillips et Smith, 1996). Elle est en théorie applicable (Cappé *et al.*, 2003) dans un cadre très général mais n'a pour l'instant été utilisée que dans des problèmes de sélection de variables. C'est le cas notamment de la solution de Stephens (2000) au problème de l'Exemple 7.2.

Cette méthode s'appuie sur les *processus de sauts* : on simule un processus de saut à temps continu sur l'espace (7.2), c'est-à-dire un processus stochastique $(\xi_t)_{t\in\mathbb{R}^+}$ qui reste dans un état donné (i,θ_i) pour une durée suivant une loi exponentielle $T \sim \mathcal{E}xp(\varphi_i(\theta_i))$, φ étant l'*intensité* du processus, puis saute vers un nouvel état j avec une probabilité $q_{i\to j}$ et simule θ_j selon une densité $h_{i\to j}(\theta_j|\theta_i)$. Ensuite, comme en temps discret (voir (6.17)), si les paramètres du processus, φ, q et h, satisfont une *condition d'équilibre ponctuel*

$$\pi(i,\theta_i)\varphi_i(\theta_i)q_{i\to j}h_{i\to j}(\theta_j|\theta_i) = \pi(j,\theta_j)\varphi_j(\theta_j)q_{j\to i}h_{j\to i}(\theta_i|\theta_j)\,,$$

alors $\pi(i,\theta_i)$ est une distribution stationnaire de ce processus markovien. Par exemple, si $h_{i\to j}(\theta_j|\theta_i) = g_j(\theta_j)$ et $q_{i\to j} = 1/k$, avec k nombre d'états, la condition d'équilibre est

$$\pi(i,\theta_i)\varphi_i(\theta_i)g_j(\theta_j) = \pi(j,\theta_j)\varphi_j(\theta_j)g_i(\theta_i)$$

et l'intensité est $\varphi_i(\theta_i) \propto g_i(\theta_i)/\pi(i,\theta_i)$. (L'intensité $\varphi_i(\theta_i)$ est l'inverse de la durée moyenne en (i,θ_i), qui est logiquement proportionnelle à $\pi(i,\theta_i)$.)

Dans le cas particulier où les déplacements sont limités aux états adjacents, c'est-à-dire lorsque $q_{i\to i+1} + q_{i\to i-1} = 1$ (avec les modifications qui conviennent aux extrêmités), le processus est appelé *processus de saut à naissances et à morts*. On écrit alors souvent $\varphi_i(\theta_i) = \beta(\theta_i) + \delta(\theta_i)$, avec $\beta(\theta_i)$ *taux de naissance* et $\delta(\theta_i)$ *taux de mort*, et on s'affranchit du paramètre $q_{i\to j}$. Le processus reste dans l'état (i,θ_i) pendant un temps exponentiel $\mathcal{E}xp[\beta(\theta_i) + \delta(\theta_i)]$, puis se déplace soit vers l'état $(i+1,\theta_{i+1})$ avec probabilité $\beta(\theta_i)/(\beta(\theta_i) + \delta(\theta_i))$, θ_{i+1} étant simulé selon $K_i^+(\theta_{i+1}|\theta_i)$, soit vers l'état $(i-1,\theta_{i-1})$, θ_{i-1} étant simulé selon $K_i^-(\theta_{i-1}|\theta_i)$.

Exemple 7.21. (Suite de l'Exemple 7.2) Pour l'exemple du mélange, les étiquettes des états i correspondent aux nombres de composantes, la naissance à l'ajout d'une composante et la mort à la suppression d'une composante. Alors $\theta_i = (p_{1i},\ldots,p_{ii},\mu_{1i},\ldots,\mu_{ii},\sigma_{1i},\ldots,\sigma_{ii})$. Dans son implémentation de l'algorithme à sauts de naissances et de morts, Stephens (2000) simule de nouvelles composantes selon la distribution a priori (dans laquelle toutes les composantes sont i.i.d.) et choisit un taux de naissance fixe $\beta(\theta_i) = b$. La condition d'équilibre devient alors

$$(i+1)\beta(\theta_{i+1})L[(i+1,\theta_{i+1})|x_1,\ldots,x_n]\pi(i+1) = bL[(i,\theta_i)|x_1,\ldots,x_n]\pi(i)\,,$$

en notant $L(\theta|x_1,\ldots,x_n)$ la vraisemblance. (Le coefficient $(i+1)$ tient au fait qu'il y a $(i+1)$ composantes et donc $(i+1)$ suppressions possibles.)

En notant $\theta_i/(p_{\ell i},\mu_{\ell i},\sigma_{\ell i})$ le paramètre du modèle à $(i-1)$ composantes où la composante $(p_{\ell i},\mu_{\ell i},\sigma_{\ell i})$ a été supprimée, l'algorithme de naissance et de mort est le suivant :

ALGORITHME 7.1. **Sauts de naissance et de mort**
Dans l'état (i,θ_i),

1. Calculer les taux de mort de chaque composante $(\ell = 1,\ldots,i)$

$$\beta_\ell(\theta_i) = \frac{L[(i-1,\theta_i/(p_{\ell i},\mu_{\ell i},\sigma_{\ell i}))|x_1,\ldots,x_n]}{L[(i,\theta_i)|x_1,\ldots,x_n]}$$

et prendre $\beta(\theta_i) = \sum_{\ell=1}^{i}\beta_\ell(\theta_i)$

2. Simuler le temps de saut $T \sim \mathcal{E}xp(\beta(\theta_i) + b)$
3. À l'instant T, supprimer

$$(p_{\ell i},\mu_{\ell i},\sigma_{\ell i})|x_1,\ldots,x_n$$

avec probabilité

$$\frac{\beta_\ell(\theta_i)}{\beta(\theta_i)+b}$$

Sinon, créer

$$(p_{(i+1)(i+1)},\mu_{(i+1)(i+1)},\sigma_{(i+1)(i+1)})$$

suivant la distribution a priori.

Remarquons que, dans l'étape 3, le nouveau poids est simulé selon la distribution marginale a priori de $p_{(i+1)(i+1)}$, qui est une distribution $\mathscr{B}e(i,1)$ si l'a priori sur $(p_{1(i+1)},\ldots,p_{(i+1)(i+1)})$ est Dirichlet $\mathscr{D}_{i+1}(1,\ldots,1)$. ‖

On pourra consulter Cappé *et al.* (2003) pour une analyse plus approfondie des liens entre l'algorithme par sauts réversibles et l'algorithme par processus de saut, leur conclusion étant que les deux méthodes diffèrent très peu.

7.8.3 Sélection de variables dans le cas de modèles linéaires généralisés

Nous présentons maintenant de façon plus détaillée la technique de sélection de variables introduite en Section 7.5. Soit, donc, une famille exponentielle générale $(i=1,\ldots,n)$

$$y_i|\theta_i \sim \exp\left[\varphi_i\left\{\theta_i y_i - \psi(\theta_i)\right\} + c(\varphi_i, y_i)\right]$$

avec une structure de *modèle linéaire généralisé* (McCullagh et Nelder, 1989) qui impose une relation entre la moyenne et le vecteur des covariables,

$$g(\psi'(\theta_i)) = x_i^t\beta\,.$$

Dans ce cadre, la divergence de Kullback-Leibler est calculable analytiquement puisque

$$d(f(\cdot\,|\theta), f(\cdot\,|\theta_0)) = \sum_{i=1}^{n} \varphi_i\,\{\psi'(\theta_i)(\theta_i-\theta_i^0) - \psi(\theta_i) + \psi(\theta_i^0)\}$$

et les équations de projection $(j=1,\ldots,p)$

$$\sum_{i=1}^{n} \varphi_i\,\psi'(\theta_i)\,\frac{\partial\theta_i^0}{\partial\beta_j} = \sum_{i=1}^{n} \varphi_i\,\psi'(\theta_i^0)\,\frac{\partial\theta_i^0}{\partial\beta_j}, \tag{7.22}$$

sont équivalentes au système des équations de vraisemblance, ce qui rend leur résolution plus facile.

Pour un modèle logit,

$$P(y_i=1|x_i,\alpha) = 1 - P(y_i=0|x_i,\alpha) = \frac{\exp(\alpha^t x_i)}{1+\exp(\alpha^t x_i)}\,,$$

la projection $\alpha^\perp$ de α sur les covariables z_i (vecteur inclus dans les x_i) est, par exemple, associé à β solution de

$$\sum_{i=1}^{n} \frac{\exp\beta^t z_i}{1+\exp\beta^t z_i} z_i = \sum_{i=1}^{n} \frac{\exp\alpha^t x_i}{1+\exp\alpha^t x_i} z_i\,,$$

ce qui donne effectivement une équivalence formelle avec les équations du maximum de vraisemblance

$$\sum_{i=1}^{n} \frac{\exp\beta^t z_i}{1+\exp\beta^t z_i} z_i = \sum_{i=1}^{n} y_i z_i\,.$$

Une conséquence de (7.22) est que les projections de Kullback-Leibler sont transitives dans la mesure où, si ω est un vecteur inclus dans z, lui-même inclus dans x, on a

$$\sum_{i=1}^{n} \frac{\exp \gamma^t \omega_i}{1+\exp \gamma^t \omega_i}\omega_i = \sum_{i=1}^{n} \frac{\exp \beta^t z_i}{1+\exp \beta^t z_i}\omega_i$$
$$= \sum_{i=1}^{n} \frac{\exp \alpha^t x_i}{1+\exp \alpha^t x_i}\omega_i$$

pour l'exemple du logit. En d'autres termes, cela signifie que la projection γ de la projection β de α est la projection de α sur un sous-espace plus petit, une version orientée "choix de modèle" du *théorème de la double projection*. Une autre propriété remarquable est l'additivité des distances entre ces projections :

$$d(f(\cdot\,|\alpha), f(\cdot\,|\gamma)) = d(f(\cdot\,|\alpha), f(\cdot\,|\beta)) + d(f(\cdot\,|\beta), f(\cdot\,|\gamma))\,.$$

Par rapport au schéma général de sélection de variables présenté en Section 7.5, cela veut dire que, une fois qu'un sous-modèle a été rejeté parce qu'il a été considéré comme trop loin du modèle entier, tous ses sous-modèles seront également rejetés. Voir Dupuis et Robert (2001) pour plus de détails.

8

Admissibilité et classes complètes

> "You can turn the worse that comes to your advantage if you only think, his father has always said, and certainly Abell Cauthon was the best horse trader in the Two Rivers (...) All because he thought about things from every side that there was."
>
> **Robert Jordan**, *The Dragon Reborn.*

8.1 Introduction

Nous avons souligné à plusieurs reprises au cours des Chapitres 1 à 3 l'intérêt des estimateurs de Bayes dans la recherche fréquentiste d'optimalité et en particulier à l'égard de l'admissibilité. Nous y revenons à présent en détail. Dans la Section 8.2, nous étudions les performances des estimateurs de Bayes et de Bayes généralisés en termes d'admissibilité. Puis la Section 8.3 établit un lien entre l'admissibilité d'un estimateur et une suite de distributions a priori grâce à la condition suffisante de Stein. La notion de *classe complète* décrite en Section 8.4 est également fondamentale, car elle permet d'obtenir une caractérisation des estimateurs admissibles ou, au moins, une réduction substantielle de la classe des estimateurs acceptables. Nous présentons des cas où l'ensemble des estimateurs de Bayes constitue une classe complète et d'autres situations dans lesquelles il est nécessaire de considérer les estimateurs de Bayes généralisés. Enfin dans la Section 8.5, nous exposons une méthode introduite par Brown (1971) et développée par Hwang (1982b), qui donne des conditions nécessaires d'admissibilité dans un cadre

plus général, mais non bayésien. Pour une analyse plus technique de ces sujets, on pourra consulter la revue de Rukhin (1995).

8.2 Admissibilité des estimateurs de Bayes

8.2.1 Caractérisations générales

Rappelons les deux résultats suivants sur l'admissibilité des estimateurs (propres) de Bayes, vus dans le Chapitre 2 (Propositions 2.34 et 2.35) :

Proposition 8.1. *Si un estimateur de Bayes est unique, il est admissible.*

Proposition 8.2. *Lorsque la fonction de risque est continue en θ pour tout estimateur δ, si π est équivalente à la mesure de Lebesgue sur Θ, c'est-à-dire si elle est absolument continue de densité positive sur Θ, un estimateur de Bayes associé à π est admissible.*

En revanche, si le support de π n'est pas l'espace entier, il est possible qu'un estimateur de Bayes associé soit inadmissible. De même, les estimateurs de Bayes sont souvent inadmissibles lorsque le risque de Bayes est infini.

Exemple 8.3. On considère une loi normale $x \sim \mathcal{N}(\theta, 1)$ avec un a priori conjugué $\theta \sim \mathcal{N}(0, \sigma^2)$. La distribution a posteriori est alors $\mathcal{N}(\frac{\sigma^2}{\sigma^2+1}x, \frac{\sigma^2}{\sigma^2+1})$ et l'estimateur de Bayes pour la fonction de coût quadratique est

$$\delta^\pi(x) = \frac{\sigma^2}{\sigma^2+1}x\,,$$

qui est admissible, comme le montre le Corollaire 8.14 ci-dessous. À l'inverse, si on change le coût quadratique en

$$\mathrm{L}_\alpha(\theta, \delta) = e^{\theta^2/2\alpha}(\theta - \delta)^2,$$

l'estimateur de Bayes correspondant est inadmissible pour α suffisamment petit. L'estimateur de Bayes généralisé associé à L_α est en fait

$$\delta^\pi_\alpha(x) = \frac{\int_{-\infty}^{\infty} \theta e^{\theta^2/2\alpha} e^{-(\theta-\delta^\pi(x))^2(\sigma^2+1)/2\sigma^2} d\theta}{\int_{-\infty}^{\infty} e^{\theta^2/2\alpha} e^{-(\theta-\delta^\pi(x))^2(\sigma^2+1)/2\sigma^2} d\theta},$$

à condition que les deux intégrales soient finies. Dans la mesure où

$$\exp\left\{\frac{\theta^2}{2\alpha} - (\theta - \delta^\pi(x))^2\frac{\sigma^2+1}{2\sigma^2}\right\}$$
$$= \exp\left\{-\frac{\theta^2}{2}(\frac{\sigma^2+1}{\sigma^2} - \frac{1}{\alpha}) + \delta^\pi(x)\theta\frac{\sigma^2+1}{\sigma^2} - \delta^\pi(x)^2\frac{\sigma^2+1}{2\sigma^2}\right\},$$

δ^π_α est défini pour $\alpha > \frac{\sigma^2}{\sigma^2+1}$ et

$$\begin{aligned}\delta^\pi_\alpha(x) &= \frac{\sigma^2+1}{\sigma^2}\left(\frac{\sigma^2+1}{\sigma^2} - \alpha^{-1}\right)^{-1}\delta^\pi(x)\\ &= \frac{\alpha}{\alpha - \frac{\sigma^2}{\sigma^2+1}}\delta^\pi(x).\end{aligned}$$

Le risque de Bayes correspondant est

$$r(\pi) = \int_{-\infty}^{+\infty} e^{\theta^2/2\alpha} e^{-\theta^2/2\sigma^2} d\theta,$$

et est donc infini pour $\alpha \leq \sigma^2$. De plus, puisque

$$\begin{aligned}\frac{\alpha}{\alpha - \frac{\sigma^2}{\sigma^2+1}}\,\delta^\pi(x) &= \frac{\alpha}{\alpha - \frac{\sigma^2}{\sigma^2+1}}\,\frac{\sigma^2}{\sigma^2+1}x\\ &= \frac{\alpha}{\alpha\frac{\sigma^2+1}{\sigma^2} - 1}x,\end{aligned}$$

l'estimateur de Bayes $\delta^\pi_\alpha(x)$ est de la forme cx avec $c > 1$ lorsque

$$\alpha > \alpha\frac{\sigma^2+1}{\sigma^2} - 1,$$

c'est-à-dire quand $\alpha < \sigma^2$. Et, dans ce cas,

$$\begin{aligned}R(\theta, \delta^\pi_\alpha) &= \mathbb{E}_\theta[(cx-\theta)^2]e^{\theta^2/2\alpha}\\ &= \{(c-1)^2\theta^2 + c^2\}e^{\theta^2/2\alpha} > e^{\theta^2/2\alpha}\end{aligned}$$

implique que δ^π_α est inadmissible, puisqu'il est dominé par $\delta_0(x) = x$, de risque égal à 1. Mais δ_0 est également un estimateur de Bayes formel sous L_α quand $\alpha < \sigma^2$, puisque le risque de Bayes est alors infini. Il est intéressant de remarquer que le cas limite $\alpha = \sigma^2$ correspond à l'estimateur admissible $\delta^\pi_{\sigma^2}(x) = x$ avec un risque de Bayes infini. ‖

Exemple 8.4. Soit $y \sim \sigma^2\chi^2_p$. La distribution a priori conjuguée de σ^2 est la distribution gamma inverse $\mathcal{IG}(\nu/2, \alpha/2)$ (voir le Chapitre 3) et $\pi(\sigma^2|y)$ est la distribution $\mathcal{IG}((\nu+p)/2, (\alpha+y)/2)$, ce qui donne l'espérance a posteriori suivante :

$$\delta^\pi_{\nu,\alpha}(y) = \mathbb{E}^\pi[\sigma^2|y] = \frac{\alpha+y}{\nu+p-2}.$$

Dans le cas particulier $\nu = 2$, $\delta^\pi(y) = (y/p) + (\alpha/p)$. Puisque y/p est un estimateur non biaisé de σ^2, les estimateurs $\delta^\pi_{2,\alpha}$ ne sont pas admissibles sous l'erreur quadratique (puisque $\alpha > 0$). Ce résultat est également vrai pour $\nu < 2$. On vérifie facilement que le risque de Bayes de δ^π est infini dans ce cas (voir Lehmann, 1983, p. 270). ‖

Exemple 8.5. Les estimateurs constants $\delta_0(x) = \theta_0$ sont les estimateurs de Bayes correspondant à une masse de Dirac a priori en θ_0 et sont presque toujours admissibles sous des erreurs quadratiques. En fait,

$$\mathbb{E}_{\theta_0}(\delta(x) - \theta_0)^2 = (\mathbb{E}_{\theta_0}[\delta(x)] - \theta_0)^2 + \text{var}_{\theta_0}(\delta(x)) = 0$$

implique $\text{var}_{\theta_0}(\delta(x)) = 0$ et donc $\delta(x) = \theta_0$ uniformément, à moins que la distribution ne soit dégénérée en θ_0 (voir l'Exercice 8.4). ‖

La Proposition 8.2 se transpose au cas discret (la démonstration est directe et laissée à titre d'exercice).

Proposition 8.6. *Si Θ est un ensemble discret et si $\pi(\theta) > 0$ pour tout $\theta \in \Theta$, alors un estimateur de Bayes associé à π est admissible.*

8.2.2 Conditions aux limites

Nous avons vu en Section 3.3 que, si la distribution de x appartient à une famille exponentielle

$$f(x|\theta) = h(x)e^{\theta.T(x)-\psi(\theta)},$$

les distributions conjuguées sont aussi membres de familles exponentielles et l'espérance a posteriori de la moyenne de $T(x)$ est affine en $T(x)$, ce qui signifie

$$\mathbb{E}^{\pi}[\nabla\psi(\theta)|x] = \frac{T(x)+t_0}{\lambda+1} = \frac{1}{\lambda+1}T(x) + \frac{\gamma_0\lambda}{\lambda+1}, \tag{8.1}$$

avec

$$\pi(\theta|t_0,\lambda) = e^{\theta.t_0-\lambda\psi(\theta)}$$

et $\gamma_0 = t_0/\lambda$. Dans le cas où $\theta \in \mathbb{R}$ et l'espace naturel des paramètres est $N = [\underline{\theta}, \bar{\theta}]$, Karlin (1958) donne une condition suffisante d'admissibilité pour ces estimateurs de la moyenne (voir aussi les Exercices 8.1 et 8.2).

Théorème 8.7. *Si $\lambda > 0$, une condition suffisante pour que l'estimateur* (8.1) *soit admissible sous l'erreur quadratique est que, pour tout $\underline{\theta} < \theta_0 < \bar{\theta}$,*

$$\int_{\theta_0}^{\bar{\theta}} e^{-\gamma_0\lambda\theta+\lambda\psi(\theta)}\,d\theta = \int_{\underline{\theta}}^{\theta_0} e^{-\gamma_0\lambda\theta+\lambda\psi(\theta)}\,d\theta = +\infty.$$

Ce théorème est une conséquence de l'*inégalité de Cramér-Rao* (Lehmann et Casella, 1998). Il s'agit également d'un corollaire de la condition nécessaire et suffisante de Stein (Section 8.3.3). Berger (1982a) considère la réciproque du Théorème 8.7 : il montre que, moyennant quelques hypothèses supplémentaires, cette condition est aussi nécessaire (voir l'Exercice 8.12).

Exemple 8.8. (Suite de l'Exemple 8.4) La paramétrisation naturelle de la distribution du khi deux est

$$\theta = \frac{1}{\sigma^2}, \quad T(y) = -\frac{1}{2}y, \quad \psi(\theta) = -\frac{p}{2}\log(\theta),$$

et

$$\int_0^c e^{-\gamma_0\lambda\theta}\theta^{-\lambda p/2}\,d\theta$$

est infinie si $\lambda p \geq 2$. De même,

$$\int_c^{+\infty} e^{-\gamma_0\lambda\theta}\theta^{-\lambda p/2}\,d\theta = +\infty$$

si $\gamma_0\lambda < 0$ ou $\gamma_0\lambda = 0$ et $\lambda p \leq 2$. Par conséquent, l'estimateur de Bayes

$$\delta^\pi(y) = \frac{\gamma_0\lambda}{1+\lambda} - \frac{1}{1+\lambda}\frac{y}{2}$$

est admissible si $\gamma_0 = 0$ et $\lambda = 2/p$ ou $\gamma_0 < 0$ et $\lambda \geq 2/p$; ces conditions suggèrent les estimateurs

$$\varphi_1(y) = \frac{p}{p+2}\left(\frac{-y}{2}\right) \quad \text{et} \quad \varphi_2(y) = \frac{\gamma_0\lambda}{1+\lambda} + \frac{1}{1+\lambda}\left(\frac{-y}{2}\right),$$

pour $\mathbb{E}_\sigma(-y/2) = -\frac{p}{2}\sigma^2$, et donc les estimateurs de Bayes admissibles suivants pour σ^2 :

$$\delta_1(y) = \frac{y}{p+2} \quad \text{et} \quad \delta_2(y) = ay + b, \quad b > 0,\ 0 \leq a \leq \frac{1}{p+2}. \qquad \|$$

Exemple 8.9. Soit $x \sim \mathscr{B}(n,p)$. La paramétrisation naturelle est donnée par $\theta = n\log(p/q)$ puisque

$$f(x|\theta) = \binom{n}{x} e^{(x/n)\theta}\left(1+e^{\theta/n}\right)^{-n}.$$

Alors les deux intégrales

$$\int_{-\infty}^{\theta_0} e^{-\gamma_0\lambda\theta}\left(1+e^{\theta/n}\right)^{\lambda n} d\theta \quad \text{et} \quad \int_{\theta_0}^{+\infty} e^{-\gamma_0\lambda\theta}\left(1+e^{\theta/n}\right)^{\lambda n} d\theta$$

ne peuvent diverger simultanément si $\lambda < 0$. Considérons le cas $\lambda > 0$. La seconde intégrale diverge en $+\infty$ si $\lambda(1-\gamma_0) > 0$, c'est-à-dire si $\gamma_0 < 1$. Et la première intégrale diverge en $-\infty$ si $\gamma_0\lambda \geq 0$. On obtient alors une classe d'estimateurs de Bayes de p admissibles par le Théorème 8.7 :

$$\delta^\pi(x) = a\frac{x}{n} + b, \qquad 0 \leq a \leq 1, \quad b \geq 0, \quad a+b \leq 1. \qquad \|$$

8.2.3 Estimateurs de Bayes généralisés inadmissibles

Nous l'avons vu, les estimateurs de Bayes ne sont pas nécessairement admissibles ; l'inadmissibilité est encore plus courante pour les estimateurs de Bayes associés à des lois impropres. Le cas particulier où le risque de Bayes d'un estimateur de Bayes associé à une loi impropre est fini (et où cet estimateur est donc admissible–voir la Proposition 2.37) est relativement rare, sauf pour les tests et d'autres cadres où le coût est borné (voir l'Exemple 2.38), et on a alors recours à des techniques plus élaborées pour prouver l'admissibilité, comme par exemple la condition de Stein (Section 8.3.3).

Exemple 8.10. On considère $x \sim \mathcal{N}_p(\theta, I_p)$ et $\delta_0(x) = x$; δ_0 est un estimateur de Bayes généralisé pour la distribution a priori $\pi(\theta) = 1$ sous le coût quadratique. L'effet Stein (Note 2.8.2) implique l'admissibilité de δ_0 si $p \leq 2$ (voir le Corollaire 8.14) et inadmissible sinon. ‖

Exemple 8.11. La distribution a priori employée dans l'Exemple 8.10 peut générer des cas d'inadmissibilité encore plus extrêmes. Par exemple, si $\pi(\theta) = 1$ et si le paramètre d'intérêt est $\eta = ||\theta||^2$, l'Exemple 3.32 montre que la distribution a posteriori de η est une loi du $\chi^2_p(||x||^2)$, ce qui amène l'estimateur de Bayes généralisé suivant :

$$\delta^\pi(x) = ||x||^2 + p.$$

Comme nous l'avons déjà vu, cet estimateur est inadmissible et dominé par $\tilde{\delta}(x) = (||x||^2 - p)^+$. L'Exemple 3.32 propose une distribution a priori alternative qui est plus appropriée dans ce contexte. ‖

Exemple 8.12. Soit $x \sim \mathcal{G}(\alpha, \theta)$ avec α supposé connu. Puisque θ est un paramètre d'échelle, $\pi(\theta) = 1/\theta$ est une distribution non informative appropriée (voir le Chapitre 9). La distribution a posteriori correspondante est $\mathcal{G}(\alpha, x)$ et donc

$$\delta^\pi(x) = \frac{\alpha}{x}$$

est l'estimateur de Bayes généralisé de θ sous coût quadratique. Pour un estimateur de la forme $\delta_c(x) = c/x$, le risque quadratique est

$$R(\theta, \delta_c) = \mathbb{E}_\theta \left(\frac{c}{x} - \theta\right)^2 = c^2 \mathbb{E}_\theta(x^{-2}) - 2c\theta \mathbb{E}_\theta(x^{-1}) + \theta^2.$$

Pour $\alpha > 2$, on a

$$\begin{aligned}
\mathbb{E}_\theta(x^{-2}) &= \frac{1}{\Gamma(\alpha)} \int_0^{+\infty} x^{-2} x^{\alpha-1} \theta^\alpha e^{-\theta x}\, dx \\
&= \frac{1}{\Gamma(\alpha)} \int_0^{+\infty} \theta^\alpha x^{\alpha-3} e^{-\theta x}\, dx \\
&= \theta^2 \frac{\Gamma(\alpha-2)}{\Gamma(\alpha)} = \frac{\theta^2}{(\alpha-1)(\alpha-2)}
\end{aligned}$$

et

$$\mathbb{E}_\theta(x^{-1}) = \frac{1}{\Gamma(\alpha)} \int_0^{+\infty} \theta^\alpha x^{\alpha-2} e^{-\theta x}\, dx$$
$$= \theta \frac{\Gamma(\alpha-1)}{\Gamma(\alpha)} = \frac{\theta}{\alpha-1}.$$

On en déduit que le meilleur estimateur de la forme δ_c est associé à

$$c^* = \frac{\theta \mathbb{E}_\theta(x^{-1})}{\mathbb{E}_\theta(x^{-2})} = \frac{\theta^2/(\alpha-1)}{\theta^2/(\alpha-1)(\alpha-2)} = \alpha - 2,$$

et donc que δ^π est dominé par δ_{c^*}. ‖

Ces trois exemples montrent bien que toutes les situations sont possibles pour les estimateurs de Bayes généralisés, de l'admissibilité de x pour $p = 1, 2$ (Exemple 8.10) à l'inadmissibilité forte des estimateurs des Exemples 8.11 et 8.12, en passant par l'inadmissibilité faible[67] de x pour $p \geq 3$ (Exemple 8.10).

8.2.4 Représentations différentielles

Pour les familles exponentielles multidimensionnelles, Brown et Hwang (1982) ont étendu le Théorème 8.7 à des distributions a priori impropres arbitraires. Soit une variable aléatoire

$$x \sim f(x|\theta) = h(x) e^{\theta.x - \psi(\theta)},$$

où θ et x appartiennent à $\mathbb{R}^p$. Rappelons que la moyenne de cette distribution est $\nabla\psi(\theta)$. Étant donné une mesure π de densité g sur Θ, on suppose que

$$I_x(\nabla g) = \int ||\nabla g(\theta)|| e^{\theta.x - \psi(\theta)}\, d\theta < +\infty. \tag{8.2}$$

Pour estimer $\nabla\psi(\theta)$ sous coût quadratique, l'estimateur de Bayes généralisé associé à g peut être représenté sous une forme différentielle

$$\delta_g(x) = x + \frac{I_x(\nabla g)}{I_x(g)}. \tag{8.3}$$

Les conditions suivantes sur g permettent d'établir l'admissibilité de δ_g :

[67] En fait, $\delta_0(x) = x$ reste un estimateur minimax quelle que soit la dimension et les estimateurs qui dominent δ_0 n'améliorent δ_0 (en termes de risque) de façon significative que dans une région relativement restreinte de l'espace d'échantillonnage (voir, par exemple, Bondar, 1987). La conséquence pratique de cette propriété est que, sans information a priori sur θ, la domination de δ_0 a une importance essentiellement formelle.

$$\int_{\{||\theta||>1\}} \frac{g(\theta)}{||\theta||^2 \log^2(||\theta|| \vee 2)} d\theta < \infty, \tag{8.4}$$

$$\int \frac{||\nabla g(\theta)||^2}{g(\theta)} d\theta < \infty, \tag{8.5}$$

et

$$\forall \theta \in \Theta, \qquad R(\theta, \delta_g) < \infty. \tag{8.6}$$

Théorème 8.13. *Sous les hypothèses (8.4), (8.5) et (8.6), l'estimateur (8.3) est admissible.*

La démonstration de ce résultat repose sur la condition de Blyth, présentée en Section 8.3.2. Elle ne sera donc développée que dans l'Exemple 8.25. Ce théorème a des conséquences importantes dans la mesure où il concerne le cas de l'estimation des paramètres d'espérance pour *toutes* les familles exponentielles continues sur $\mathbb{R}^p$. Entre autres, un cas particulier est l'obtention de l'admissibilité de Stein (1955b) pour toute famille exponentielle. Cela généralise aussi Zidek (1970), qui s'intéressait seulement au cas monodimensionnel (voir l'Exercice 8.8).

Corollaire 8.14. *Si $\Theta = \mathbb{R}^p$ et $p \leq 2$, l'estimateur $\delta_0(x) = x$ est admissible.*

Preuve. Considérons le cas $g \equiv 1$, alors $\nabla g \equiv 0$ et $\delta_g(x) = x$. Les conditions (8.4), (8.5) et (8.6) étant satisfaites, δ_g est admissible. □

Exemple 8.15. (Suite de l'Exemple 8.10) Si $x \sim \mathcal{N}_p(\theta, I_p)$, θ est le paramètre naturel de la distribution et le résultat original de Stein (1955a) est en fait le Corollaire 8.14. Remarquons que le Théorème 8.13 propose également une solution pour tester l'admissibilité d'autres estimateurs de Bayes généralisés de θ, notamment ceux qui sont étudiés par Strawderman (1971, Exercice 10.5) et Berger (1980b). ‖

Exemple 8.16. Soient x_1, x_2 deux variables aléatoires indépendantes de même loi $\mathcal{P}(\lambda_i)$ $(i = 1, 2)$. Si $\theta_i = \log(\lambda_i)$, $\delta_0(x) = (x_1, x_2)$ est un estimateur admissible de $(\lambda_1, \lambda_2) = (e_1^\theta, e_2^\theta)$. Ce résultat n'est pas vrai pour plus de deux dimensions, comme le montrent Hwang (1982a) et Johnstone (1984). ‖

Brown et Hwang (1982) présentent plusieurs généralisations du Théorème 8.13, couvrant des cas où $\Theta \neq \mathbb{R}^p$, comme les distributions gamma et géométrique. Ils démontrent également que, dans le cas particulier de p observations x_i issues de distributions de Poisson indépendantes, $\mathcal{P}(\lambda_i)$, l'estimateur de Bayes généralisé

$$\delta_{\mathrm{CZ}}(x) = \left[1 - \frac{\beta + p - 1}{\beta + p - 1 + S}\right] x,$$

avec $S = \sum_i x_i$, proposé par Clevenson et Zidek (1975) pour améliorer $x = (x_1, \ldots, x_p)$, est admissible pour $\beta > 0$ et $p \geq 2$ avec la fonction de coût

$$\mathrm{L}(\theta, \delta) = \sum_{i=1}^{p} \frac{1}{\lambda_i}(\delta - \lambda_i)^2.$$

Das Gupta et Sinha (1986) donnent aussi des conditions suffisantes d'admissibilité pour l'estimation de moyennes de lois gamma indépendantes.

8.2.5 Conditions de récurrence

Lorsqu'on se restreint au cas d'une distribution normale multidimensionnelle $\mathcal{N}_p(\theta, \Sigma)$, avec Σ connu, Brown (1971) parvient à donner une caractérisation plus précise des estimateurs de Bayes admissibles sous coût quadratique par le biais d'une condition nécessaire et suffisante, grâce à une représentation markovienne du problème d'estimation. (Ajoutons que Shinozaki, 1975, établit que le choix $\Sigma = I_p$ se fait sans perte de généralité, voir la Section 2.5.1 et Exercice 2.39.)

Théorème 8.17. *Soit $x \sim \mathcal{N}_p(\theta, I_p)$. Un estimateur de Bayes généralisé de la forme*

$$\delta(x) = (1 - h(||x||))x$$

est

(i) *inadmissible s'il existe $\epsilon > 0$ et $K < +\infty$ tels que, pour $||x|| > K$,*

$$||x||^2 h(||x||) < p - 2 - \epsilon;$$

et

(ii) *admissible s'il existe K_1 et K_2 tels que $h(||x||)||x|| \leq K_1$ pour tout x et, pour $||x|| > K_2$,*

$$||x||^2 h(||x||) \geq p - 2.$$

La démonstration de ce résultat est assez difficile. Le raisonnement conduisant à (i) et (ii) inclut la preuve de la récurrence ou de la transience d'un processus aléatoire[68] associé à δ. (Voir Srinivasan, 1981, pour une description simplifiée.) La partie (i) peut aussi être vue comme une conséquence du Lemme 8.38 ci-dessous. Remarquons la présence du facteur $(p - 2)$, qui indiquait déjà la limite entre admissibilité et inadmissibilité de l'estimateur usuel

[68] Les marches aléatoires sont généralement récurrentes en dimension 1 ou 2 et transientes dans des dimensions plus grandes (voir Feller, 1971, ou Meyn et Tweedie, 1994). Le lien établi par Brown (1971) prouve que le fait que $p = 3$ soit un cas limite dans les deux problèmes n'est pas une coïncidence.

$\delta_0(x) = x$. La relation entre ce résultat et le phénomène de Stein est détaillée en Section 8.5.

Johnstone (1984) donne un équivalent du Théorème 8.17 pour le modèle de Poisson. Si $x_i \sim \mathscr{P}(\lambda_i)$ $(i = 1, \ldots, p)$, le paramètre $\lambda = (\lambda_1, \ldots, \lambda_p)$ est estimé sous le coût

$$\sum_{i=1}^{p} \frac{1}{\lambda_i}(\delta_i - \lambda_i)^2.$$

Alors :

Théorème 8.18. *Un estimateur de Bayes généralisé de la forme*

$$\delta(x) = (1 - h(s))x,$$

avec $s = \sum_i x_i$, *est*

(i) *inadmissible s'il existe* $\epsilon > 0$ *et* $K < +\infty$ *tels que, pour* $s > K$,

$$sh(s) < (p - 1 - \epsilon);$$

et

(ii) *admissible s'il existe* K_1 *et* K_2 *tels que* $\sqrt{s}\, h(s) \leq K_1$ *pour tout* s *et, pour* $s > K_2$,

$$sh(s) \geq (p - 1).$$

Eaton (1992) dresse des parallèles similaires à ceux décrits par Brown (1971) entre l'admissibilité d'un estimateur et la récurrence d'une chaîne de Markov associée. Nous citons ci-dessous les principaux résultats de cet article mais encourageons les lecteurs à le consulter non seulement pour les démonstrations complètes, mais aussi pour les développements intéressants sur les conséquences de ces résultats. Le problème considéré par Eaton (1992) est de chercher si, pour une fonction *bornée* $g(\theta)$, un estimateur de Bayes généralisé associé à une mesure a priori π est admissible sous coût quadratique. En supposant que la distribution a posteriori $\pi(\theta|x)$ soit bien définie, nous considérons le noyau de transition

$$K(\theta|\eta) = \int_{\mathscr{X}} \pi(\theta|x) f(x|\eta)\, dx, \tag{8.7}$$

associé à la chaîne de Markov $(\theta^{(n)})$ définie comme suit. La transition de $\theta^{(n)}$ à $\theta^{(n+1)}$ correspond d'abord à la simulation de $x \sim f(x|\theta^{(n)})$, puis à celle de $\theta^{(n+1)} \sim \pi(\theta|x)$. (Concernant l'utilisation de ce noyau dans des méthodes de Monte Carlo par chaînes de Markov et pour de plus amples détails sur la théorie des chaînes de Markov, voir le Chapitre 6.) Pour tout ensemble mesurable C tel que $\pi(C) < +\infty$, on définit :

$$V(C) = \left\{h \in \mathscr{L}^2(\pi); h(\theta) \geq 0 \text{ et } h(\theta) \geq 1 \text{ lorsque } \theta \in C\right\}$$

et

$$\Delta(h) = \int\int \{h(\theta) - h(\eta)\}^2 K(\theta|\eta)\pi(\eta)\, d\theta\, d\eta.$$

Le résultat suivant permet alors de caractériser l'admissibilité pour *toute fonction bornée* en fonction de Δ et $V(C)$ et donc indépendamment des fonctions estimées g :

Théorème 8.19. *Si, pour tout C tel que $\pi(C) < +\infty$,*

$$\inf_{h \in V(C)} \Delta(h) = 0, \tag{8.8}$$

alors l'estimateur de Bayes $\mathbb{E}^\pi[g(\theta)|x]$ est admissible sous coût quadratique pour toute fonction bornée g.

Ce résultat est naturellement assez général, mais n'est que modérément utile dans la mesure où la vérification pratique de (8.8) pour tout ensemble C peut être très lourde. Il faut également noter que (8.8) est toujours vraie lorsque π est une distribution a priori propre, puisque $h \equiv 1$ appartient à $\mathscr{L}^2(\pi)$ et $\Delta(1) = 0$ dans ce cas. L'extension aux lois a priori impropres s'appuie sur des approximations de 1 par des fonctions de $V(C)$. (Voir le Chapitre 9 pour un lien analogue entre difficultés de calcul et minimaxité.) Eaton (1992) donne une condition équivalente au Théorème 8.19 en s'appuyant sur la chaîne de Markov $(\theta^{(n)})$. Pour un ensemble donné C, une condition d'arrêt σ_C est définie comme le premier entier $n > 0$ tel que $(\theta^{(n)})$ appartienne à C (et $+\infty$ sinon). On dit que la chaîne $(\theta^{(n)})$ est *π-récurrente* si la probabilité que σ_C soit finie vaut 1 pour π-presque tout point de départ $\theta^{(0)}$.

Théorème 8.20. *Pour tout ensemble C tel que $\pi(C) < +\infty$,*

$$\inf_{h \in V(C)} \Delta(h) = \int_C \left\{1 - P(\sigma_C < +\infty | \theta^{(0)} = \eta)\right\} \pi(\eta)\, d\eta.$$

Par conséquent, les estimateurs de Bayes généralisés de fonctions bornées de θ sont admissibles si la chaîne de Markov associée $(\theta^{(n)})$ est π-récurrente.

Des extensions, exemples et commentaires sur ce résultat se trouvent dans la Note 8.7.1 et dans Eaton (1992, 1999). Son intérêt essentiel, outre son élégance mathématique, est que la vérification de la récurrence de la chaîne de Markov $(\theta^{(n)})$ est beaucoup plus aisée que la détermination de la borne inférieure de $\Delta(h)$. De plus, ce théorème permet d'obtenir une vérification numérique d'admissibilité en simulant une chaîne $(\theta^{(n)})$, ce qui rappelle la vérification numérique de minimaxité proposée par Berger et Robert (1990).

8.3 Conditions nécessaires et suffisantes d'admissibilité

Les résultats présentés dans la section précédente ne concernent que les estimateurs de Bayes généralisés. En outre, certaines conditions sont très difficiles à vérifier–on pense notamment à (8.4) ou (8.5). Nous introduisons dans

cette section une condition générale nécessaire et suffisante d'admissibilité qui n'exige pas que les estimateurs soient de Bayes généralisés. Elle formalise en quelque sorte l'affirmation déjà énoncée que *"les estimateurs admissibles sont des limites d'estimateurs de Bayes...."*. Une première version de la condition de Stein concerne uniquement les estimateurs à risque continu ; dans la Section 8.3.1, nous expliquons pourquoi il est généralement suffisant de ne considérer que ceux-ci.

8.3.1 Risques continus

Il est souvent nécessaire de restreindre le cadre d'étude aux estimateurs à fonctions de risque continues pour obtenir une condition suffisante d'admissibilité. Toutefois, dans certains cas, tous les estimateurs sont à risque continu. Dans d'autres situations, les estimateurs admissibles sont nécessairement à risque continu.

Lemme 8.21. *Soit $\Theta \subset \mathbb{R}^m$. La fonction de coût $\mathrm{L}(\theta, \delta)$ est supposée bornée et continue en θ pour tout $\delta \in \mathscr{D}$. Si $f(x|\theta)$ est continue en θ pour tout x, la fonction de risque de tout estimateur est continue.*

Preuve. Étant donné un estimateur δ, la différence des risques en θ et $\theta' \in \Theta$ est

$$\begin{aligned} |R(\theta,\delta) - R(\theta',\delta)| &= \left| \int \mathrm{L}(\theta,\delta(x)) f(x|\theta)\, dx - \int \mathrm{L}(\theta',\delta(x)) f(x|\theta')\, dx \right| \\ &\le \int \big|\mathrm{L}(\theta,\delta(x)) - \mathrm{L}(\theta',\delta(x))\big| f(x|\theta)\, dx \\ &\quad + \left| \int \mathrm{L}(\theta,\delta(x))(f(x|\theta) - f(x|\theta'))\, dx \right|. \end{aligned}$$

Puisque L est continue et bornée par C, il existe $\eta_0 > 0$ et un ensemble compact K_0 tels que

$$\int_{K_0^c} f(x|\theta)\, dx < \frac{\epsilon}{8C} \quad \text{et} \quad \int_{K_0} \big|\mathrm{L}(\theta,\delta(x)) - \mathrm{L}(\theta',\delta(x))\big| f(x|\theta)\, dx < \frac{\epsilon}{4}$$

avec $||\theta - \theta'|| < \eta_0$. Ainsi,

$$\int \big|\mathrm{L}(\theta,\delta(x)) - \mathrm{L}(\theta',\delta(x))\big| f(x|\theta)\, dx < \frac{\epsilon}{2}.$$

De plus, $f(x|\theta)$ étant une fonction continue de θ, un argument équivalent permet d'écrire qu'il existe $\eta_1 > 0$ et un ensemble compact K_1 tels que

$$\begin{aligned} \left| \int \mathrm{L}(\theta,\delta(x))(f(x|\theta) - f(x|\theta'))\, dx \right| &\le C \int_{K_1} \big|f(x|\theta) - f(x|\theta')\big|\, dx \\ &\quad + C \int_{K_1^c} [f(x|\theta) + f(x|\theta')]\, dx < \frac{\epsilon}{2} \end{aligned}$$

et

$$\int_{K_1^c} f(x|\theta)\,dx < \frac{\epsilon}{8C},$$

avec $||\theta - \theta'|| < \eta_1$. Donc $R(\theta, \delta)$ est continue. □

L'intérêt du Lemme 8.21 est plus ou moins limité puisque les problèmes d'admissibilité les plus difficiles concernent justement les cas où L n'est pas bornée. Dans certains contextes, on peut cependant réduire la classe des estimateurs à considérer à la classe des estimateurs à risque continu. On parle de caractérisation de *classe complète*.

Définition 8.22. *Une classe $\mathscr{C}$ d'estimateurs est dite* complète *si, quel que soit $\delta' \notin \mathscr{C}$, il existe $\delta \in \mathscr{C}$ qui domine δ'. La classe est* essentiellement complète *si, quel que soit $\delta' \notin \mathscr{C}$, il existe $\delta \in \mathscr{C}$ au moins aussi bon que δ'.*

Si on excepte les cas triviaux comme celui de la classe de tous les estimateurs, il n'est pas toujours possible de déterminer des classes complètes utiles. Par exemple, il existe des cas, bien que rares, où la classe des estimateurs admissibles *n'est pas* une classe complète (voir Blackwell et Girshick, 1954, Théorème 5.7.1, ou Brown, 1976). La Section 8.4 analyse les relations entre les estimateurs de Bayes, les estimateurs de Bayes généralisés et les classes complètes. Le résultat suivant est un lemme de classe complète énonçant des conditions suffisantes pour n'avoir à considérer que les estimateurs à risque continu.

Lemme 8.23. *Soit un modèle de décision statistique $\mathscr{X}, \Theta \subset \mathbb{R}$ avec un espace de décision fermé $\mathscr{D} \subset \mathbb{R}$. On suppose que $f(x|\theta)$ vérifie la propriété de rapport de vraisemblances monotone et est continue en θ. Si*

(i) $\mathrm{L}(\theta, d)$ *est une fonction continue de θ pour tout $d \in \mathscr{D}$;*

(ii) L *est décroissante en d pour $d < \theta$ et croissante pour $d > \theta$; et*

(iii) *il existe deux fonctions K_1 et K_2 bornées sur les sous-ensembles compacts de Θ, telles que*

$$\mathrm{L}(\theta_1, d) \leq K_1(\theta_1, \theta_2)\mathrm{L}(\theta_2, d) + K_2(\theta_1, \theta_2),$$

alors les estimateurs à risque fini et continu forment une classe complète.

Voir Ferguson (1967) et Brown (1976) pour d'autres résultats. Par exemple, il est possible de montrer que si le problème est *monotone*, alors les *estimateurs monotones* constituent une classe complète (Exercice 8.23 et Théorème 5.43).

8.3.2 Condition suffisante de Blyth

Avant que Stein (1955b) n'établisse sa condition nécessaire et suffisante (Section 8.3.3), Blyth (1951) propose une condition suffisante d'admissibilité, qui fait un lien entre l'admissibilité d'un estimateur et l'existence d'une suite de distributions a priori approchant cet estimateur.

Théorème 8.24. *Soit un ensemble ouvert non vide $\Theta \subset \mathbb{R}^p$. On suppose que les estimateurs à risque continu forment une classe complète. Si, pour un estimateur à risque continu δ_0, il existe une suite (π_n) de distributions a priori généralisées telles que*

(i) $r(\pi_n, \delta_0)$ *est fini quel que soit n ;*

(ii) *pour tout ensemble ouvert non vide $C \subset \Theta$, il existe $K > 0$ et N tels que, pour tout $n \geq N$, $\pi_n(C) \geq K$; et*

(iii) $\lim_{n \to +\infty} r(\pi_n, \delta_0) - r(\pi_n) = 0$;

alors l'estimateur δ_0 est admissible.

Preuve. Si δ_0 n'est pas admissible, il existe un estimateur δ' dominant δ_0, c'est-à-dire tel que $R(\theta, \delta) - R(\theta, \delta') \geq 0$ et

$$R(\theta, \delta) - R(\theta, \delta') > \epsilon$$

sur un ensemble ouvert $C \subset \Theta$ (pour ϵ suffisamment petit). Il découle ensuite des hypothèses (i) et (ii), que, pour $n \geq N$,

$$\begin{aligned} r(\pi_n, \delta_0) - r(\pi_n) &\geq r(\pi_n, \delta_0) - r(\pi_n, \delta') \\ &= \mathbb{E}^{\pi}[R(\theta, \delta_0) - R(\theta, \delta')] \\ &\geq \int_C (R(\theta, \delta_0) - R(\theta, \delta'))\pi_n(\theta)\, d\theta \\ &\geq \epsilon \int_C \pi_n(\theta)\, d\theta \geq \epsilon K. \end{aligned}$$

□

Ce résultat est utile pour établir l'admissibilité d'estimateurs de Bayes généralisés, puisque les mesures π associées à ces estimateurs peuvent s'écrire comme des limites de suites de distributions propres π_n. Cela dit, le choix de telles suites n'est pas toujours évident, comme le montrent Berger (1982a) ou Brown et Hwang (1982). Le Théorème 8.24 s'applique également à d'autres estimateurs, dans des contextes où il existe des estimateurs admissibles qui ne sont pas de Bayes généralisés (voir la Section 8.4).

Exemple 8.25. La preuve du Théorème 8.13 est une première illustration de la condition de Blyth. Soit h_n à valeurs dans $[0, 1]$, dérivable et telle que $h_n(\theta) = 0$ si $||\theta|| > n$ et $h_n(\theta) = 1$ sur un ensemble S vérifiant

$$\int_S g(\theta)\,d\theta > 0.$$

Nous définissons à présent une suite de mesures associées de densités $g_n(\theta) = h_n^2(\theta)g(\theta)$ et les estimateurs de Bayes correspondants δ_n. En repassant à la notation $I_x(.)$ adoptée en (8.2), la différence des risques de Bayes intégrés est

$$\begin{aligned} r(\pi_n, \delta_g) - r(\pi_n) &= \int ||\delta_g(x) - \delta_n(x)||^2 I_x(g_n)\,dx \\ &= \int \left\| \frac{I_x(\nabla g)}{I_x(g)} - \frac{I_x(h_n^2 \nabla g)}{I_x(g_n)} - \frac{I_x(g\nabla h_n)}{I_x(g_n)} \right\|^2 I_x(g_n)\,dx, \end{aligned}$$

avec la notation de (8.3). Par conséquent,

$$\begin{aligned} r(\pi_n, \delta_g) - r(\pi_n) &\leq 2\int \left\| \frac{I_x(\nabla g)}{I_x(g)} - \frac{I_x(h_n^2 \nabla g)}{I_x(g_n)} \right\|^2 I_x(g_n)\,dx \\ &\quad + 2\int \left\| \frac{I_x(g\nabla h_n)}{I_x(g_n)} \right\|^2 I_x(g_n)\,dx \\ &= B_n + A_n. \end{aligned}$$

Le second terme, A_n, admet pour borne supérieure

$$4\int ||\nabla h_n(\theta)||^2 g(\theta)\,d\theta.$$

Dans le cas particulier

$$h_n(\theta) = \begin{cases} 1 & \text{pour } ||\theta|| < 1, \\ 1 - \dfrac{\log(||\theta||)}{\log(n)} & \text{pour } 1 < ||\theta|| < n, \\ 0 & \text{sinon,} \end{cases}$$

on obtient en fait

$$||\nabla h_n(\theta)||^2 \leq \frac{1}{||\theta||^2 \log^2(\max(||\theta||, 2))} \mathbb{I}_{||\theta||>1}(\theta)\,,$$

et la condition (8.4) implique que A_n converge vers 0 quand n tend vers l'infini. Le premier terme satisfait

$$\begin{aligned} B_n &= \int \left\| I_x\left(g_n \frac{I_x(\nabla g)}{I_x(g)} - h_n^2 \nabla g \right) \right\|^2 \Big/ (I_x(g_n))\,dx \\ &= \int \left\| I_x\left(g_n \left[\frac{I_x(\nabla g)}{I_x(g)} - \frac{\nabla g}{g} \right] \right) \right\|^2 \Big/ (I_x(g_n))\,dx \\ &\leq \int I_x\left(g \left\| \frac{I_x(\nabla g)}{I_x(g)} - \frac{\nabla g}{g} \right\|^2 \right) dx. \end{aligned}$$

En utilisant (8.5), on obtient par le théorème de convergence dominée que B_n a pour limite 0, puisque g_n tend vers g. Ceci achève la démonstration du Théorème 8.13. ‖

En pratique, une méthode typique d'utilisation de la condition de Blyth pour un estimateur de Bayes généralisé, δ_0, est de construire une suite d'estimateurs de Bayes propres qui tend vers δ_0, puis de "dénormaliser" la suite de distributions a priori associées par un poids adéquat.

Exemple 8.26. On considère $x \sim \mathcal{N}(\theta, 1)$ et $\delta_0(x) = x$, un estimateur de θ. Parce que δ_0 correspond à $\pi(\theta) = 1$ sous coût quadratique, nous choisissons pour mesure π_n avec une densité

$$g_n(x) = e^{-\theta^2/2n},$$

c'est-à-dire la densité d'une distribution normale $\mathcal{N}(0, n)$ sans le facteur de normalisation $1/\sqrt{2\pi n}$. Comme les densités g_n sont croissantes en n, la condition (ii) du Théorème 8.24 est satisfaite, ainsi que (i) : l'estimateur de Bayes pour π_n est toujours

$$\delta_n(x) = \frac{nx}{n+1},$$

puisque l'absence du facteur de normalisation n'a pas de conséquence directe dans ce cas, et

$$\begin{aligned} r(\pi_n) &= \int_{\mathbb{R}} \left[\frac{\theta^2}{(n+1)^2} + \frac{n^2}{(n+1)^2} \right] g_n(\theta)\, d\theta \\ &= \sqrt{2\pi n}\, \frac{n}{n+1}, \end{aligned}$$

ainsi que

$$r(\pi_n, \delta_0) = \int_{\mathbb{R}} 1\ g_n(\theta)\, d\theta = \sqrt{2\pi n}.$$

Les deux risques sont donc finis. De plus,

$$r(\pi_n, \delta_0) - r(\pi_n) = \sqrt{2\pi n}/(n+1)$$

tend vers 0. La condition de Blyth fournit donc une autre preuve d'admissibilité de $\delta_0(x) = x$ dans le cas normal. En revanche, la preuve d'admissibilité de δ_0 en dimension deux requiert une suite plus compliquée (voir Stein, 1955a). ‖

Exemple 8.27. Soit $x \sim \mathcal{B}(m, \theta)$. Le problème d'inférence est de tester l'hypothèse nulle $H_0 : \theta \leq \theta_0$ sous la fonction de coût quadratique décrite en Section 5.4,

$$\left(\mathbb{I}_{[0,\theta_0]}(\theta) - \gamma(x)\right)^2.$$

La *p-value* est alors

$$\varphi(x) = P_{\theta_0}(X \geq x) = \sum_{k=x}^{m} \binom{m}{k} \theta_0^k (1-\theta_0)^{m-k}.$$

Les distributions conjuguées naturelles sont ici des distributions bêta. L'idée est donc d'approcher $\varphi(x)$ par une suite d'estimateurs associée à une suite de distributions bêta convenablement choisies. En fait, $\varphi(x)$ peut s'écrire (pour $x \neq 0$)

$$\varphi(x) = \frac{1}{B(x, m-x+1)} \int_0^{\theta_0} t^{x-1}(1-t)^{m-x}\, dt = P(T \leq \theta_0 | x)$$

lorsque $T \sim \mathscr{B}e(x, m-x+1)$, ce qui correspond à la distribution a priori généralisée

$$\pi(\theta) = \theta^{-1} \qquad (0 < \theta < 1).$$

On considère π_n de densité

$$g_n(\theta) = \theta^{\alpha_n - 1}$$

sur $[0,1]$ avec la suite (α_n) qui décroît vers 0. Dans ce cas, la procédure bayésienne classique est

$$\gamma^{\pi_n}(x) = P^{\pi_n}(\theta \leq \theta_0 | x) = \frac{1}{B(x+\alpha_n, m-x+1)} \int_0^{\theta_0} t^{x+\alpha_n-1}(1-t)^{m-x}\, dt$$

et

$$r(\pi_n) = \sum_{k=0}^{m} B(k+\alpha_n, m-k+1)\gamma^{\pi_n}(k)(1-\gamma^{\pi_n}(k)),$$
$$r(\pi_n, \varphi) = \sum_{k=0}^{m} B(k+\alpha_n, m-k+1)(\gamma^{\pi_n}(k) - 2\gamma^{\pi_n}\varphi(k) + \varphi^2(k)).$$

On en déduit que

$$r(\pi_n, \varphi) - r(\pi_n) = \sum_{k=0}^{m} B(k+\alpha_n, m-k+1)(\gamma^{\pi_n}(k) - \varphi(k))^2.$$

Si $k \neq 0$, on vérifie sans difficulté que

$$\lim_{\alpha_n \to 0} (\varphi(k) - \gamma^{\pi_n}(k))^2 = 0.$$

De même, on a

$$\lim_{\alpha\to 0} \frac{\int_0^{\theta_0} t^{\alpha-1}(1-t)^{m-1}dt}{\int_0^1 t^{\alpha-1}(1-t)^{m-1}dt} = 1\,,$$

pour le cas $k = 0$. En outre, la condition (ii) est également satisfaite. La *p-value* φ est alors admissible dans ce cadre. L'Exemple 5.45 donne une preuve plus directe de ce résultat tirant profit du fait que le risque de Bayes est fini. ‖

Les Exemples 8.25 et 8.27 illustrent un résultat général : sous coût quadratique, la condition (iii) du Théorème 8.24 implique la convergence quadratique des estimateurs de Bayes vers δ_0 au sens des mesures marginales.

Proposition 8.28. *Si* L *est une fonction de coût quadratique et s'il existe une suite* (π_n) *vérifiant les conditions* (i), (ii) *et* (iii) *du Théorème 8.24, alors les estimateurs de Bayes* δ^{π_n} *tendent quadratiquement vers* δ_0 *pour les mesures marginales*

$$m_n(x) = \int_\Theta f(x|\theta)\pi_n(\theta)\,d\theta.$$

Preuve. La différence des risques s'écrit naturellement

$$\begin{aligned} r(\pi_n,\delta_0) - r(\pi_n) &= \int_{\mathscr{X}}\int_\Theta (||\delta_0(x)-\theta||^2 - ||\delta^{\pi_n}(x)-\theta||^2)\pi_n(\theta|x)\,d\theta\, m_n(x)\,dx \\ &= \int_{\mathscr{X}} \Bigg[||\delta_0(x)-\delta^{\pi_n}(x)||^2 \\ &\quad + 2(\delta_0(x)-\delta^{\pi_n}(x))\cdot\int_\Theta (\delta^{\pi_n}(x)-\theta)\pi_n(\theta|x)\,d\theta\Bigg] m_n(x)\,dx \\ &= \int_{\mathscr{X}} ||\delta_0(x)-\delta^{\pi_n}(x)||^2 m_n(x)\,dx, \end{aligned}$$

puisque

$$\int_\Theta (\delta^{\pi_n}(x)-\theta)\pi_n(\theta|x)\,d\theta = 0.$$

□

Malheureusement, ce résultat de convergence dépend de la suite (m_n), sauf s'il est possible d'établir une équivalence uniforme avec la mesure de Lebesgue, ou une autre mesure fixée, auquel cas il y a convergence quadratique au sens classique. C'est par exemple ce qui se passe lorsque la suite (m_n) est croissante, comme dans les Exemples 8.25, 8.26 et 8.27. La Section 8.3.4 décrit un résultat plus fondamental dû à Brown (1986b), qui montre que la convergence ponctuelle des δ^{π_n} vers δ_0, indépendamment des mesures m_n, est en réalité nécessaire.

8.3.3 Condition nécessaire et suffisante de Stein

Les compléments apportés par Stein (1955b) et Farrell (1968a) à la condition précédente permettent de déduire un résultat encore plus important que le Théorème 8.24, puisqu'il établit que *tous les estimateurs admissibles sont des limites de suites d'estimateurs de Bayes* (au sens du risque de Bayes). Les hypothèses de Farrell (1968a) sont

(i) $f(x|\theta)$ est continu en θ et strictement positive sur Θ ; et

(ii) le coût L est strictement convexe, continu et, si $E \subset \Theta$ est compact,

$$\lim_{\|\delta\| \to +\infty} \inf_{\theta \in E} \mathrm{L}(\theta, \delta) = +\infty.$$

Remarquons que cette seconde hypothèse élimine nécessairement les fonctions de coût bornées.

Théorème 8.29. *Sous les hypothèses* (i) *et* (ii)*, un estimateur δ est admissible si et seulement si il existe une suite (F_n) d'ensembles compacts croissants tels que $\Theta = \bigcup_n F_n$, une suite (π_n) de mesures finies de supports F_n et une suite (δ_n) d'estimateurs de Bayes associés à π_n tels que*

(i) *il existe un ensemble compact $E_0 \subset \Theta$ tel que* $\inf_n \pi_n(E_0) \geq 1$;

(ii) *si $E \subset \Theta$ est compact,* $\sup_n \pi_n(E) < +\infty$;

(iii) $\lim_n r(\pi_n, \delta) - r(\pi_n) = 0$; *et*

(iv) $\lim_n R(\theta, \delta_n) = R(\theta, \delta)$.

De ce théorème fondamental découlent la plupart des résultats d'admissibilité et de classe complète présentés en Section 8.4. Une démonstration du Théorème 8.29 dépasse le cadre de ce livre ; voir Farrell (1968a). La *suffisance* est liée à la condition de Blyth, mais la réciproque *nécessaire* permet d'exclure de nombreux estimateurs inadmissibles.

8.3.4 Un autre théorème limite

Brown (1986b) donne une caractérisation alternative, assez générale, des estimateurs admissibles. Soit $x \sim f(x|\theta)$, avec $f(x|\theta) > 0$. On suppose que $\mathscr{D}$ est un ensemble fermé convexe. De plus, on suppose que la fonction de coût L est semi-continue inférieurement et telle que

$$\lim_{||\delta|| \to +\infty} \mathrm{L}(\theta, \delta) = +\infty.$$

(Cela correspond plus ou moins à l'hypothèse (ii) de Farrell, 1968a.) Le résultat principal de Brown (1986b) consiste à montrer que, sous ces hypothèses, l'adhérence (au sens de la convergence ponctuelle) de l'ensemble

des estimateurs de Bayes est une classe complète. Le résultat de convergence qui suit reformule cette propriété (voir Brown, 1986b, p. 254-267).

Proposition 8.30. *Si* L *est strictement convexe, tout estimateur admissible de θ est une limite ponctuelle d'estimateurs de Bayes pour une suite de distributions a priori à supports finis.*

Ce résultat est à comparer aux résultats de Dalal et Hall (1983) et Diaconis et Ylvisaker (1985), présentés en Section 3.4 et qui montrent que, pour une famille exponentielle, toute distribution a priori est une limite de mélanges de distributions a priori conjuguées. Par conséquent, pour les familles exponentielles, un estimateur admissible est aussi la limite d'estimateurs de Bayes associés à un mélange de distributions a priori conjuguées. Lorsque le modèle est invariant par transformation sphérique, les distributions à support fini peuvent être remplacées par des distributions sur des sphères imbriquées, puisque celles-ci préservent la symétrie. Dans ce cas, si π_c est la distribution uniforme sur la sphère de rayon c,

$$\mathscr{S}_c = \{\theta;\ ||\theta|| = c\},$$

et si δ_c est l'estimateur de Bayes associé sous coût quadratique, c'est-à-dire la moyenne a posteriori, Robert (1990) dérive le théorème limite suivant.

Proposition 8.31. *Si $x \sim \mathscr{N}_p(\theta, I_p)$ et si π est une distribution a priori à symétrie sphérique de centre* 0*, alors il existe deux suites, (q_n^i) et (c_n^i), telles que $\sum_{i=1}^n q_n^i = 1$ et*

$$m^\pi(x) = \int_{\mathbb{R}^p} f(x|\theta)\pi(\theta)\,d\theta = \lim_{n\to+\infty} \sum_{i=1}^n q_n^i m_{c_n^i}(x),$$

avec

$$m_{c_n^i} = \int_{\mathbb{R}^p} f(x|\theta)\pi_{c_n^i}(\theta)\,d\theta.$$

De plus, sous coût quadratique,

$$\delta^\pi(x) = \lim_{n\to+\infty} \sum_{i=1}^n \frac{q_n^i m_{c_n^i}(x)}{\sum_j q_n^j m_{c_n^j}(x)}\ \delta_{c_n^i}(x). \tag{8.9}$$

Par conséquent, dans le cas normal, tout estimateur de Bayes associé à une distribution a priori à symétrie sphérique est une limite ponctuelle d'estimateurs de Bayes associés à des distributions uniformes sur des sphères. On rappelle que les estimateurs δ_c peuvent s'écrire

$$\delta_c(x) = c\ \frac{I_{p/2}(||x||c)}{I_{p/2-1}(||x||c)}\ \frac{x}{||x||}, \tag{8.10}$$

où I_ν est la *fonction de Bessel modifiée* (Exercices 4.36 et 4.37). Une conséquence établie par Kempthorne (1988) est en fait que *tout estimateur admissible $\delta(x)$ peut être écrit sous la forme* (8.10) *ou alors il existe un estimateur*

δ' *de la forme* (8.10) *équivalent à* δ (en termes de risque).

8.4 Classes complètes

Nous venons de voir dans un cadre général que les estimateurs admissibles peuvent être considérés comme des limites d'estimateurs de Bayes de plusieurs points de vue. Dans certains cas particuliers, il est possible de décrire plus précisément ces estimateurs admissibles et de montrer qu'ils sont des estimateurs de Bayes généralisés. L'intérêt de ces résultats est multiple. D'une part, ils permettent de réduire la classe des estimateurs à considérer. D'autre part, ils illustrent l'avantage de ne faire appel qu'à des estimateurs de Bayes ou de Bayes généralisés d'un point de vue fréquentiste. Cela concerne, par exemple, le cas de l'évaluation de procédures de test sous coût quadratique, vue en Section 5.4 (Théorèmes 5.42 et 5.43). Cette section donne des résultats analogues pour l'estimation ponctuelle. On trouvera d'autres références dans Brown (1986b) et Rukhin (1995).

En guise d'introduction, considérons l'exemple très simple où $\Theta = \{\theta_1, \theta_2\}$, qui a l'avantage de permettre une représentation graphique de l'*ensemble de risque*,

$$\mathscr{R} = \{r = (R(\theta_1, \delta), R(\theta_2, \delta)),\ \delta \in \mathscr{D}^*\},$$

en notant $\mathscr{D}^*$ l'ensemble des estimateurs randomisés. On suppose que l'ensemble de risque $\mathscr{R}$ est *borné et fermé inférieurement*, c'est-à-dire tel que tous les risques sur la frontière inférieure de $\mathscr{R}$ appartiennent à $\mathscr{R}$ et ont des composantes finies. Cette hypothèse est vérifiée lorsque le coût est positif. Cette frontière inférieure, que nous noterons $\Gamma(\mathscr{R})$, est importante dans la mesure où elle contient en fait les points *admissibles* de $\mathscr{R}$. En effet, si $r \in \Gamma(\mathscr{R})$, il ne peut exister $r' \in \mathscr{R}$ tel que $r'_1 \leq r_1$ et $r'_2 \leq r_2$ avec inégalité stricte sur l'un des deux axes. Par ailleurs, pour tout $r \in \Gamma(\mathscr{R})$, il existe une tangente à $\mathscr{R}$ passant par r, avec une pente positive et d'équation

$$p_1 r_1 + p_2 r_2 = k,$$

c'est-à-dire telle que tout $r' \in \mathscr{R}$ vérifie $p_1 r'_1 + p_2 r'_2 \geq k$, ce que montre la Figure 8.1. (Il s'agit en réalité d'une conséquence de la convexité de $\mathscr{R}$.) Cette propriété implique que r est un estimateur de Bayes pour la distribution a priori $\pi(\theta_i) = p_i$ $(i = 1, 2)$, puisqu'il minimise le risque de Bayes $p_1 r_1 + p_2 r_2$. On en déduit le résultat général suivant.

Proposition 8.32. *Si* Θ *est fini et si l'ensemble de risque* $\mathscr{R}$ *est borné et fermé inférieurement, alors l'ensemble des estimateurs de Bayes forme une classe complète.*

Cette caractérisation repose sur le théorème de l'*hyperplan séparateur* puisque, sous les hypothèses du théorème, il existe un hyperplan tangent à l'ensemble de risque pour tout point de la frontière inférieure et que cet hyperplan définit une distribution a priori sur Θ par dualité. L'extension de

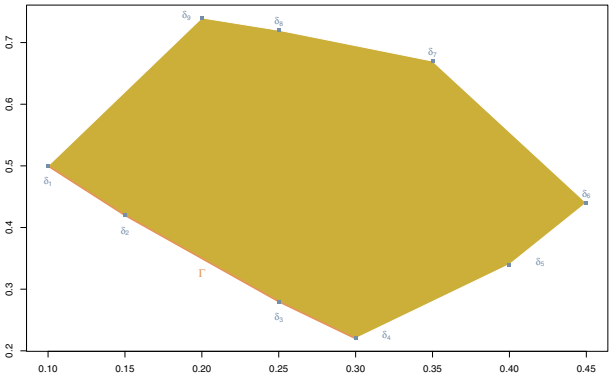

Fig. 8.1. Ensemble de risque et estimateurs admissibles pour $\Theta = \{\theta_1, \theta_2\}$.

ce résultat de classe complète à des espaces de paramètres Θ dénombrables et non dénombrables nécessite une généralisation équivalente des théorèmes d'hyperplan séparateur aux espaces de fonctions sur Θ. Par exemple, Brown (1976) donne le résultat suivant, en notant $\mathring{S}$ l'intérieur de S.

Lemme 8.33. *Soit S un sous-ensemble convexe d'un espace vectoriel topologique $\mathscr{E}$. Si $\mathring{S} \neq \emptyset$ et $y_0 \notin \mathring{S}$, il existe $f \in \mathscr{E}^*$ telle que S soit incluse dans $\{y;\ f(y) \geq f(y_0)\}$.*

On déduit de ce lemme le résultat de classe complète suivant, dû à Wald (1950) et qui généralise la Proposition 8.32.

Théorème 8.34. *On suppose que Θ est compact et que l'ensemble de risque $\mathscr{R}$ est convexe. Si tous les estimateurs ont une fonction de risque continue, les estimateurs de Bayes constituent une classe complète.*

Preuve. Ce résultat est bien une conséquence du Lemme 8.33 puisque, si δ_0 est admissible, la fonction de risque $R(\theta, \delta_0)$ appartient à la frontière inférieure de l'ensemble de risque. Par conséquent, il existe une fonction linéaire sur $\mathscr{R}$, ψ^*, telle que, pour tout estimateur δ,

$$\psi^*(R(\cdot, \delta)) \geq \psi^*(R(\cdot, \delta_0)).$$

Il vient alors du *théorème de représentation de Riesz* qu'il existe une mesure finie π sur Θ telle que

$$\psi^*(R(\cdot, \delta)) = \int_\Theta R(\theta, \delta)\pi(\theta)d\theta,$$

et que cette mesure peut être renormalisée ainsi $\tilde{\pi}(\theta) = \pi(\theta)/\pi(\Theta)$, définissant par là-même une distribution a priori. L'inégalité ci-dessus devient donc

$$\int R(\theta, \delta)\tilde{\pi}(\theta)\, d\theta \geq \int R(\theta, \delta_0)\tilde{\pi}(\theta)\, d\theta$$

et entraîne que δ_0 est un estimateur de Bayes pour $\tilde{\pi}$. □

Dans le cas où Θ n'est pas compact, nous avons déjà vu des exemples où les estimateurs de Bayes ne peuvent former une classe complète. Ainsi, lorsqu'on s'intéresse à la moyenne θ d'une variable aléatoire normale $x \sim \mathcal{N}(\theta, 1)$, l'estimateur $\delta(x) = x$ est admissible mais n'est pas un estimateur de Bayes. Cela dit, dans bien des cas, les classes complètes sont tout de même constituées d'estimateurs de Bayes généralisés (en regroupant, bien sûr, les estimateurs *de Bayes et de Bayes généralisés*). Par exemple, Berger et Srinivasan (1978) démontrent, dans le cadre de l'estimation du paramètre naturel θ d'une famille exponentielle

$$x \sim f(x|\theta) = e^{\theta \cdot x - \psi(\theta)} h(x), \quad x, \theta \in \mathbb{R}^k,$$

sous coût quadratique, que tout estimateur admissible est un estimateur de Bayes généralisé. Il s'agit donc d'une extension de Brown (1971), qui avait traité le cas normal.

Exemple 8.35. Dans le cas normal, $x \sim \mathcal{N}_p(\theta, I_p)$, nous avons parlé à plusieurs reprises de l'estimateur tronqué de James-Stein,

$$\delta^{\mathrm{JS}}(x) = \left(1 - \frac{p-2}{||x||^2}\right)^+ x. \tag{8.11}$$

Bien que satisfaisant, cet estimateur n'est pas admissible. En effet, s'il l'était, ce serait un estimateur de Bayes généralisé, ce qui est impossible puisque la fonction $\delta^{\mathrm{JS}}(x)$ n'est pas analytique (voir l'Exercice 8.26). ‖

Chow (1987) montre un résultat analogue pour les familles avec paramètres de non-centralité, $\chi^2_p(\lambda)$ et $\mathcal{F}_{p,q}(\lambda)$, illustrant ainsi la complétude des règles de Bayes généralisées en dehors du cadre des familles exponentielles. Ce théorème de classe complète a pour conséquence particulière l'inadmissibilité de l'estimateur classique $(x-p)^+$, pour la distribution $\chi^2_p(\lambda)$, bien que Saxena et Alam (1982) aient prouvé l'efficacité de cet estimateur, dans la mesure où il domine l'estimateur du maximum de vraisemblance (voir également l'Exercice 3.25).

Fraisse *et al.* (1990) établissent un résultat similaire à Berger et Srinivasan (1978) en présence d'un *paramètre de nuisance*. Soit $x = (u, z)$ avec $u \in \mathbb{R}^k$ et $z \in \mathbb{R}$. La densité de x par rapport à ν est

$$f(x|\theta, \delta) = h(u, z) e^{\theta . u + \delta z - \psi(\theta, \delta)},$$

avec $\theta \in \Theta \subset \mathbb{R}^k$ et $\delta \in \Delta$, intervalle compact de $\mathbb{R}^*_+$. Comme dans le cas normal, le problème posé consiste à estimer θ/δ sous coût quadratique. Pour ce modèle, le théorème de la classe complète est donné par :

Proposition 8.36. *Si φ est un estimateur admissible de θ/δ, il existe une mesure π sur $\Theta \times \Delta$ telle que, pour ν-presque tout (u, z),*

$$\varphi(u,z) = \frac{\int_{\Theta\times\Delta} \theta e^{\theta.u+\delta z-\psi(\theta,\delta)}\pi(d\theta,d\delta)}{\int_{\Theta\times\Delta} \delta e^{\theta.u+\delta z-\psi(\theta,\delta)}\pi(d\theta,d\delta)} \ . \tag{8.12}$$

Par conséquence, le théorème de classe complète de Berger et Srinivasan (1978) n'est pas invalidé en présence de paramètres de nuisance. La preuve de la Proposition 8.36 repose en fait sur la Proposition 8.30 (voir l'Exercice 8.27).

Exemple 8.37. Soient $x \sim \mathcal{N}_p(\theta, \sigma^2 I_p)$ et $s^2 \sim \sigma^2\chi^2_q$, indépendant de x. Dans ce cadre, $\delta_0(x, s^2) = x$ est aussi inadmissible pour $p \geq 3$. Nous considérons des extensions de l'estimateur de James-Stein (8.11) de la forme

$$\varphi(x, s^2) = (I_p - h(||x||^2_C, s^2)B)x,$$

où B et C sont des matrices $(p \times p)$, h est différentiable presque partout et $||x||^2_C = x^tCx$. On appelle ces estimateurs *estimateurs à rétrécisseur matriciels* (voir Judge et Bock, 1978). La Proposition 8.36 implique qu'une condition nécessaire d'admissibilité sur φ est que h soit infiniment différentiable et que B et C soient proportionnelles (voir l'Exercice 8.28). ‖

Brown (1988) considère le problème d'estimation de la moyenne d'une famille exponentielle, $\xi(\theta)$. En dimension un, il établit que les estimateurs admissibles ont une expression intégrale proche de (8.12). En fait, les estimateurs admissibles sont alors égaux par intervalles à des estimateurs de Bayes généralisés.

Dans le cas des distributions à *support discret*, la complétude des estimateurs de Bayes généralisés n'est pas toujours vraie et les classes complètes font intervenir des procédures bayésiennes par morceaux (voir Berger et Srinivasan, 1978, Brown, 1981, et Brown et Farrell, 1985). Les résultats sur les classes complètes obtenus en Section 5.4 pour le test sous coût quadratique sont de ce type, puisque nous avons vu que les estimateurs admissibles sont identiques aux estimateurs de Bayes généralisés sur des intervalles de *troncature*.

8.5 Conditions nécessaires d'admissibilité

Lorsqu'on ne dispose pas d'un théorème de classe complète limitant le choix d'estimateurs à l'ensemble des estimateurs de Bayes généralisés, il faut trouver une autre façon d'exclure le plus d'estimateurs inadmissibles possibles. Bien que nécessaire, la condition de Stein n'est pas, en général, un outil très utile pour une telle tâche d'élimination, car son intérêt pratique principal réside dans la condition suffisante de Blyth. Par ailleurs, les résultats

de la Section 8.3 ne sont pas applicables dans ce contexte général puisqu'ils ne concernent que les estimateurs de Bayes généralisés. Pour les coûts quadratiques, Hwang (1982b) développe une technique proposée par Brown (1971), appelée STUB (pour *semi-tail upper bounds*, bornes inférieures de demi-queue), qui donne une condition nécessaire d'admissibilité utilisable en pratique. Elle trouve sa source dans le lemme suivant.

Lemme 8.38. *Soient deux estimateurs δ_1 et δ_2 à risques finis, tels que*

$$R(\theta,\delta_1) = \mathbb{E}_\theta\left[(\delta_1(x)-\theta)^t Q(\delta_1(x)-\theta)\right] < R(\theta,\delta_2)$$

pour tout $\theta \in \Theta$ et pour une matrice définie positive donnée Q. Alors, tout estimateur δ satisfaisant presque partout l'inégalité

$$\delta(x)^t Q(\delta_1(x)-\delta_2(x)) < \delta_2(x)^t Q(\delta_1(x)-\delta_2(x))$$

est inadmissible sous n'importe quel coût quadratique.

Preuve. On considère le nouvel estimateur $\delta'(x) = \delta(x)+\delta_1(x)-\delta_2(x)$. Alors

$$\begin{aligned}
R(\theta,\delta') &= \mathbb{E}_\theta\left[(\delta'(x)-\theta)^t Q(\delta'(x)-\theta)\right] \\
&= R(\theta,\delta) + 2\mathbb{E}_\theta\left[(\delta_1(x)-\delta_2(x))^t Q(\delta(x)-\theta)\right] \\
&\quad +\mathbb{E}_\theta\left[(\delta_1(x)-\delta_2(x))^t Q(\delta_1(x)-\delta_2(x))\right] \\
&\le R(\theta,\delta) + 2\mathbb{E}_\theta\left[(\delta_1(x)-\delta_2(x))^t Q(\delta_2(x)-\theta)\right] \\
&\quad +\mathbb{E}_\theta\left[(\delta_1(x)-\delta_2(x))^t Q(\delta_1(x)-\delta_2(x))\right] \\
&= R(\theta,\delta) + R(\theta,\delta_1) - R(\theta,\delta_2) < R(\theta,\delta)
\end{aligned}$$

et δ' domine δ. □

Ce lemme peut sembler simple à première vue mais il est en fait relativement puissant puisqu'il donne une nouvelle condition nécessaire d'admissibilité pour *tout* couple (δ_1,δ_2) ordonné (par le risque). De plus, comme l'admissibilité ne dépend pas de la matrice Q, on obtient une gamme étendue de critères d'inadmissibilité. Elle couvre en particulier la condition nécessaire d'admissibilité (i) du Théorème 8.17.

Exemple 8.39. On considère $x \sim \mathcal{N}_p(\theta, I_p)$. Il découle de James et Stein (1961) (voir la Note 2.8.2) que, parmi tous les estimateurs

$$\delta_a(x) = \left(1 - \frac{a}{||x||^2}\right) x,$$

δ_{p-2} est optimal pour les coûts quadratiques usuels. Par conséquent, le Lemme 8.38 implique que tout estimateur δ vérifiant

$$\delta(x)^t x \frac{a-(p-2)}{||x||^2} \leq \left(1 - \frac{a}{||x||^2}\right)(a-(p-2)) \tag{8.13}$$

est inadmissible. On considère l'estimateur δ de la forme

$$\delta(x) = \left(1 - \frac{h(x)}{||x||^2}\right) x.$$

Alors (8.13) implique que δ est inadmissible si

$$h(x) \leq a < p-2 \qquad \text{ou} \qquad h(x) \geq a > p-2.$$

Donc, tout estimateur tel que h soit uniformément plus grand ou plus petit que $(p-2)$ est inadmissible. La condition nécessaire du Théorème 8.17 s'obtient en considérant ensuite les estimateurs tronqués ($a \leq p-2$),

$$\varphi_a(x) = \left(1 - \frac{a}{||x||^2}\mathbb{I}_{[K,+\infty[}(||x||^2)\right) x\,,$$

et en montrant que $a^* = p-2$ correspond à l'estimateur optimal de cette classe (voir l'Exercice 8.20). Le Lemme 8.38 implique alors que, si

$$h(x) \leq a < p-2$$

pour $||x||^2 > K$, alors l'estimateur δ est également inadmissible. ‖

Remarquons aussi que dans le Lemme 8.38, l'inégalité *stricte* $R(\theta,\delta_1) < R(\theta,\delta_2)$ n'a pas à être satisfaite pour tout θ, mais seulement pour certains θ, du moment que $R(\theta,\delta_1) \leq R(\theta,\delta_2)$ est vérifiée quel que soit $\theta \in \Theta$.

Exemple 8.40. Das Gupta (1958) déduit du Lemme 8.38 une condition nécessaire d'admissibilité pour les distributions exponentielles. Si $x_1,\ldots,x_p$ sont des variables aléatoires $\mathcal{E}xp(\theta_i)$, tout estimateur δ de $(\theta_1^{-1},\ldots,\theta_p^{-1})$ satisfaisant

$$\sum_{i=1}^p x_i^{-3}\delta_i(x) \leq \sum_{i=1}^p x_i^{-3}\delta_{c,i}^{\mathrm{B}}(x)$$

pour $x_i \leq M$, $x=(x_1,\ldots,x_n)$, et

$$\delta_{c,i}^{\mathrm{B}}(x) = \frac{x_i}{2}\left[1 + \frac{c x_i^{-4}}{2\left(\sum_{j=1}^p x_j^{-2}\right)^2}\right],$$

$0 < c < 2(p-1)$, est inadmissible. L'estimateur $\delta_{c,i}^{\mathrm{B}}$ a été suggéré par Berger (1980b) pour améliorer l'estimateur habituel, $x/2$, pour $p \geq 2$. Constatons que $x/2$ domine l'estimateur du maximum de vraisemblance, x (voir l'Exercice 8.32). ‖

Il est également possible d'établir une condition nécessaire d'admissibilité, à partir du Lemme 8.38, pour l'estimation d'un vecteur moyen normal quand la variance est connue à un facteur multiplicatif près, σ^2, comme dans l'Exemple 8.37. Soient $x \sim \mathcal{N}_p(\theta, \sigma^2 I_p)$ et $s^2 \sim \sigma^2 \chi^2_q$ une observation indépendante de x de σ^2. Le résultat suivant donne une condition nécessaire d'admissibilité (Robert, 1998).

Proposition 8.41. *Si, étant donné l'estimateur*

$$\delta(x) = (1 - h(||x||^2, s^2))x,$$

il existe α, M_1 *et* M_2 *tels que*

(i) *pour* $t \geq M_1$ *et* $u \leq M_2$,

$$\frac{t}{u} h(t,u) \leq \alpha < \frac{p-2}{q+2};$$

ou

(ii) *pour* $t \leq M_1$ *et* $u \geq M_2$,

$$\frac{t}{u} h(t,u) \geq \alpha > \frac{p-2}{q+2};$$

δ *est inadmissible sous coût quadratique.*

La démonstration de ce résultat utilise l'existence d'un estimateur optimal dans la classe

$$\varphi_c(x, s^2) = x - \frac{cs^2}{||x||^2} \mathbb{I}_A(||x||^2, s^2)x,$$

avec $A = [K_1, +\infty) \times [0, M_2]$ ou $A = [0, K_1] \times [M_2, +\infty)$. Pour étayer la Proposition 8.41, rappelons que, dans ce cadre, les estimateurs de James-Stein sont de la forme :

$$\delta^{\mathrm{JS}}_a(x, s^2) = \left(1 - \frac{as^2}{||x||^2}\right) x$$

et que

$$a^* = \frac{p-2}{q+2}$$

donne un estimateur optimal dans la classe δ^{JS}_a. Par conséquent, $\delta^{\mathrm{JS}}_{a^*}$ correspond au facteur de rétrécissement minimal pour $||x||^2/s^2$ grand et le rétrécissement maximal pour $||x||^2/s^2$ petit. L'estimateur $\delta^{\mathrm{JS}}_{a^*}$ est néanmoins inadmissible (Exemple 8.35). Fraisse *et al.* (1998) étendent ce résultat aux familles exponentielles avec un paramètre de nuisance de la même façon que dans la Proposition 8.36.

Exemple 8.42. (Suite de l'Exemple 8.37) Parmi les estimateurs à rétrécisseur matriciels, les seuls estimateurs d'intérêt sont de la forme

$$\varphi(x, s^2) = (I_p - h(x^t Bx, s^2)B)x, \tag{8.14}$$

puisque les autres sont inadmissibles. La Proposition 8.41 entraîne que, si, quel que soit (t, u),

$$\frac{t}{u}h(t, u) \leq \alpha < \frac{p-2}{q+2},$$

ces estimateurs sont également inadmissibles. De plus, une condition nécessaire de minimaxité sous coût quadratique est

$$\frac{t}{u}h(t, u) \leq 2\,\frac{\mathrm{tr}(B) - 2\lambda_{\max}(B)}{\lambda_{\max}(B)}\,\frac{1}{q+2},$$

où $\mathrm{tr}(B)$ désigne la trace et $\lambda_{\max}(B)$ la plus grande valeur propre de B (voir Brown, 1975 , et Cellier *et al.*, 1989). Une condition nécessaire pour l'existence d'un estimateur vérifiant à la fois les critères d'admissibilité et de minimaxité est donc

$$\mathrm{tr}(B) > \lambda_{\max}(B)\ \frac{p+2}{2},$$

ce qui exclut de fait les estimateurs rétrécissant vers des sous-espaces de faibles dimensions.

Ce résultat met aussi en évidence le fait que l'admissibilité et la minimaxité ne sont pas totalement compatibles. En fait, un estimateur admissible sous coût quadratique l'est pour n'importe quel coût quadratique. À l'inverse, Brown (1975) montre que le seul estimateur de la forme (8.14), qui est minimax pour tous les coûts quadratiques, est $\delta_0(x) = x$. Cela est également lié à l'admissibilité universelle de l'estimateur δ_0 établie dans Brown et Hwang (1989) (voir la Section 2.6). ‖

8.6 Exercices

Section 8.2.1

8.1 (Lehmann, 1986) Soit une variable aléatoire x de moyenne μ et de variance σ^2.

a. Montrer que $\delta(x) = ax + b$ est un estimateur inadmissible de μ sous coût quadratique si
   (a) $a > 1$; ou
   (b) $a < 0$; ou
   (c) $a = 1$ et $b \neq 0$.

b. Généraliser au cas où $\delta(x) = (1 + h(x))x$ avec $h(x) > 0$.

8.2 (Suite de l'Exercice 8.1) En déduire qu'il suffit de considérer $\lambda \geq 0$ pour les estimateurs (8.1) utilisés dans le Théorème 8.7.

8.3 On considère $x \sim \mathscr{U}_{[-\theta,\theta]}$ et $\pi(\theta)$ est la distribution uniforme $\mathscr{U}_{[0,1]}$.

a. Montrer que

$$\delta_1^\pi(x) = \begin{cases} \dfrac{1-|x|}{\log(1/|x|)} & \text{si } |x| \leq 1, \\ 0 & \text{sinon,} \end{cases}$$

est un estimateur de Bayes inadmissible et dominé, sous coût quadratique usuel, par

$$\delta_2^\pi(x) = \begin{cases} \delta_1^\pi(x) & \text{si } |x| \leq 1, \\ |x| & \text{sinon,} \end{cases}$$

b. Montrer que δ_2^π est aussi un estimateur de Bayes de π.

8.4 Soit $x \sim \mathscr{B}(n,p)$. Déterminer si $\delta_0 \equiv 0$ est un estimateur admissible de p sous coût quadratique.

8.5 (Johnson, 1971) On considère $x \sim \mathscr{B}(n,\theta)$.

a. Montrer que $\delta_0(x) = x$ est l'estimateur du maximum de vraisemblance de θ et également un estimateur de Bayes sous coût quadratique pour $\pi(\theta) = 1/\theta(1-\theta)$.

b. Montrer que $(\delta_0, 1-\delta_0)$ est admissible sous le coût

$$\mathrm{L}(\theta,\delta) = (\theta-\delta_1)^2 + (1-\theta-\delta_2)^2. \tag{8.15}$$

(*Indication :* Utiliser la représentation bayésienne de δ_0 pour montrer que

$$\int [R(\theta,\delta) - R(\theta,(\delta_0,1-\delta_0))] \frac{d\theta}{\theta(1-\theta)} \geq 0$$

et en déduire que le seul cas d'égalité est $\delta_1 = \delta_0$, $\delta_2 = 1-\delta_0$.)

c. Montrer qu'une classe complète pour le coût (8.15) est constituée des estimateurs tels que $\delta_1 = 1-\delta_2$.

d. Généraliser le résultat b. au cas multinomial $x \sim \mathscr{M}_k(n, p_1, \ldots, p_k)$. (*Indication :* Procéder par récurrence.)

Section 8.2.2

8.6 Déterminer les lois a priori bêta $\mathscr{B}e(\alpha,\beta)$ correspondant aux estimateurs admissibles de l'Exemple 8.9.

Section 8.2.3

8.7 Dans le cadre de l'Exemple 8.12, montrer que le risque de Bayes de δ^π est infini et déterminer si δ_{c^*} est un estimateur de Bayes.

8.8 *(Zidek, 1970) Pour $x \sim f(x|\theta)$, $\theta \in \mathbb{R}$, tel que $\{\theta; f(x|\theta) > 0\}$ soit un intervalle, on étudie l'estimation de $g(\theta)$ sous coût quadratique. On cherche une condition suffisante d'admissibilité pour l'estimateur de Bayes généralisé

$$\delta^\pi(x) = \frac{\int g(\theta) f(x|\theta)\pi(\theta)\,d\theta}{\int f(x|\theta)\pi(\theta)\,d\theta}$$

avec π une mesure et

$$\int R(\theta,\delta^\pi)\pi(\theta)\,d\theta = +\infty.$$

a. On définit

$$M(x,\theta) = \int_{\theta}^{+\infty} [g(t) - \delta^{\pi}(x)]^2 f(x|t)\pi(t)\,dt$$

et

$$h(\theta) = \int \left[\frac{M(x,\theta)}{f(x|\theta)\pi(\theta)}\right]^2 f(x|\theta)\,dx.$$

Montrer qu'il existe une fonction $q(\theta)$ telle que $\tilde{\pi}(\theta) = q(\theta)\pi(\theta)$ soit une densité de probabilités et que

$$\int R(\theta,\delta^{\pi})\tilde{\pi}(\theta)\,d\theta < +\infty.$$

b. Soit $\tilde{\delta}$ l'estimateur de Bayes associé à $\tilde{\pi}$. Montrer que

$$r = \int [R(\theta,\delta^{\pi}) - R(\theta,\delta)]\tilde{\pi}(\theta)\,d\theta = \int \frac{[\int q'(\theta)M(x,\theta)\,d\theta]^2}{\int f(x|\theta)\pi(\theta)\,d\theta}dx.$$

c. En désignant $q(\theta)$ par $f^2(\theta)$, utiliser l'inégalité de Cauchy-Schwarz pour montrer que

$$r \le 4\int [f'(\theta)]^2 h(\theta)\pi(\theta)\,d\theta.$$

d. Montrer que si, pour tout (θ_0,θ_1) et $\epsilon > 0$, il existe une fonction q telle que $q(t) = 1$ sur (θ_0,θ_1) et un nombre réel $r < \epsilon$, alors l'estimateur δ^{π} est admissible.

e. On considère la condition (E) : Si

$$\int_t^{+\infty} R(\theta,\delta^{\pi})\pi(\theta)\,d\theta = +\infty, \qquad \text{alors} \quad \int_t^{+\infty} \frac{1}{h(\theta)\pi(\theta)}\,d\theta = +\infty.$$

Soit

$$y(\theta) = \int_{\theta_1}^{\theta} \frac{1}{h(t)\pi(t)}\,dt$$

et

$$f(t) = \left(1 - \frac{y(t)}{F}\right)\mathbb{I}_{0\le y(t)\le F}.$$

Montrer que

$$f'(t) = -\frac{1}{Fh(t)\pi(t)} \qquad (0 \le y(t) \le F),$$

et que

$$\int_{\theta_1}^{+\infty} [f'(t)]^2 h(t)\pi(t)\,dt = \frac{1}{F}.$$

Déduire de (E) qu'il est possible de choisir F tel que $r < \epsilon$. Conclure en donnant une condition suffisante d'admissibilité.

f. Recommencer la question e. sous l'hypothèse symétrique, c'est-à-dire si

$$\int_{-\infty}^{t} R(\theta,\delta^{\pi})\pi(\theta)\,d\theta = +\infty, \quad \text{alors} \quad \int_{-\infty}^{t} \frac{1}{h(\theta)\pi(\theta)}\,d\theta = +\infty.$$

8.9 On considère le coût borné

$$\mathrm{L}(\theta,\delta) = 1 - e^{-a(\theta-\delta)^2} \qquad (a > 0),$$

pour l'estimation de θ avec $x \sim \mathcal{N}(\theta,1)$.

a. Déterminer les estimateurs de Bayes associés à l'a priori conjugué $\theta \sim \mathcal{N}(\mu, \tau^2)$.
b. Déterminer les estimateurs de Bayes associés à l'a priori $\pi(\theta) \propto \exp(-\lambda|\theta - \mu|)$.
c. Étudier l'admissibilité de l'estimateur de Bayes généralisé associé à l'a priori de Jeffreys $\pi(\theta) = 1$ quand a varie. (*Indication :* Déterminer si le risque de Bayes est fini et appliquer la méthode de Blyth si nécessaire.)

Section 8.2.4

8.10 Établir la formule de représentation (8.3) et vérifier les égalités de l'Exemple 8.25.

8.11 Montrer que les estimateurs δ_{CZ} proposés en Section 8.2 dans un cadre poissonnien sont effectivement des estimateurs de Bayes généralisés en trouvant les lois a priori correspondantes.

8.12 *(Berger, 1982a) Soit x de loi

$$x \sim f(x|\theta) = h(x)e^{\theta x - \psi(\theta)}$$

pour $x \in [a, b]$. Étant donné deux fonctions positives différentiables, m_0 et d, on pose

$$\delta_0(x) = \frac{m_0'(x)}{m_0(x)} - \frac{h'(x)}{h(x)}, \quad \gamma(x) = 2\frac{d'(x)}{d(x)}, \quad \text{et} \quad \delta(x) = \delta_0(x) + \gamma(x).$$

a. Montrer que, sous coût quadratique,

$$R(\theta, \delta) - R(\theta, \delta_0) = \mathbb{E}_\theta \left(\frac{4}{d(x)} \left[d''(x) + d'(x)\frac{m_0'(x)}{m_0(x)} \right] \right),$$

moyennant certaines conditions de régularité comme

$$\lim_{x \to a} h(x)\gamma(x)e^{\theta x} = \lim_{x \to b} h(x)\gamma(x)e^{\theta x} = 0.$$

b. On suppose que l'une des fonctions

$$g_1(x) = \int_a^x \frac{1}{m_0(y)}dy \qquad \text{ou} \qquad g_2(x) = \int_x^b \frac{1}{m_0(y)}dy$$

est finie sur $[a, b]$. On note g_i cette fonction. Montrer que si, en outre,

$$\mathbb{E}_\theta \left| \frac{d}{dx} \log g_i \right|^2 < +\infty$$

et

$$\lim_{x \to a} h(x)e^{\theta x}\frac{g_i'(x)}{g_i(x)} = \lim_{x \to b} h(x)e^{\theta x}\frac{g_i'(x)}{g_i(x)} = 0,$$

alors δ_0 est inadmissible et dominé par δ pour $\gamma(x) = 2\alpha g_i'(x)/g_i(x)$ si $0 \leq \alpha \leq 1$.
c. Appliquer au cas $x \sim \mathcal{G}(\nu, \theta)$ et

$$\pi(\theta) = \frac{1}{\pi}\frac{1}{1 + \theta^2}.$$

Section 8.2.5

8.13 Montrer que le noyau de transition (8.7) est lié à la mesure stationnaire π. (*Indication :* Montrer que la condition d'équilibre ponctuel s'applique.) En déduire que la chaîne associée est soit récurrente nulle, soit transiente lorsque l'a priori π est impropre.

8.14 Soient $x \sim \mathcal{N}(\theta, 1)$ et $\pi(\theta) \propto \exp\{-b\theta^2/2 + ab\theta\}$.

a. Donner des conditions nécessaires et suffisantes sur (a, b) pour que la distribution a posteriori soit définie. Montrer que, dans ce cas, la distribution a posteriori est normale de moyenne $(x + ab)/(1 + b)$ et de variance $1/(b + 1)$.

b. Montrer que le noyau de transition (8.7) est alors donné par

$$\eta|\theta \sim \mathcal{N}\left(\frac{\theta + ab}{1 + b}, \frac{b + 2}{(1 + b)^2}\right).$$

c. En déduire que la chaîne de Markov est un modèle AR(1) (Section 4.5.2)

$$\theta^{(t+1)} = \frac{1}{1 + b}\theta^{(t)} + \frac{ab}{1 + b} + \frac{\sqrt{b + 2}}{1 + b}\epsilon_t\,.$$

En conclure qu'elle est transiente lorsque $b < 0$ et récurrente si $b = 0$.

Section 8.3.1

8.15 Vérifier que les trois conditions du Lemme 8.23 sont bien satisfaites pour une fonction de coût quadratique,

$$\mathrm{L}(\theta, \delta) = (\delta - \theta)^t Q(\delta - \theta),$$

pour toute matrice définie positive Q.

Section 8.3.2

8.16 *(Clevenson et Zidek, 1975) Soient $(x_1, \ldots, x_n)$ des variables aléatoires indépendantes de Poisson, $x_i \sim \mathcal{P}(\lambda_i)$.

a. Utiliser une suite de lois a priori conjuguées et la méthode de Blyth pour montrer que $\delta_0(x_i) = x_i$ est un estimateur admissible de λ_i sous coût quadratique.

b. Pour $n \geq 2$, montrer que

$$\mathbb{E}_\lambda\left[\sum_{i=1}^n \frac{1}{\lambda_i}\left\{x_i\left(1 + \frac{n-1}{\sum_{i=1}^n x_i}\right)^{-1} - \lambda_i\right\}^2\right] \leq \mathbb{E}_\lambda\left[\sum_{i=1}^n \frac{1}{\lambda_i}(x_i - \lambda_i)^2\right]$$

et en déduire que $\delta_0(x_1, \ldots, x_n) = (x_1, \ldots, x_n)$ est un estimateur inadmissible de $\lambda = (\lambda_1, \ldots, \lambda_n)$. (*Indication :* Minimiser (en λ) $\mathbb{E}_\lambda[\sum_i \lambda_i^{-1}(ax_i - \lambda_i)^2]$ et remplacer la solution a par $\sum_i x_i / \sum_i x_i + n - 1$.)

8.17 Transposer la démarche de l'Exemple 8.27 au cas Poisson : montrer que, si $H_0 :\ \lambda \leq \lambda_0$ et $\varphi(x) = P_{\lambda_0}(X \geq x)$, avec $X \sim \mathcal{P}(\lambda_0)$, alors φ est admissible sous coût quadratique. (*Indication :* Utiliser la condition de Blyth.)

8.18 Reprendre l'Exercice 8.17 avec la distribution gamma, $\mathcal{G}(\nu, \theta)$ et $H_0 :\ \theta \leq \theta_0$.

8.19 Soit $x \sim \mathcal{N}_2(\theta, I_2)$. Déterminer si la condition de Blyth pour l'admissibilité de $\delta_0(x) = x$ est vérifiée par la suite $\pi_n(\theta)$ qui vaut

$$\pi_n(\theta) = \exp\{-||\theta||^2/2n\}.$$

Si cette suite ne convient pas, en proposer une autre.

8.20 *(Hwang et Brown, 1991) On considère $x \sim \mathcal{N}_p(\theta, I_p)$. La région de confiance standard est

$$C_x = \{\theta;\ \|\theta - x\| < c\},$$

avec $P_\theta(\theta \in C_x) = 1-\alpha$. Grâce à la méthode de Blyth, montrer que l'évaluation $\gamma_0(x) = 1 - \alpha$ est admissible sous coût quadratique

$$\mathrm{L}(\theta, \gamma) = \left(\gamma - \mathbb{I}_{C_x}(\theta)\right)^2,$$

pour $p \leq 4$. [*Note :* Robert et Casella, 1993, montrent en outre que cet estimateur constant est inadmissible pour $p \geq 5$. À l'inverse, Hwang et Brown, 1991, établissent, par validité fréquentiste, que γ_0 est admissible quel que soit p (voir la Section 5.5).]

8.21 Dans le cadre de l'Exemple 8.27, montrer que la loi marginale m_n associée à g_n est croissante. Pour $\pi(\theta) = 1/\theta$, montrer que, pour $x \neq 0$,

$$\varphi(x) = \frac{1}{B(x, m-x+1)} \int_0^{\theta_0} t^{x-1}(1-t)^{m-x}\, dt$$

puis traiter le cas $x = 0$.

Section 8.4

8.22 Une classe $\mathscr{C}$ est dite *complète minimale* si $\mathscr{C}$ est complète et si aucun sous-ensemble propre de $\mathscr{C}$ n'est complet.

a. Montrer que toute classe complète contient tous les estimateurs admissibles.

b. Montrer que, s'il existe une classe complète minimale, il s'agit exactement des estimateurs admissibles.

8.23 *(Karlin et Rubin, 1956) On suppose que $f(x|\theta)$ satisfait la propriété des rapports de vraisemblance monotones (en $x \in \mathbb{R}$), avec $\theta \in \Theta$. Le problème d'estimation est dit *monotone* si $\mathrm{L}(\theta, \delta)$ est minimal pour $\delta = q(\theta)$, avec q croissante en θ, et si $\mathrm{L}(\theta, \delta)$ est une fonction croissante de $|\delta - q(\theta)|$.

a. Montrer que, si L est convexe, les estimateurs qui sont des fonctions croissantes de x constituent une classe complète.

b. Montrer que, si δ_0 n'est pas monotone, l'estimateur monotone δ_{M}, défini par

$$P_{q^{-1}(a)}(\delta_{\mathrm{M}}(X) \leq a) = P_{q^{-1}(a)}(\delta_0(X) \leq a), \quad \forall a$$

domine δ_0.

c. Si δ_{M} est strictement croissante, montrer que la relation ci-dessus implique que $\delta_{\mathrm{M}}(x)$ est un nombre a tel que

$$F(x|q^{-1}(a)) = P_{q^{-1}(a)}(\delta_0(X) \leq a).$$

8.24 Appliquer l'Exercice 8.23 au cas où $x \sim \mathcal{N}(\theta, 1)$, $\mathrm{L}(\theta, \delta) = (\theta - \delta)^2$ et $\delta_0(x) = -cx + b$, avec $c > 0$.

8.25 (Berger, 1985b) Soit Θ un ensemble fini de cardinal p. On suppose que l'ensemble de risque $\mathscr{R}$ est borné et fermé inférieurement. On note $\Gamma(\mathscr{R})$ la frontière inférieure de $\mathscr{R}$, c'est-à-dire

$$\Gamma(\mathscr{R}) = \{r \in \mathscr{R};\ \nexists\, r' \in \mathscr{R},\ r' \neq r \text{ et } r'_i \leq r_i,\ 1 \leq i \leq p\} \subset \mathscr{R}.$$

Le coût L est supposé convexe.

a. Montrer que l'ensemble des estimateurs dont le vecteur de risque est dans $\Gamma(\mathscr{R})$ forme une classe complète minimale.

b. Montrer que l'ensemble des estimateurs de Bayes forme une classe complète et que l'ensemble des estimateurs de Bayes généralisés forme une classe complète minimale.

c. *Généraliser au cas où L n'est pas convexe.

8.26 *(Berger et Srinivasan, 1978) On considère $x \sim \mathscr{N}_p(\theta, \Sigma)$ avec Σ connu. La moyenne θ est estimée sous coût quadratique. Montrer qu'un estimateur δ_0 est un estimateur de Bayes généralisé si et seulement si

(i) $g(x) = \Sigma^{-1}\delta_0(x)$ est continûment différentiable, avec un jacobien symétrique $J_g(x) = \nabla\nabla^t g(x)$; et

(ii) pour $g(x) = \nabla r(x)$, $\exp\{r(x)\}$ peut être exprimée comme une transformée de Laplace.

8.27 *(Fraisse *et al.*, 1990) Soit $x = (u, z)$ avec $u \in \mathbb{R}^k$ et $z \in \mathbb{R}$, de densité

$$f(x|\theta, \delta) = \exp\{\theta \cdot u + \delta z - K(\theta, \delta)\}$$

par rapport à une mesure ν σ-finie, avec $\theta \in \Theta \subset \mathbb{R}^k$ et $\delta \in \Delta$, sous-ensemble compact de $\mathbb{R}_+^*$.

a. Montrer (ou admettre) le lemme suivant : *Si* (μ_n) *est une suite de mesures à supports finis telles que, pour presque tout* (u, z),

$$\sup_n \|\nabla\psi_{\mu_n}(z)\| < +\infty,$$

alors il existe une mesure μ *et une sous-suite* (n_k) *telles que*

$$\lim_{k\to+\infty} \psi_{\mu_{n_k}}(u, z) = \psi_\mu(u, z) \quad \textit{et} \quad \lim_{k\to+\infty} \nabla\psi_{\mu_{n_k}}(u, z) = \nabla\psi_\mu(u, z),$$

avec

$$\psi_\mu(u, z) = \int_{\Theta\times\Delta} e^{\theta\cdot u+\delta z}\mu(d\theta, d\delta).$$

b. Déduire de la Proposition 8.30 que, pour tout estimateur admissible φ de θ/δ sous coût d'erreur quadratique $\delta^2||\varphi - \theta/\delta||^2$, il existe une suite de mesures (ϱ_n) à supports finis sur $\Theta \times \Delta$ telles que

$$\varphi(u, z) = \lim_{n\to+\infty} \frac{\int \theta e^{\theta\cdot u+\delta z}\mu_n(d\theta, d\delta)}{\int \delta e^{\theta\cdot u+\delta z}\mu_n(d\theta, d\delta)},$$

avec $\mu_n(d\theta, d\delta) = e^{-K(\theta,\delta)}\varrho_n(\theta, \delta)$.

c. Montrer que l'hypothèse du lemme ci-dessus est satisfaite et que, pour tout estimateur admissible φ, il existe μ_0 tel qu'on ait presque partout

$$\varphi(u, z) = \frac{\int \theta e^{\theta\cdot u+\delta z}\mu_0(d\theta, d\delta)}{\int \delta e^{\theta\cdot u+\delta z}\mu_0(d\theta, d\delta)},$$

c'est-à-dire que φ est un estimateur de Bayes généralisé associé à μ_0.

8.28 (Fraisse *et al.*, 1990) Soient $x \sim \mathcal{N}_p(\theta, \sigma^2 I_p)$ et $s^2 \sim \sigma^2 \chi^2_q$. La moyenne θ est estimée sous coût quadratique, avec $\sigma \in [a, b]$.

a. Montrer que ce modèle s'insère dans le cadre de l'Exercice 8.26.

b. On considère l'estimateur

$$\varphi(x, s^2) = (I_p - h(x^t Bx, s^2)C)x.$$

Montrer que, si φ est admissible, il existe $\varrho \in \mathbb{R}^*_+$ tel que $B = \varrho C$.

c. Comparer aux résultats de l'Exercice 8.26.

8.29 *(Moors, 1981) Soit $x \sim \mathcal{B}e(p)$, avec $0.2 \leq p \leq 0.8$. On estime le paramètre p sous coût quadratique.

a. Montrer que $\delta^\pi(1) = 1 - \delta^\pi(0)$ lorsque l'a priori $\pi(p)$ est symétrique centré sur $1/2$.

b. Montrer que $\delta^\pi(1) \leq \max_p[1 - 2p(1-p)] = 0.68$.

c. En déduire que, si un estimateur vérifie $\delta(1) = 1 - \delta(0)$ et $\delta(1) > 0.68$, il est inadmissible.

8.30 *(Johnson, 1971) On considère $x \sim \mathcal{B}(n, \theta)$, avec θ à estimer sous coût quadratique.

a. Rappeler pourquoi tout estimateur admissible est nécessairement un estimateur de Bayes.

b. Montrer que la réciproque est fausse, ce qui revient à proposer des estimateurs de Bayes inadmissibles.

c. Montrer que l'ensemble des estimateurs de Bayes admissibles est constitué des estimateurs

$$\delta_\tau(x) = \begin{cases} 0 & \text{si } 0 \leq x \leq \underline{n}, \\ \dfrac{\displaystyle\int_0^1 \theta^{x-\underline{n}}(1-\theta)^{n-x-\overline{n}-1}\, d\tau(\theta)}{\displaystyle\int_0^1 \theta^{x-\underline{n}-1}(1-\theta)^{n-x-\overline{n}-1}\, d\tau(\theta)} & \text{si } \underline{n} < x < \overline{n}, \\ 1 & \text{si } \overline{n} \leq x \leq n\,. \end{cases}$$

d. Expliquer pourquoi δ_τ est un estimateur de Bayes pour une classe entière de distributions a priori τ.

8.31 (Hwang *et al.*, 1992) On considère $H_0 : \theta \in \Theta_0$. Les procédures de test γ sont comparées sous un coût strictement convexe, $\mathrm{L}(\mathbb{I}_{\Theta_0}(\theta), \gamma)$.

a. Montrer que $\gamma_0(x) = 1/2$ est le seul estimateur minimax.

b. En déduire que γ_0 est admissible.

c. Peut-on écrire γ_0 comme un estimateur de Bayes généralisé ? Ce phénomène contredit-il les théorèmes de classe complète (5.42 et 5.43) de la Section 5.4 ?

8.32 Étant donné $x \sim \mathcal{E}xp(\theta)$ et $\delta_c(x) = cx$, déterminer le meilleur estimateur δ_c de θ^{-1} sous coût quadratique. Montrer que cet estimateur est un estimateur de Bayes généralisé et discuter son admissibilité.

Section 8.5

8.33 (Robert and Casella, 1994) On considère $x \sim \mathcal{N}(\theta, 1)$ et l'ensemble de confiance usuel $C_x = [x - c, x + c]$. Au lieu d'utiliser la probabilité fixe de confiance $\alpha = P_\theta(\theta \in C_x)$, on propose une procédure φ qu'on évalue sous le coût

$$\mathrm{L}(\theta, \varphi) = d(\theta, C_x)(\mathbb{I}_{C_x}(\theta) - \varphi)^2,$$

le poids $d(\theta, C_x)$ étant une mesure de distance entre θ et la frontière de C_x.

a. Expliquer la pertinence d'un choix d'intervalle de confiance dépendant des données et justifier le choix du poids de distance.

b. Dans le cas particulier

$$d(\theta, C_x) = 1 - e^{-\omega(\theta - x)^2}\left(1 - e^{-\omega(\theta - x)^2}\right),$$

montrer que cette distance est minimale pour $|\theta - x| = c$ si $\omega = \log(2)/c^2$.

c. Donner la forme générale des estimateurs de Bayes sous cette fonction de coût et montrer que le résultat de classe complète du Théorème 5.42 s'applique dans ce cas.

d. Pour les classes de distances symétriques de la forme $h(|\theta - x|)$ et $\pi(\theta) = 1$, montrer que les estimateurs de Bayes sont constants et pas forcément égaux à α. Ces estimateurs sont-ils admissibles ?

8.34 Déduire la Proposition 8.41 du Lemme 8.38.

8.35 Montrer que, si $x \sim \mathcal{N}_p(\theta, I_p)$, $\delta_0(x) = x$ est admissible pour la classe de coûts

$$\mathrm{L}(\theta, \delta) = (\theta - \delta)^t Q(\theta - \delta),$$

avec Q dans l'ensemble des matrices symétriques définies positives.

8.36 (Hwang, 1982b) Soit $x \sim \mathcal{N}_p(\theta, I_p)$. Une classe d'estimateurs de θ est donnée par

$$\phi_a(x) = x - \frac{a}{||x||^2}\mathbb{I}_{[K, +\infty[}(||x||^2)x,$$

pour $0 \leq a \leq (p - 2)$.

a. Montrer que φ_{a^*} associé à $a^* = p - 2$ est optimal parmi les estimateurs φ_a sous coût quadratique classique.

b. Démontrer le résultat de l'Exemple 8.39.

c. Appliquer la même technique à

$$\varphi_b(x) = x - \frac{b}{||x||^2}\mathbb{I}_{[0, K]}(||x||^2)x$$

et $b \geq (p - 2)$. En déduire une condition STUB.

Note 8.7.1

8.37 *Montrer que K dans (8.7) et K^* dans (8.16) sont soit tous deux récurrents, soit tous deux transients. (*Indication :* Utiliser la variable indicatrice $\sum_{t=1}^{\infty} \mathbb{I}_B(x_t)$ pour un ensemble arbitraire B.)

8.38 *Dans le cadre de l'Exemple 8.43,

a. Montrer que l'équation (8.17) est vérifiée.

b. Montrer que la chaîne de Markov associée à K^* peut s'écrire $x_{t+1} = (x_t + b)z_{t+1}$, où les z_t sont indépendants et de densité

$$f(z) = \frac{\Gamma(2\alpha + a)}{\Gamma(\alpha + a)\Gamma(\alpha)} \frac{z^{\alpha-1}}{(z+1)^{2\alpha+a}} .$$

c. Montrer que z_t a une moyenne infinie quand $a + \alpha \leq 1$ et que $\mathbb{E}[\log(z_t)]$ est négative quand $a < 0$, nulle si $a = 0$, et positive quand $a < 0$.

d. Dans le cas $b = 0$, montrer que (x_t) est récurrente si et seulement si $a = 0$.

8.39 *(Hobert et Robert, 1999) Soit $x \sim \mathscr{P}(\theta)$ avec $\pi(\theta) \propto \theta^{a-1} \exp\{-b\theta\}$.

a. Énoncer des conditions nécessaires et suffisantes sur (a, b) pour que la distribution a posteriori soit définie et que la distribution a priori soit impropre.

b. Expliciter le noyau de transition (8.7) dans ce cas. Si $b = -1/2$ et $a = k/2$, montrer que la transition est une loi du khi deux décentré.

c. Montrer que le noyau de transition (8.16) est

$$K^*(x, y) = \frac{\Gamma(y + x + a)}{y!\Gamma(x + a)} p^{x+a}(1 - p)^y ,$$

avec $p = (b + 1)/(b + 2)$.

d. Montrer que cette distribution correspond à la loi binomiale négative standard lorsque a est un entier naturel.

e. Dans le cas général, la distribution de fonction de masse

$$P(Z = z) = \frac{\Gamma(z + c)}{z!\Gamma(c)} p^c(1 - p)^z$$

est appelée distribution *binomiale négative généralisée* $\mathscr{N}eg(c, p)$. Déduire de la fonction génératrice que, si $z_1, \ldots, z_n$ sont indépendants de $z_i \sim \mathscr{NB}(c_i, p)$, $z_1 + \ldots + z_n \sim \mathscr{N}eg(c_1 + \ldots + c_n, p)$.

f. En déduire que (x_t) associé au noyau K^* est un *processus de branchement*,

$$x_{t+1} = \sum_{i=1}^{x_t} \eta_{i,t} + \omega_{t+1}$$

avec $\eta_{i,t} \sim \mathscr{N}eg(1, p)$ et $\omega_{t+1} \sim \mathscr{N}eg(a, p)$.

g. En conclure que la chaîne (x_t) est récurrente si $b = 0$ et $0 < a < 1$, et transiente sinon.

8.40 *(Hobert et Robert, 1999) Soit $x \sim \mathscr{N}eg(k, \theta)$ avec $\pi(\theta) \propto \theta^{a-1}(1 - \theta)^{b-1}$.

a. Formuler des conditions nécessaires et suffisantes sur (a, b) pour que la distribution a posteriori soit définie et que la distribution a priori soit impropre.

b. Expliciter le noyau de transition (8.7) pour $b = k$.

c. Montrer que la chaîne de Markov correspondante $(\theta^{(t)})$ peut s'écrire $\theta^{(t+1)} = \theta^{(t)}/(\omega_t + \theta^{(t)})$, les ω_t étant i.i.d.

d. Montrer que $\mathbb{E}[\log \omega_t] = 0$ si et seulement si $a = 0$ et en déduire que la chaîne $(\theta^{(t)})$ est récurrente lorsque $a = 0$ et transiente sinon ;

8.41 Pour le noyau de transition (8.18),

a. Montrer que la mesure stationnaire est $\pi(\theta)\Psi(\theta)$, où π est la distribution a priori et $\Psi(\theta)$ est la constante de normalisation en $T(\theta,\eta)$. (*Indication :* Prouver que la condition de balance ponctuelle est vérifiée et utiliser l'égalité $\pi(\theta)f(x|\theta)\pi(\eta|x) = \pi(\eta)f(x|\eta)\pi(\theta|x)$.)

b. En déduire que la chaîne est récurrente positive lorsque le risque de Bayes

$$\int\int \left(\varphi(\theta) - \mathbb{E}[\varphi(\theta)|x]\right)^2 f(x|\theta)\pi(\theta)dxd\theta$$

est fini.

8.7 Notes

8.7.1 Compléments sur la condition suffisante d'admissibilité d'Eaton

Hobert et Robert (1999) montrent que la condition suffisante de Eaton (1992) s'applique également à une *chaîne duale*, avec noyau de transition

$$K^*(x,y) = \int_\Theta f(x|\theta)\pi(\theta|x)\,d\theta, \tag{8.16}$$

puisque les deux noyaux K dans (8.7) et K^* sont de même nature, c'est-à-dire qu'ils sont soit tous deux récurrents, soit tous deux transients. Ce résultat de dualité est particulièrement intéressant lorsque l'espace d'échantillonnage est plus simple que l'espace des paramètres, par exemple lorsque K^* porte sur un espace d'états fini. (Cette propriété a été utilisée dans un contexte complètement différent par Diebolt et Robert, 1994, pour affiner les propriétés de convergence d'un échantillonneur de Gibbs dans des modèles à variables latentes. Voir Robert et Casella, 2004, Section 9.2.3.) Hobert et Robert (1999) illustrent l'intérêt de cette condition dans des cadres classiques (Exercices 8.39 et 8.40).

Exemple 8.43. Soit un modèle gamma, $x \sim \mathcal{G}a(\alpha,\theta)$, avec $\pi(\theta) \propto \theta^{\alpha-1}\exp\{-b\theta\}$. La loi a posteriori est définie lorsque $b \geq 0$ et $a > -\alpha$. Alors que le noyau de transition K ne peut pas être calculé analytiquement, K^* s'écrit

$$K^*(x,y) = \frac{\Gamma(2\alpha+a)\Gamma(b+x)^{\alpha+a}}{\Gamma(\alpha+a)\Gamma(\alpha)}\,\frac{y^{\alpha-1}}{(x+y+b)^{2\alpha+a}}\,. \tag{8.17}$$

Hobert et Robert (1999) montrent ensuite que ce noyau est récurrent si, et seulement si, $a = 0$ (Exercice 8.38). ‖

Eaton (1999) généralise le résultat publié par Eaton (1992) à des fonctions arbitraires de θ, $\varphi(\theta)$, estimées sous coût quadratique, à l'aide d'une autre représentation markovienne. Au lieu du $K(\theta|\eta)$ défini par (8.7), Eaton (1999) propose d'utiliser le noyau de transition

$$T(\theta|\eta) = \Psi(\eta)^{-1}(\varphi(\theta) - \varphi(\eta))^2 K(\theta|\eta)\,, \tag{8.18}$$

où $\Psi(\theta)$ est le facteur de normalisation de cette densité. Un équivalent du Théorème 8.19 dans ce cas est que l'estimateur de Bayes $\mathbb{E}^\pi[\varphi(\theta)|x]$ est admissible lorsque la chaîne de Markov associée à T est récurrente. Bien que la procédure ne soit pas complètement générale, puisque une étude de la chaîne

de Markov est nécessaire pour chaque fonction φ remarquable, l'extension aux fonctions φ non bornées lui confère un intérêt indéniable. (Comme le suggère Eaton, 1999, ce résultat s'étend à l'estimation de fonctions vectorielles $\varphi(\theta)$ sous coût quadratique.)

Exemple 8.44. Dans le cas particulier d'une famille de position, si $f(x|\theta) = g(x-\theta)$ et si θ est estimé sous la fonction de coût $\mathrm{L}(\theta, d) = (\theta - d)^2$, alors le noyau de Markov K associé à l'a priori plat $\pi(\theta) = c$ est

$$K(\theta|\eta) = \int g(x-\theta)g(x-\eta)dx = r(\theta-\eta)\,,$$

par un changement adéquat d'intégrande dans l'intégrale. Le noyau de transition est

$$T(\theta,\eta) \propto (\theta-\eta)^2 r(\theta-\eta) = t(\theta-\eta)$$

et le facteur de proportionnalité est indépendant de θ. Par conséquent, la chaîne de Markov associée à T est une marche aléatoire. Elle est récurrente en dimension un si le premier moment de t existe. Eaton (1999) montre que cette condition est équivalente à l'existence d'un troisième moment de g. ‖

9

Invariance, mesures de Haar et estimateurs équivariants

> "The ring certainly looked like stone, but it felt harder than steel and heavier than lead. And the circle of it was twisted. If she ran a finger along one edge, it would go around twice, inside as well as out ; it only had one edge."
>
> **Robert Jordan**, *The Dragon Reborn.*

9.1 Principes d'invariance

La notion d'invariance a été introduite dans un cadre fréquentiste essentiellement pour réduire de façon considérable le nombre d'estimateurs acceptables, de sorte qu'un estimateur optimal puisse être trouvé. De ce point de vue, cette notion est une alternative au concept d'estimation sans biais et n'est donc pas pertinente pour le paradigme bayésien. Cependant, il existe une autre raison de faire appel à l'invariance qui n'est pas liée à la Théorie de la Décision : on peut exiger des estimateurs recherchés des propriétés de convergence à l'égard d'un certain nombre de transformations et il est alors logique de s'intéresser à cette notion. En outre, les estimateurs optimaux (équivariants) sont toujours de Bayes ou de Bayes généralisés. Les mesures correspondantes peuvent alors être considérées comme des lois a priori non informatives découlant de la structure d'invariance. Au-delà du fait que l'optimalité classique suggère de nouveau des estimateurs de Bayes, c'est donc surtout le lien entre structures d'invariance et distributions non informatives qui nous pousse à étudier l'invariance d'un point de vue bayésien.

Une première version du *principe d'invariance* est de considérer que les propriétés d'une procédure statistique ne devraient pas dépendre de l'*unité de mesure* utilisée. Si x et θ sont mesurés en une unité u_1 et si y et η sont les transformés de x et θ pour une autre unité u_2, alors un estimateur $\delta_2(y)$ de η devrait correspondre à l'estimateur $\delta_1(x)$ de θ par le même changement d'unité. Insistons sur la généralité de cette notion d'*unité de mesure* : par exemple, cela peut être un simple choix d'échelle (*cm* contre *m*)–auquel cas on exige des estimateurs qu'ils soient *équivariants par changement d'échelle*–le choix d'une origine particulière–on parle dans ce cas d'*équivariance par changement de position*–ou encore le choix de l'ordonnancement des observations dans un échantillon x–on se restreint alors aux estimateurs *symétriques*[69].

Exemple 9.1. On considère le problème d'estimation de la vitesse de la lumière, θ, à partir d'une observation x, distribuée selon $\mathscr{U}([\theta-\epsilon,\theta+\epsilon])$, et mesurée en mètres par seconde. Un changement d'unité typique dans ce contexte est le *changement d'échelle*, $y=\tau x$, avec, par exemple, $\tau=10^{-3}$ pour une conversion de mètres en kilomètres. Dans ce cas, $y\sim\mathscr{U}([\eta-\epsilon',\eta+\epsilon'])$ avec $\eta=\tau\theta$, $\epsilon'=\tau\epsilon$, mais η représente toujours la même quantité intrinsèque, à savoir la vitesse de la lumière. Si δ_0 est un estimateur de θ dans l'unité initiale, il semble légitime d'exiger que l'estimateur dans le problème transformé δ^* vérifie la propriété *d'équivariance d'échelle*

$$\delta^*(y)=\tau\delta_0(y/\tau).$$

En outre, on suppose que le coût est le coût quadratique ajusté

$$\mathrm{L}(\theta,d)=\left(1-\frac{d}{\theta}\right)^2.$$

Il satisfait

$$\mathrm{L}(\theta,d)=\left(1-\frac{\tau d}{\tau\theta}\right)^2=\left(1-\frac{d^*}{\eta}\right)^2=\mathrm{L}(\eta,d^*)$$

lorsque $d^*=\tau\,d$. Par conséquent, le coût est invariant par ce changement d'unité et les deux problèmes d'estimation sont *formellement identiques*. Il est alors logique de choisir le même estimateur pour les deux problèmes, $\delta^*(y)=\delta_0(y)$. On obtient en regroupant les deux équations

$$\delta_0(\tau y)=\tau\delta_0(y)$$

quels que soient τ et y. Finalement, les règles de décision compatibles avec les exigences d'invariance sont nécessairement de la forme $\delta_0(x)=ax$, avec a constante positive. ‖

[69] Comme nous avons discuté au Chapitre 2, il est rare de pouvoir obtenir une invariance absolue où l'estimateur de toute transformation est la transformation d'un même estimateur, à moins d'imposer des coûts invariants fonctionnels comme dans la Section 2.5.4.

Ce principe est souvent élargi en un *principe d'invariance formelle*, stipulant que deux problèmes de *structure* formelle identique, $(\mathscr{X}, f(x|\theta), \mathrm{L})$, devraient être traités avec la même règle de décision, qu'ils soient physiquement liés, comme dans le principe d'invariance restreint, ou pas. Dans l'exemple précédent, la vitesse de la lumière *est toujours le même objet*. Cette extension n'est pas forcément naturelle d'un point de vue bayésien dans la mesure où l'information a priori n'a aucune raison d'être identique dans les deux problèmes. C'est donc uniquement dans les cadres non informatifs que les deux approches sont compatibles.

Une approche bayésienne de l'invariance est justifiée par les trois raisons suivantes :

(i) Le meilleur estimateur invariant (ou *équivariant*) est un estimateur de Bayes généralisé pour une mesure particulière, dite *mesure de Haar*.

(ii) Cette mesure convient d'autant mieux dans les contextes non informatifs que l'invariance suggère une méthode alternative pour construire des distributions a priori non informatives.

(iii) La méthode la plus efficace pour obtenir les meilleurs estimateurs équivariants est l'approche bayésienne.

Par conséquent, les considérations d'invariance vont dans le sens du paradigme bayésien, puisqu'il recouvre une fois de plus un critère fréquentiste d'optimalité. Dans ce chapitre, nous détaillons le lien entre invariance et approche bayésienne dans le cas des paramètres de position dans la Section 9.2, avant de présenter le cadre général de l'invariance au cours de la Section 9.3, puis la mesure de Haar en tant qu'a priori non informatif potentiel en Section 9.4 et le théorème de Hunt-Stein, qui relie invariance et minimaxité, en Section 9.5. (Pour des études plus détaillées sur l'invariance et des points de vue généraux ou bayésiens, les lecteurs pourront se reporter à Berger, 1985b, Chapitre 6, Eaton, 1989, et Wijsman, 1990.)

9.2 Le cas particulier des paramètres de position

Soit $(x_1, \ldots, x_n)$ de densité $f(x_1-\theta, \ldots, x_n-\theta)$, avec un *paramètre de position* inconnu $\theta \in \mathbb{R}$. Le degré naturel d'invariance du problème est l'*invariance par translation*. Si $(x_1, \ldots, x_n)$ subit la transformation

$$(y_1, \ldots, y_n) = (x_1 + a, \ldots, x_n + a)\,,$$

la nouvelle variable aléatoire $(y_1, \ldots, y_n)$ est distribuée selon $f(y_1 - \theta - a, \ldots, y_n - \theta - a)$ et $\eta = \theta + a$ est le paramètre de position correspondant. Par conséquent, le vecteur transformé a le même type de densité et le problème est invariant par translation. Il semble naturel de reproduire l'invariance en exigeant la condition suivante sur les estimateurs δ de θ :

$$\delta(x_1 + a, \ldots, x_n + a) = \delta(x_1, \ldots, x_n) + a. \tag{9.1}$$

Cette condition est satisfaite, par exemple, par $\delta_0(x_1, \ldots, x_n) = \bar{x}$. En outre, il paraît également logique d'imposer la même restriction d'invariance sur la fonction de coût, à savoir que $\mathrm{L}(\theta + a, d + a) = \mathrm{L}(\theta, d)$ pour tout a. Une fonction de coût compatible avec la structure d'invariance devrait donc être de la forme

$$\mathrm{L}(\theta, d) = \mathrm{L}(0, d - \theta) = \varrho(d - \theta). \tag{9.2}$$

Les estimateurs vérifiant (9.1) sont appelés *équivariants* et les fonctions de coût répondant à (9.2) *invariantes*, sous l'action d'un groupe de translations. Le but de ces contraintes est de réduire la classe des estimateurs "acceptables" de façon suffisamment significative pour qu'il n'y ait finalement plus qu'un seul *meilleur estimateur équivariant* sous le coût (9.2), puisque cela n'est pas possible autrement (Exercice 2.36). Le lemme suivant explique pourquoi on peut envisager l'unicité d'un estimateur optimal.

Lemme 9.2. *Pour des fonctions de coût de la forme* (9.2)*, les estimateurs équivariants ont un risque constant.*

Preuve. On a

$$\begin{aligned} R(\delta, \theta) &= \mathbb{E}_\theta[\varrho(\delta(x) - \theta)] \\ &= \mathbb{E}_\theta[\varrho(\delta(x_1 - \theta, \ldots, x_n - \theta))] \\ &= \mathbb{E}_0[\varrho(\delta(x_1, \ldots, x_n))] = R(\delta, 0). \end{aligned}$$

□

Par conséquent, on retrouve dans le cas particulier des estimateurs équivariants sous coût équivariant une situation analogue à celle de l'ensemble des estimateurs considérés sous risque bayésien : il existe un *ordre total* sur cette classe restreinte, puisque comparer deux estimateurs est équivalent à comparer deux nombres réels. C'est la raison pour laquelle un meilleur estimateur équivariant peut exister.

Ce meilleur estimateur est classiquement déduit en conditionnant par rapport à une statistique libre maximale, telle que $y = (x_1 - x_n, \ldots, x_{n-1} - x_n)$. On vérifie alors de façon immédiate que tout estimateur équivariant peut s'écrire $\delta_0(x) + v(y)$, où δ_0 est un estimateur équivariant particulier, par exemple $\delta_0(x) = x_n$.

Lemme 9.3. *S'il existe une fonction $v^*(y)$ qui minimise*

$$\mathbb{E}_0[\varrho(\delta_0(x) + v(y))|y],$$

le meilleur estimateur équivariant sous le coût (9.2) *est*

$$\delta^*(x) = \delta_0(x) + v^*(y).$$

Preuve. Par définition, le meilleur estimateur équivariant minimise (en v) le risque constant

$$R(\delta,\theta) = \mathbb{E}_0[\varrho(\delta(x))] = \mathbb{E}_0[\varrho(\delta_0(x) + v(y))].$$

On peut conditionner par rapport à y, puisque c'est une statistique libre, en décomposant le risque en

$$\mathbb{E}_0[\varrho(\delta_0(x) + v(y))] = \mathbb{E}\left[\mathbb{E}_0[\varrho(\delta_0(x) + v(y))|y]\right].$$

Si $v^*(y)$ minimise l'intégrande pour tout y, δ^* minimise le risque dans la classe des estimateurs équivariants. □

Dans le cas particulier où $\varrho(\delta-\theta) = (\delta-\theta)^2$, le facteur v^* optimal est donné par

$$v^*(y) = -\mathbb{E}_0[\delta_0(x)|y].$$

Nous avons ainsi obtenu le meilleur estimateur équivariant de Pitman (1939).

Corollaire 9.4. *Pour le coût quadratique* $\mathrm{L}(\theta,d) = (\theta-d)^2$*, le meilleur estimateur équivariant de* θ *est*

$$\delta^*(x_1,\ldots,x_n) = \frac{\int_{-\infty}^{+\infty} \theta f(x_1-\theta,\ldots,x_n-\theta)\,d\theta}{\int_{-\infty}^{+\infty} f(x_1-\theta,\ldots,x_n-\theta)\,d\theta}.$$

Preuve. On prend $\delta_0(x) = x_n$ en tant qu'estimateur équivariant particulier pour le Lemme 9.3 et on note $y_n = x_n$ pour compléter y. La densité de $(y_1,\ldots,y_n)$ est alors (pour $\theta = 0$)

$$g_Y(y_1,\ldots,y_n) = f(y_1+y_n,\ldots,y_{n-1}+y_n,y_n),$$

car $y_i = x_i - x_n$ $(i \neq n)$ et le déterminant du jacobien est égal à 1. De plus,

$$\begin{aligned}
\mathbb{E}_0[y_n|y_1,\ldots,y_{n-1}] &= \frac{\int_{-\infty}^{+\infty} t f(y_1+t,\ldots,y_{n-1}+t,t)\,dt}{\int_{-\infty}^{+\infty} f(y_1+t,\ldots,y_{n-1}+t,t)\,dt} \\
&= \frac{\int_{-\infty}^{+\infty} t f(x_1-x_n+t,\ldots,x_{n-1}-x_n+t,t)\,dt}{\int_{-\infty}^{+\infty} f(x_1-x_n+t,\ldots,x_{n-1}-x_n+t,t)\,dt} \\
&= x_n - \frac{\int_{-\infty}^{+\infty} \theta f(x_1-\theta,\ldots,x_{n-1}-\theta,x_n-\theta)\,d\theta}{\int_{-\infty}^{+\infty} f(x_1-\theta,\ldots,x_n-\theta)\,d\theta},
\end{aligned}$$

par le changement de variable $\theta = x_n - t$. Avec

$$\delta^*(x_1,\ldots,x_n) = x_n - \mathbb{E}_0[y_n|y_1,\ldots,y_{n-1}],$$

on déduit l'expression donnée ci-dessus pour δ^*. □

L'intérêt principal du Corollaire 9.4 est, outre qu'il fournit le meilleur estimateur équivariant, de présenter cet estimateur comme un estimateur de Bayes, bien que le calcul ne fasse intervenir aucune technique bayésienne. L'estimateur de Pitman est bien un estimateur de Bayes associé à la distribution a priori $\pi(\theta) = 1$, c'est-à-dire la distribution non informative habituelle des paramètres de position (Chapitre 3). Ce résultat est d'ailleurs valable pour d'autres coûts invariants (9.2), comme nous le verrons en Section 9.4. Par conséquent, le meilleur estimateur équivariant peut être déterminé comme l'estimateur δ qui minimise le coût a posteriori

$$\mathbb{E}^{\pi}[\mathrm{L}(\theta,\delta)|x] = \frac{\int_{-\infty}^{+\infty} \varrho(\theta-\delta) f(x_1-\theta,\ldots,x_n-\theta)\,d\theta}{\int_{-\infty}^{+\infty} f(x_1-\theta,\ldots,x_n-\theta)\,d\theta},$$

ou, de manière équivalente,

$$\int_{-\infty}^{+\infty} \varrho(\theta-\delta) f(x_1-\theta,\ldots,x_n-\theta)\,d\theta,$$

et cette représentation simplifie grandement le calcul des meilleurs estimateurs équivariants. Nous verrons dans la Section 9.4 que le lien entre meilleurs estimateurs équivariants et une classe particulière d'estimateurs de Bayes existe de façon beaucoup plus générale que pour le problème d'estimation de paramètres de position.

9.3 Problèmes de décision invariants

Nous présentons à présent une description abstraite du concept d'invariance au moyen de *groupes d'invariance* qui nous permettra de généraliser le lien entre problèmes invariants et analyse bayésienne. Soient un modèle statistique $(\mathscr{X},\ \Theta,\ f(x|\theta))$ et un problème d'inférence sur θ représenté par un *espace de décision*, $\mathscr{D}$. En outre, on suppose donné un *groupe* $\mathscr{G}$ de *transformations sur* $\mathscr{X}$. (On peut aussi le voir comme une forme particulière d'information a priori.) Nous allons maintenant illustrer l'importance de l'existence d'un tel groupe dans plusieurs cas.

Définition 9.5. *Le modèle statistique est dit* invariant *(ou* fermé*) sous l'action du groupe* $\mathscr{G}$ *si, pour tout* $g \in \mathscr{G}$*, il existe un unique* $\theta^* \in \Theta$ *tel que* $y = g(x)$ *soit distribué selon la densité* $f(y|\theta^*)$*. On note* $\theta^* = \bar{g}(\theta)$*.*

Exemple 9.6. On considère $x \sim f(x-\theta)$ et le *groupe des translations*

$$\mathscr{G} = \{g_c;\ g_c(x) = x + c, c \in \mathbb{R}\}\,.$$

Le modèle statistique est invariant sous l'action de $\mathscr{G}$. Ce n'est pas le cas avec le groupe multiplicatif

$$\mathscr{G}' = \{g_c;\ g_c(x) = cx, c > 0\},$$

puisque

$$y = cx \sim \frac{1}{c} f\left(\frac{y - c\theta}{c}\right). \qquad \|$$

Lorsque le groupe $\mathscr{G}$ a une action globalement invariante sur le modèle, on peut définir naturellement un ensemble $\bar{\mathscr{G}}$ de transformations sur Θ. La preuve que $\bar{\mathscr{G}}$ est aussi un *groupe* est laissée aux lecteurs à titre d'exercice. Dans un souci de simplification, nous adopterons les notations suivantes dans ce qui suit : gx pour $g(x)$ et $\bar{g}\theta$ pour $\bar{g}(\theta)$.

On suppose de plus que la fonction de coût associée au modèle, L de $\Theta \times \mathscr{D}$ dans $\mathbb{R}^+$, est discriminante, c'est-à-dire que deux décisions différentes se verront affectées des coûts différents. En outre, on fait l'hypothèse de compatibilité suivante avec la structure d'invariance.

Définition 9.7. *Si le modèle est invariant sous l'action de $\mathscr{G}$, le coût* L *est dit* invariant sous $\mathscr{G}$ *si, quels que soient $g \in \mathscr{G}$ et $d \in \mathscr{D}$, il existe une unique décision $d^* \in \mathscr{D}$ telle que* $\mathrm{L}(\theta, d) = \mathrm{L}(\bar{g}\theta, d^*)$ *pour tout $\theta \in \Theta$. On note cette décision $d^* = \tilde{g}(d)$ et on dit que le problème de décision est* invariant sous $\mathscr{G}$.

Dans ce cas, le groupe $\mathscr{G}$ induit un second groupe $\tilde{\mathscr{G}}$, agissant sur $\mathscr{D}$. Étant donné ces trois groupes $\mathscr{G}$, $\bar{\mathscr{G}}$, et $\tilde{\mathscr{G}}$, et les hypothèses ci-dessus sur le problème de décision, il semble logique de restreindre la classe des estimateurs considérés aux *estimateurs équivariants*, c'est-à-dire à ceux qui satisfont

$$\delta(gx) = \tilde{g}\delta(x).$$

Dans des références plus anciennes, on qualifie parfois également ces estimateurs d'*invariants*. Un cas particulier intéressant est l'estimation de θ, où $\mathscr{D} = \Theta$, puisque $\mathscr{G} = \tilde{\mathscr{G}}$ dans ce contexte.

Exemple 9.8. (Suite de l'Exemple 9.6) On cherche à estimer θ sous coût quadratique $(\theta - d)^2$. Le problème décisionnel est alors invariant et $\bar{\mathscr{G}} = \tilde{\mathscr{G}} = \mathscr{G}$. $\|$

Exemple 9.9. Soit $x \sim \mathscr{N}(0, \sigma^2)$. La variance σ^2 est estimée sous le coût *entropique*,

$$\mathrm{L}(\sigma, \delta) = \frac{\delta}{\sigma^2} - \log(\delta/\sigma^2) - 1,$$

présenté au Chapitre 2. Si on s'intéresse au groupe des *transformations d'échelle*,

$$\mathscr{G} = \{g_c;\ g_c(x) = cx, c > 0\},$$

les groupes associés sont

$$\bar{\mathscr{G}} = \tilde{\mathscr{G}} = \left\{\bar{g}_c(\sigma^2) = c^2\sigma^2, c > 0\right\}$$

et le coût, donc le problème de décision, est également invariant sous l'action de $\mathscr{G}$. ‖

Exemple 9.10. Soient $x \sim \mathscr{T}_p(\nu, \theta, I_p)$ et $||\theta||^2$ le paramètre d'intérêt. Une structure naturelle d'invariance est l'invariance sous *transformations orthogonales*,

$$\mathscr{G} = \bar{\mathscr{G}} = \left\{g_A;\ g_A(x) = Ax,\ A^tA = I_p\right\},$$

et le problème est invariant si le coût peut s'écrire

$$\mathrm{L}(\theta, \delta) = \tilde{\mathrm{L}}(||\theta||^2, \delta),$$

puisqu'il existe toujours une matrice orthogonale A telle que $A\theta = ||\theta||(1, 0, \ldots, 0)^t$ et $\tilde{\mathscr{G}}$ contient juste la transformation identité. Dans ce cas, les estimateurs équivariants dépendent uniquement de $||x||^2$. ‖

Définition 9.11. *Quand $\bar{\mathscr{G}}$ est un groupe agissant sur Θ, θ_1 et θ_2 sont dits* équivalents *s'il existe $\bar{g} \in \bar{\mathscr{G}}$ avec $\theta_2 = \bar{g}\theta_1$. Une* orbite *de Θ est une classe d'équivalence pour cette relation et le groupe $\bar{\mathscr{G}}$ est dit* transitif *si Θ n'a qu'une seule orbite.*

Si le groupe $\mathscr{G}$ est suffisamment petit, il peut y avoir de nombreuses orbites. Par exemple, quand $x \sim \mathscr{B}(n, p)$ et $\mathscr{G}$ est restreint à $g_0(x) = x$ et $g_1(x) = n - x$, $\bar{\mathscr{G}} = \{\bar{g}_0, \bar{g}_1\}$ avec $\bar{g}_1(p) = 1 - p$. Il y a donc une orbite associée à chaque $p \in [0, 0.5]$. Quand $\mathscr{G}$ est plus grand, cette notion permet souvent de généraliser le phénomène observé pour les paramètres de position.

Théorème 9.12. *Le risque d'un estimateur équivariant est constant dans toute orbite de Θ, c'est-à-dire que*

$$R(\delta, \theta) = R(\delta, \bar{g}\theta)$$

quel que soit $g \in \mathscr{G}$.

Preuve. Comme pour les estimateurs de paramètres de position, on a

$$\begin{aligned} R(\delta, \theta) = \mathbb{E}_\theta\left[\mathrm{L}(\theta, \delta(x))\right] &= \mathbb{E}_\theta\left[\mathrm{L}(\bar{g}\theta, \tilde{g}\delta(x))\right] \\ &= \mathbb{E}_\theta\left[\mathrm{L}(\bar{g}\theta, \delta(gx))\right] \\ &= \mathbb{E}_{\bar{g}\theta}\left[\mathrm{L}(\bar{g}\theta, \delta(x))\right] \\ &= R(\delta, \bar{g}\theta) \end{aligned}$$

pour tout $g \in \mathscr{G}$. □

Le résultat suivant est une conséquence immédiate du Théorème 9.12.

Corollaire 9.13. *Si $\bar{\mathscr{G}}$ est transitif, tout estimateur équivariant a un risque constant.*

Pour des groupes transitifs, il est donc légitime de chercher le *meilleur estimateur équivariant* en minimisant le risque *constant* $R(\delta, \theta_0)$ dans la classe des estimateurs équivariants. Néanmoins, il n'est pas pour autant toujours facile d'avoir recours aux *statistiques libres*, comme dans la Section 9.2. Une solution classique est de considérer la statistique *invariante maximale* pour réduire la dimension du problème, de même qu'on utilise des statistiques exhaustives minimales avec des coûts convexes.

Définition 9.14. *Pour un groupe de transformations $\mathscr{G}$, une statistique $T(x)$ est* invariante *si $T(gx) = T(x)$ quels que soient $x \in \mathscr{X}$ et $g \in \mathscr{G}$. On parle de* statistique invariante maximale *si T est invariante et si $T(x_1) = T(x_2)$ implique l'équivalence de x_1 et x_2.*

En d'autres termes, une statistique invariante maximale indexe les orbites de $\bar{\mathscr{G}}$. En particulier, si $\bar{\mathscr{G}}$ est transitif, les seules statistiques invariantes maximales sont constantes. De plus, il est alors clair que *toute statistique invariante est une fonction d'une statistique invariante maximale.* Remarquons enfin que, si $\bar{\mathscr{G}}$ est transitif, $T(x)$ est nécessairement libre.

Exemple 9.15. Soit une distribution munie d'un paramètre d'échelle σ,

$$x = (x_1, \ldots, x_n) \sim \frac{1}{\sigma^n} f\left(\frac{x_1}{\sigma}, \ldots, \frac{x_n}{\sigma}\right),$$

et $\mathscr{G}$ désigne le groupe multiplicatif, constitué des transformations

$$g_c(x_1, \ldots, x_n) = (cx_1, \ldots, cx_n) \qquad (c > 0).$$

Alors, si $z = ||x||$,

$$T(x) = \begin{cases} 0 & \text{si } z = 0, \\ \frac{x}{z} & \text{sinon,} \end{cases}$$

est une statistique invariante maximale. ‖

Exemple 9.16. (Suite de l'Exemple 9.10) De façon analogue, si $z = ||x||$, la statistique

$$T(x) = \begin{cases} 0 & \text{si } z = 0, \\ \frac{x}{z} & \text{sinon,} \end{cases}$$

est également invariante maximale pour ce problème. ‖

Pour déterminer le meilleur estimateur équivariant, on peut faire appel à la statistique invariante maximale par *conditionnement*. (Il faut noter que le choix d'une statistique invariante maximale particulière n'a pas d'importance puisque, toutes les statistiques invariantes maximales étant en bijection, elles génèrent toutes la même σ-algèbre.) En fait, si $\bar{\mathscr{G}}$ est transitif et T est une statistique invariante maximale, tout estimateur équivariant δ satisfait

$$\begin{aligned} R(\delta,\theta) &= R(\delta,\theta_0) \\ &= \mathbb{E}_{\theta_0}[\mathrm{L}(\theta_0,\delta(x))] \\ &= \mathbb{E}^T_{\theta_0}\{\mathbb{E}_{\theta_0}[\mathrm{L}(\theta_0,\delta(x))|T(x)=t]\} \end{aligned}$$

pour une valeur arbitraire de θ_0 (puisque le risque est constant). Comme T est invariante maximale, tout x tel que $T(x)=t$ peut s'écrire gx_t, où x_t est un membre adéquat de l'orbite de x (en supposant l'axiome du choix). Alors, pour un estimateur équivariant, $\delta(x)=\tilde{g}\delta(x_t)$. Il est donc suffisant de minimiser la quantité ci-dessus en $\delta(x_t)$, conditionnellement à T, pour obtenir le meilleur estimateur équivariant. Bien que simple, le conditionnement ci-dessus se révèle essentiel dans la détermination des meilleurs estimateurs équivariants et sera réutilisé dans la suite.

Exemple 9.17. (Suite de l'Exemple 9.15) Remarquons que, dans ce cas, T est également une statistique libre. Pour le coût entropique, le problème de minimisation (en δ) est

$$\begin{aligned} \mathbb{E}_1[\delta(x)-\log\delta(x)|T(x)=t] &= \mathbb{E}_1[\delta(zt)-\log\delta(zt)|T(x)=t] \\ &= \mathbb{E}_1[z\delta(t)-\log\delta(t)-\log(z)|T(x)=t] \end{aligned}$$

(avec $z=||x||$). Par linéarité de l'espérance, $\delta(t)$ minimise

$$\mathbb{E}_1[z|T=t]\delta(t)-\log\delta(t),$$

et vérifie donc

$$\delta^*(t)=\frac{1}{\mathbb{E}_1[z|T=t]}.$$

Le meilleur estimateur équivariant de σ est donc

$$\delta^*(x)=\frac{||x||}{\mathbb{E}_1[z|T=x/||x||]}\,.$$

Dans le cas particulier où $x_i\sim\mathscr{N}(0,\sigma^2)$, on en déduit que le meilleur estimateur équivariant de σ est

$$\delta^*(x)=\frac{||x||}{\mathbb{E}_1(||x||)}=\frac{\Gamma(p/2)}{\sqrt{2}\Gamma(p+1/2)}\,||x||. \qquad \|$$

De plus amples détails sur cette technique figurent dans Berger (1985b, Section 6.5) et Eaton (1989, Section 2.3).

9.4 Meilleurs estimateurs équivariants et distributions non informatives

Nous nous intéressons à présent à la généralisation du résultat de la Section 9.2, obtenu dans le cas particulier des paramètres de position. Nous montrons qu'il est effectivement possible d'établir un lien entre le meilleur estimateur équivariant et une mesure σ-finie sur Θ qui est en fait la *mesure de Haar invariante à droite*. Pour une étude plus détaillée et rigoureuse, les lecteurs pourront consulter Eaton (1989) et Wijsman (1990).

Supposons d'abord que, pour un problème statistique invariant sous l'action de $\mathscr{G}$, il existe une densité de probabilité π^* sur Θ également *invariante sous l'action de* $\bar{\mathscr{G}}$, c'est-à-dire telle que

$$\pi^*(\bar{g}A) = \pi^*(A)$$

pour tout ensemble mesurable de Θ, soit pour tout $A \in \mathscr{B}(\Theta)$, et pour tout $g \in \mathscr{G}$. Dans ce cas, l'estimateur de Bayes associé à π^*, δ^*, minimise

$$\begin{aligned}\int_\Theta R(\delta^*, \theta)\, d\pi^*(\theta) &= \int_\Theta R(\delta^*, \bar{g}\theta)\, d\pi^*(\theta) \\ &= \int_\Theta \mathbb{E}_\theta \left[\mathrm{L}(\theta, \tilde{g}^{-1}\delta^*(gx))\right] d\pi^*(\theta)\end{aligned}$$

et, si l'estimateur de Bayes est unique, il vérifie

$$\delta^*(x) = \tilde{g}^{-1}\delta^*(gx)$$

π-presque partout, l'ensemble de mesure nulle sur lequel l'égalité n'est pas satisfaite dépendant de la fonction g. Par conséquent, *un estimateur de Bayes associé à un a priori invariant et à un coût invariant strictement convexe est presque équivariant*. Lorsque $\mathscr{G}$ n'est pas dénombrable, la réunion (sur tous les g) des ensembles de mesure nulle ci-dessus n'est pas nécessairement de mesure nulle mais il est possible de montrer, moyennant quelques hypothèses supplémentaires (Lehmann, 1986, Chapitre 6, Théorème 4), qu'il existe *un estimateur équivariant qui est un estimateur de Bayes pour* π^* (voir aussi Strasser, 1985).

Exemple 9.18. On considère $\delta^\pi(x) = \mathbb{E}^\pi[\theta|x]$ sous coût propre invariant. Si π^* est une distribution de probabilité invariante, l'estimateur de Bayes associé à π^* vérifie

$$\begin{aligned}\delta^\pi(gx) &= \frac{\int_\Theta \theta f(gx|\theta)\, d\pi^*(\theta)}{\int_\Theta f(gx|\theta)\, d\pi^*(\theta)} \\ &= \frac{\int_\Theta \theta f(x|\bar{g}^{-1}\theta)\, d\pi^*(\theta)}{\int_\Theta f(x|\bar{g}^{-1}\theta)\, d\pi^*(\theta)} \\ &= \frac{\int_\Theta \bar{g}\eta f(x|\eta)\, d\pi^*(\eta)}{\int_\Theta f(x|\eta)\, d\pi^*(\eta)}.\end{aligned}$$

Donc, si

$$\int_\Theta \bar{g}\eta f(x|\eta)\, d\pi^*(\eta) = \bar{g} \int_\Theta \eta f(x|\eta)\, d\pi^*(\eta),$$

quel que soit $g \in \mathscr{G}$, δ^* est bien équivariant. ‖

Les distributions de probabilité invariantes sont plutôt rares en pratique, puisqu'elles ne peuvent exister que sur des groupes compacts[70] $\bar{\mathscr{G}}$ (voir Lehmann, 1983, Chapitre 4, Exemple 4.2 pour un exemple sur un groupe non dénombrable). Dans d'autres contextes, il est nécessaire de considérer des *mesures invariantes*, pour lesquelles les résultats développés ci-dessus ne sont pas toujours vrais (car les estimateurs de Bayes formels ne sont pas toujours définis).

Exemple 9.19. (Suite de l'Exemple 9.6) Si π est invariante sous l'action du groupe de translation, elle vérifie $\pi(\theta) = \pi(\theta + c)$ quels que soient θ et c, ce qui implique en particulier $\pi(\theta) = \pi(0)$ uniformément sur $\mathbb{R}$ et fait donc de la mesure de Lebesgue une mesure invariante. ‖

Exemple 9.20. Soit $x_1, \ldots, x_n$ un échantillon de $\mathscr{N}(\theta, \sigma^2)$, avec θ et σ^2 inconnus. En utilisant un argument d'exhaustivité, on peut étudier uniquement le couple $(\bar{x}, s)$, avec $\bar{x}$ moyenne empirique et s^2 somme des erreurs quadratiques. Dans ce cadre, le groupe à considérer est le *groupe affine*

$$\mathscr{G} = \{g_{a,b};\ g_{a,b}(\bar{x}, s) = (a\bar{x} + b, as),\ a > 0, b \in \mathbb{R}\}\ ,$$

et $\bar{\mathscr{G}} = \tilde{\mathscr{G}} = \mathscr{G}$ si le paramètre à estimer est (θ, σ). Si π est une mesure invariante, sa densité vérifie

$$a^2\pi(a\theta + b, a\sigma) = \pi(\theta, \sigma), \qquad \forall a > 0,\ \forall b \in \mathbb{R}\,,$$

ce qui implique

$$\pi(\theta, \sigma) = \pi(0, 1)/\sigma^2.$$

Par conséquent, une mesure invariante est proportionnelle à $\pi(\theta, \sigma) = 1/\sigma^2$ et rappelle la mesure de Jeffreys vue dans le Chapitre 3. ‖

D'une façon générale, étant donné un groupe topologique localement compact $\mathscr{G}$ et en notant $K(\mathscr{G})$ l'ensemble des fonctions réelles continues sur $\mathscr{G}$ à support compact, on définit, pour tout $g \in \mathscr{G}$, la transformation L_g sur $K(\mathscr{G})$

[70] Quand $\bar{\mathscr{G}}$ n'est pas un sous-ensemble de $\mathbb{R}^p$, la structure topologique induite par $\bar{\mathscr{G}}$ est la topologie induite par la composition de groupe et l'inversion, c'est-à-dire la plus petite collection d'ensembles ouverts telle que la composition de groupe et l'inversion soient continues (voir Rudin, 1976).

$$(L_g f)(x) = f(gx) \qquad \text{pour } f \in K(\mathscr{G}),\ x \in \mathscr{G}.$$

Une intégrale J sur $K(\mathscr{G})$ est dite *invariante à gauche* si

$$J(L_g f) = J(f)$$

quels que soient $f \in K(\mathscr{G})$ et $g \in \mathscr{G}$. La mesure de Radon ν_ℓ associée à J est dite *mesure de Haar à gauche*, et on peut montrer (Nachbin, 1965) que cette mesure est unique à une constante multiplicative près. On définit R_g sur $K(\mathscr{G})$ par

$$(R_g f)(x) = f(xg), \qquad \text{pour } f \in K(\mathscr{G}),\ x \in \mathscr{G},$$

et on dérive de manière analogue des *intégrales invariantes à droite* et une *mesure de Haar à droite* ν_r, également définie à un facteur près. Comme nous l'avons énoncé ci-dessus, la *finitude de la mesure de Haar*, c'est-à-dire l'existence d'une distribution de probabilité invariante, est en fait *équivalente à la compacité de* $\mathscr{G}$. Voir Eaton (1989, Chapitre 1) pour des exemples de mesures de Haar ; Berger (1985b) s'intéresse au cas où $\mathscr{G} \subset \mathbb{R}^k$.

La définition du *module de* $\mathscr{G}$ est le *multiplicateur* Δ–qui est une fonction réelle vérifiant $\Delta(g_1 g_2) = \Delta(g_1)\Delta(g_2)$–reliant ainsi les mesures de Haar à droite et à gauche :

$$\nu_r(dx) = \Delta(x^{-1})\nu_\ell(dx)$$

(Exercices 9.13 et 9.15). On suppose l'existence d'une mesure de Radon μ sur $\mathscr{X}$ telle que, pour tout f,

$$\int_{\mathscr{X}} f(g^{-1}x)\mu(dx) = \Delta^{-1}(g) \int_{\mathscr{X}} f(x)\mu(dx).$$

Cette relation établit une connexion entre le module de $\mathscr{G}$ et le jacobien de la transformation de x en gx. Soient les distributions P_θ, $\theta \in \Theta$, de densité $f(x|\theta)$ par rapport à μ. Alors, pour tout $g \in \mathscr{G}$,

$$f(x|\theta) = f(gx|\bar{g}\theta)\Delta^{-1}(g).$$

On fait également l'hypothèse que $\bar{\mathscr{G}}$ agit *transitivement* sur Θ. En ajoutant quelques conditions, Eaton (1989, p. 84) démontre alors un théorème apparenté à celui de Fubini : *Si* ν_r *est la mesure de Haar à droite sur* $\mathscr{G}$*, si* Q *est la projection de* $\mathscr{X}$ *sur* $\mathscr{X}/\mathscr{G}$ *et si* (Tf) *est défini sur* $\mathscr{X}/\mathscr{G}$ *par*

$$(Tf)(Q(x)) = \int_{\mathscr{G}} f(gx)\nu_r(dg),$$

alors il existe une intégrale J_1 *définie sur* $K(\mathscr{X}/\mathscr{G})$ *telle que*

$$J_1(Tf) = \int_{\mathscr{X}} f(x)\mu(dx).$$

Ainsi l'intégrale de f par rapport à μ est l'intégrale sur toutes les orbites de $\mathscr{X}$ (c'est-à-dire sur $\mathscr{X}/\mathscr{G}$) de la moyenne de f par rapport à la mesure de Haar à droite sur chaque orbite, Tf.

Soit un estimateur δ et, pour $\theta \in \Theta$ fixé, posons

$$f_0(x) = \mathrm{L}(\theta, \delta(x)) f(x|\theta),$$

alors

$$R(\delta, \theta) = \int_{\mathscr{X}} f_0(x) \mu(dx).$$

Il vient du théorème ci-dessous qu'il existe une intégrale J_1 sur $K(\mathscr{X}/\mathscr{G})$ telle que

$$R(\delta, \theta) = J_1(Tf_0),$$

avec

$$\begin{aligned}(Tf_0)(Q(x)) &= \int_{\mathscr{G}} \mathrm{L}(\theta, \delta(gx)) f(gx|\theta) \nu_r(dg) \\ &= \int_{\mathscr{G}} \mathrm{L}(\bar{g}\theta, \delta(x)) f(x|\bar{g}\theta) \nu_r(dg)\end{aligned}$$

(voir Eaton, 1989, p. 85). Définissons aussi

$$H(a, x) = \int_{\mathscr{G}} \mathrm{L}(\bar{g}\theta, a) f(x|\bar{g}\theta) \nu_r(dg),$$

qui ne dépend pas de θ (puisque $\bar{\mathscr{G}}$ agit transitivement sur Θ). Remarquons que $H(\delta(x), x)$ donne le risque de δ conditionnellement à l'orbite de x. Ce constat est utile pour la détermination du meilleur estimateur équivariant.

Théorème 9.21. *S'il existe $a_0(x)$ tel que*
(i) *$H(a, x) \geq H(a_0(x), x)$ pour tout $a \in \mathscr{D}$, $x \in \mathscr{X}$; et*
(ii) *$a_0(gx) = \tilde{g} a_0(x)$ pour tout $g \in \mathscr{G}$, $x \in \mathscr{X}$,*
alors $\delta_0(x) = a_0(x)$ est un meilleur estimateur équivariant.

Preuve. Soit un estimateur équivariant δ. Alors

$$\int_{\mathscr{G}} \mathrm{L}(\bar{g}\theta, \delta(x)) f(x|\bar{g}\theta) \nu_r(dg) \geq \int_{\mathscr{G}} \mathrm{L}(\bar{g}\theta, a_0(x)) f(x|\bar{g}\theta) \nu_r(dg).$$

En intégrant par rapport à J_1, on déduit que $R(\delta, \theta) \geq R(\delta_0, \theta)$. L'estimateur δ_0 domine alors δ. □

Ce théorème met en évidence la relation entre le meilleur estimateur équivariant et un estimateur de Bayes particulier, puisque $H(a, x)$ peut également être interprété comme un risque de Bayes a posteriori. Si on sélectionne arbitrairement un $\theta_0 \in \Theta$, la fonction $\tau(g) = \bar{g}\theta_0$ définit en fait une surjection de $\mathscr{G}$ dans Θ eu égard à la transitivité de $\bar{\mathscr{G}}$. Elle induit donc

une mesure sur Θ, appelée *mesure de Haar à droite sur* Θ et définie par $\pi^*(B) = \nu_r(\tau^{-1}(B))$ pour tout $B \in \mathscr{B}(\Theta)$. Elle est manifestement invariante sous l'action de $\mathscr{G}$. En outre,

$$H(a,x) = \int_{\Theta} \mathrm{L}(\theta,a) f(x|\theta)\, d\pi^*(\theta).$$

Cette extension de la mesure de Haar à droite à Θ donne une expression du meilleur estimateur équivariant sous la forme d'un estimateur de Bayes pour tout groupe transitif agissant sur le modèle statistique.

Corollaire 9.22. *Le meilleur estimateur équivariant de θ est l'estimateur de Bayes associé à la mesure de Haar à droite sur Θ, π^*, et au coût invariant correspondant.*

Nous avons donc une méthode qui permet d'obtenir les meilleurs estimateurs équivariants directement à partir de la mesure de Haar à droite. (Voir Stein, 1965, et Zidek, 1965, pour des résultats analogues.)

Dans le raisonnement ci-dessus, la mesure dominante est μ et il s'agit donc d'une mesure *relativement invariante avec pour multiplicateur le module* Δ^{-1}. En fait, si la mesure μ était relativement invariante avec un multiplicateur arbitraire χ, c'est-à-dire que, pour tout $f \in K(\mathscr{G})$,

$$\int_{\mathscr{X}} f(gx)\mu(dx) = \chi(g) \int_{\mathscr{X}} f(x)\mu(dx),$$

le Corollaire 9.22 serait toujours vrai (Eaton, 1989, p. 87).

Exemple 9.23. (Suite de l'Exemple 9.20) Nous avons la *mesure de Haar à gauche* suivante sur Θ :

$$\pi^{\ell}(\theta,\sigma) = 1/\sigma^2.$$

La *mesure de Haar à droite* peut en être déduite par inversion : si $g = (a,b)$ et $g_0 = (a_0,b_0)$, $gg_0 = (aa_0, ab_0 + b)$ pour la composition de groupe. En prenant le jacobien en compte, nous voulons que la mesure de Haar à droite vérifie

$$a_0 \pi^r(b_0\sigma + \theta, a_0\sigma) = \pi^r(\theta,\sigma)$$

pour tout (θ,σ) et uniformément sur a_0, b_0 ; ceci entraîne

$$\pi^r(\theta,\sigma) = 1/\sigma,$$

à un facteur multiplicatif près. Par conséquent, la mesure de Haar à droite est différente de la mesure de Haar à gauche et donne une alternative non informative à l'a priori de Jeffreys (Section 3.6). Pour le coût quadratique invariant,

$$\mathrm{L}((\theta,\sigma),\delta) = \frac{(\theta-\delta_1)^2}{\sigma^2} + \left(\frac{\delta_2}{\sigma} - 1\right)^2, \tag{9.3}$$

le meilleur estimateur équivariant est l'estimateur de Bayes associé à la distribution a priori π^r, soit,

$$\delta_1^*(\bar{x},s)=\frac{\mathbb{E}^{\pi^r}[\theta/\sigma^2|\bar{x},s]}{\mathbb{E}^{\pi^r}[1/\sigma^2|\bar{x},s]},\qquad \delta_2^*(\bar{x},s)=\frac{\mathbb{E}^{\pi^r}[1/\sigma|\bar{x},s]}{\mathbb{E}^{\pi^r}[1/\sigma^2|\bar{x},s]}.$$

Puisque

$$\pi^r(\theta,\sigma|\bar{x},s)\propto\sigma^{-(n+1)}e^{-n(\bar{x}-\theta)^2/2\sigma^2}e^{-s^2/2\sigma^2},$$

il s'agit d'un cas particulier de distribution conjuguée sur (θ,σ) et

$$\delta_1^*(\bar{x},s)=\bar{x},\qquad \delta_2^*(\bar{x},s)=\frac{\Gamma(n/2)}{\sqrt{2}\Gamma((n+1)/2)}s.$$

Remarquons que δ_2 est aussi l'estimateur obtenu dans l'Exemple 9.15. ‖

Exemple 9.24. (Eaton, 1989) Soit un *modèle multiplicatif* $\mathcal{N}(\theta,\theta^2)$, à n observations $x_1,\ldots,x_n$. Ce modèle apparaît dans des contextes où la difficulté de mesure d'un objet augmente avec sa magnitude (Physique des particules, Astronomie, etc.). Si nous estimons θ sous le coût

$$\mathrm{L}(\theta,d)=\frac{(\theta-d)^2}{\theta^2},$$

le problème est invariant sous l'action du groupe multiplicatif. La mesure de Haar à droite est alors $\pi(\theta)=1/|\theta|$. (Il s'agit aussi de la mesure de Haar à gauche puisque le groupe est commutatif.)

Le meilleur estimateur équivariant de θ est donc

$$\delta^*(x_1,\ldots,x_n)=\frac{\mathbb{E}^{\pi}[1/\theta|x_1,\ldots,x_n]}{\mathbb{E}^{\pi}[1/\theta^2|x_1,\ldots,x_n]}$$

et

$$\begin{aligned}\pi(\theta|x)&\propto\frac{1}{\theta^2}\exp\left\{-\sum_{i=1}^n(x_i-\theta)^2/2\theta^2\right\}\\&\propto\frac{1}{\theta^2}\exp\left\{-\frac{1}{2}\left(\frac{n\bar{x}}{s^2}-\frac{1}{\theta}\right)^2s^2\right\},\end{aligned}$$

pour $s^2=\sum_{i=1}^n x_i^2$. La distribution a posteriori est alors *inverse normale généralisée* $\mathcal{IN}(2,n\bar{x}/s^2,1/s^2)$ (Robert, 1991) et

$$\mathbb{E}^{\pi}[1/\theta|\bar{x},s^2]=\sqrt{2}s\,\frac{{}_1F_1(1;1/2;n^2\bar{x}^2/2s^2)}{\Gamma(1/2){}_1F_1(1/2;1/2;n^2\bar{x}^2/2s^2)}.$$

Par conséquent,

$$\delta^*(x_1,\ldots,x_n) = \sqrt{2}\ \Gamma(3/2)\frac{{}_1F_1(3/2;1/2;n^2\bar{x}^2/2s^2)}{\Gamma(1/2){}_1F_1(1;1/2;n^2\bar{x}^2/2s^2)}\ s.$$

Dans ce cas, le meilleur estimateur équivariant domine l'estimateur du maximum de vraisemblance

$$\hat{\delta}(\bar{x},s) = \frac{-\bar{x} + (\bar{x}^2 + 4s^2)^{1/2}}{2},$$

qui est également équivariant. Pour plus de résultats sur les modèles multiplicatifs, voir Gleser et Healy (1976), Kariya *et al.* (1988) et Perron et Giri (1990). ‖

Les lecteurs pourront se référer à Eaton (1989), Lehmann (1986) et Berger (1985b) pour d'autres exemples d'utilisation des mesures de Haar concernant la détermination de meilleurs estimateurs équivariants dans les cadres des tests et de calculs de régions de confiance. Pour un traité général de mathématiques sur les mesures de Haar, voir Nachbin (1965).

9.5 Le théorème de Hunt-Stein

Replaçons-nous dans le cas évoqué en début de la section précédente, c'est-à-dire celui où $\mathscr{G}$ est compact et où il existe une distribution de probabilité invariante sur Θ. Alors le meilleur estimateur équivariant est un estimateur (propre) de Bayes et est donc admissible la plupart du temps. Comme le risque est constant lorsque $\bar{\mathscr{G}}$ est transitif, le meilleur estimateur équivariant est aussi *minimax*. Si $\mathscr{G}$ n'est pas compact, le meilleur estimateur équivariant est un estimateur de Bayes généralisé associé à la mesure de Haar à droite et n'est donc pas nécessairement admissible. L'effet Stein (Note 2.8.2) illustre cette possible sous-optimalité en montrant que le meilleur estimateur équivariant d'un paramètre de position, x, est inadmissible pour le coût quadratique en dimension 3 et plus. Par conséquent, il est vain d'espérer une réponse générale à la question de l'admissibilité du meilleur estimateur équivariant pour des groupes non compacts.

En revanche, il est possible d'étendre la propriété de minimaxité au-delà du cas compact, grâce au *théorème de Hunt-Stein*[71]. Ce résultat est conforme à l'intuition puisque, quand un problème est invariant, il existe un estimateur équivariant à risque constant qui atteint la borne inférieure du risque maximal

$$\inf_{\delta}\sup_{\theta} R(\delta,\theta).$$

[71] Ce théorème est également célèbre pour être resté longtemps sans démonstration publiée, bien que Kiefer (1957) en ait fourni une dans un cas particulier.

En outre, il semble logique de tirer partie de la structure naturelle d'invariance du modèle pour améliorer un estimateur δ en le "moyennant" par intégration sur $\mathscr{G}$

$$\delta^*(x) = \int_{\mathscr{G}} \delta(gx)\nu_r(dg),$$

si $\mathrm{L}(\theta, d)$ est convexe en d et si le théorème inspiré de celui de Fubini, présenté en Section 9.4, s'applique (en supposant que δ^* est bien définie). De façon informelle, nous obtiendrions alors en fait

$$\begin{aligned} R(\delta, \theta) &= \mathbb{E}_\theta[\mathrm{L}(\theta, \delta(x))] \\ &= \mathbb{E}^T(\mathbb{E}_\theta[\mathrm{L}(\theta, \delta(x))|Q(x) = T]) \\ &\geq \mathbb{E}^T[L(\theta, \delta^*(t))] = R(\delta^*, \theta). \end{aligned}$$

Cette amélioration rappelle le résultat de domination du théorème de Rao-Blackwell, dans le cas du conditionnement à une statistique exhaustive.

Nous poussons un pas plus loin la formalisation de la démonstration en introduisant la notion de *groupe moyennable* présentée en détail par Bondar et Milnes (1981). Présentons tout d'abord un contre-exemple qui montre que l'intuition n'a pas toujours raison, en particulier lorsque les structures d'invariance sont trop fortes, c'est-à-dire lorsque $\mathscr{G}$ est trop grand.

Exemple 9.25. (Stein, 1965) Soient $x \sim \mathscr{N}_p(0, \Sigma)$ et $y \sim \mathscr{N}_p(0, \varrho\Sigma)$ avec $p \geq 2$. Le paramètre ϱ est estimé sous la fonction de coût

$$\mathrm{L}((\varrho, \Sigma), d) = \mathbb{I}_{[1/2, +\infty)}\left(\left|1 - \frac{d}{\varrho}\right|\right).$$

Le problème est alors invariant sous l'action du groupe linéaire GL_p parce que, si B est une matrice régulière, $Bx \sim \mathscr{N}_p(0, B\Sigma B^t)$ et $By \sim \mathscr{N}_p(0, \varrho B\Sigma B^t)$. Puisque $\bar{g}_B(\varrho, \Sigma) = (\varrho, B\Sigma B^t)$, les estimateurs équivariants sont en réalité invariants

$$\delta(Bx, By) = \delta(x, y)$$

quels que soient x, y et B. Si x et y sont linéairement indépendants (ce qui est vrai avec probabilité 1), on peut trouver B telle que

$$Bx = (1, 0, \ldots, 0)^t \qquad \text{et} \qquad By = (0, 1, 0, \ldots, 0)^t,$$

ce qui implique que les estimateurs équivariants sont *constants presque partout*. Comme

$$R(\delta_0, (\varrho, \Sigma)) = 1 \qquad \text{si} \quad \left|1 - \frac{\delta_0}{\varrho}\right| > 1/2$$

pour une constante donnée δ_0, le risque minimax des estimateurs équivariants est de 1.

En posant

$$\delta_1(x, y) = \left| \frac{y_2}{x_1} \right|,$$

le risque de δ_1 est

$$\begin{aligned} R(\delta_1, \theta) &= P_{\varrho,\Sigma}\left(\left| 1 - \left| \frac{y_2}{x_1 \varrho} \right| \right| \geq 1/2 \right) \\ &= P\left(\left| 1 - \left| \frac{z_1}{z_2} \right| \right| \geq 1/2 \right), \end{aligned}$$

où z_1, z_2 sont i.i.d. $\mathcal{N}(0, 1)$. Par conséquent, le risque est également *constant*, mais strictement plus petit que 1. On peut remarquer que δ_1 est aussi un estimateur équivariant pour le groupe multiplicatif, qui semble une structure d'invariance mieux appropriée. ‖

Afin d'obtenir une approche plus générale du problème, on considère à présent un groupe localement compact de transformations $\mathscr{G}$, avec une mesure de Haar à droite ν_r. Soit $\mathscr{V}$ une algèbre de fonctions mesurables essentiellement bornées à valeurs réelles sur $\mathscr{G}$, telle que la fonction constante $\mathbf{1}$ soit dans $\mathscr{V}$.

Définition 9.26. *Une* moyenne *sur $\mathscr{V}$ est une fonctionnelle m, linéaire et continue sur $\mathscr{V}$, telle que*
(i) $m(\mathbf{1}) = 1$; et
(ii) $m(f) \geq 0$ si $f \in \mathscr{V}$ et $f \geq 0$ (presque sûrement).

L'existence d'une telle fonctionnelle m est en fait une condition nécessaire et suffisante pour le théorème de Hunt-Stein. Si m existe, il est possible de moyenner sur les orbites de $\mathscr{X}$ par rapport à $\mathscr{G}$, comme nous l'avons évoqué en début de section.

Exemple 9.27. (Bondar et Milnes, 1981) Pour $\mathscr{G} = \mathbb{R}$ et $n \in \mathbb{N}$, on considère

$$m_n(f) = \frac{1}{2n} \int_{-n}^{n} f(x)\, dx;$$

alors m_n définit une moyenne sur $\mathscr{L}_\infty(\mathbb{R})$. De plus, la suite (m_n) a un point d'accumulation m dans la *topologie faible* sur $\mathscr{L}_\infty$: pour tout $f \in \mathscr{L}_\infty$, $\epsilon > 0$ et $n_0 \in \mathbb{N}$, il existe $n \geq n_0$ tel que

$$|m_n(f) - m(f)| < \epsilon.$$

En particulier, ce point d'accumulation vérifie $m(f) = 0$ quelle que soit f telle que $f(x)$ tende vers 0 quand x tend vers $\pm\infty$. Notons également que m n'est pas σ-additive et que la suite (m_n) ne converge pas vers m au sens de la topologie faible. ‖

Définition 9.28. *La moyenne m est* invariante à droite *si, pour toutes $f \in \mathscr{V}$ et $g \in \mathscr{G}$, $m(f_g) = m(f)$, avec $f_g(x) = f(xg)$. Le groupe $\mathscr{G}$ est dit* moyennable *s'il existe une moyenne invariante à droite sur $\mathscr{L}_\infty(\mathscr{G})$ ou, de façon équivalente, sur $\mathscr{C}_B(\mathscr{G})$, l'espace des fonctions continues bornées sur $\mathscr{G}$.*

Comme le montrent Bondar et Milnes (1981), l'existence d'un groupe moyennable est équivalente à *l'existence d'une suite de mesures de probabilité presque invariantes à droite* : il existe alors une suite (P_n) de mesures de probabilité sur $\mathscr{G}$ telle que, pour tous $B \in \mathscr{B}(\mathscr{G})$ et $g \in \mathscr{G}$,

$$\lim_{n \to +\infty} |P_n(Bg) - P_n(B)| = 0.$$

En outre, il existe une suite (G_n) d'ensembles compacts imbriqués tels que la densité de P_n soit $\nu_r(G_n)^{-1}\mathbb{I}_{G_n}(g)$ (par rapport à ν_r). La suite (G_n) conduit donc à une approximation de la mesure de Haar ν_r par une suite de distributions de probabilité et ces distributions sont presque invariantes au sens où

$$B \cap G_n = Bg \cap G_n, \quad P_n(B) = P_n(Bg)$$

(on peut également consulter Strasser, 1985, et Lehmann, 1986). L'Exemple 9.27 est une illustration directe de ce résultat.

Des exemples de groupes moyennables sont les groupes additifs et multiplicatifs, le groupe de transformations position-échelle (Exemple 9.18) et le groupe T_p des matrices triangulaires supérieures inversibles. À l'inverse, le groupe linéaire GL_p et le groupe SL_p des matrices de déterminant 1 ne sont pas moyennables. Bondar et Milnes (1981) donnent de nombreux exemples de groupes moyennables et de groupes non moyennables.

Le théorème de Hunt-Stein établit la minimaxité du meilleur estimateur équivariant.

Théorème 9.29. *Si le groupe $\mathscr{G}$ est moyennable et si le problème statistique $(\mathscr{X}, f(x|\theta), \mathscr{D}, \mathrm{L})$ est invariant sous l'action de $\mathscr{G}$, l'existence d'un estimateur minimax implique celle d'un estimateur minimax équivariant. De plus, un estimateur équivariant qui est minimax dans l'ensemble des estimateurs équivariants est minimax.*

Des preuves de ce théorème figurent dans Berger (1985b, Section 6.7) pour le cas où $\mathscr{G}$ est fini, Lehmann (1983, Section 9.5) pour les tests et Le Cam (1986, Section 8.6) dans des cadres plus généraux, en tant que conséquence du théorème du point fixe de Markov-Kakutani. Comme précisé plus haut, le théorème de Hunt-Stein repose sur une version modifiée du théorème de Fubini. Nous nous contentons ici de donner une idée générale de la démonstration. On suppose que L est convexe. Pour un estimateur δ à valeurs réelles, on pose

$$\delta^*(x) = m(\tilde{\delta}_x),$$

où m est la moyenne invariante à droite et $\tilde{\delta}_x(g) = \delta(gx)$. L'estimateur δ^* est alors équivariant puisque, si $g_0 \in \mathscr{G}$,

$$\begin{aligned}\delta^*(g_0 x) &= \int_{\mathscr{G}} \tilde{g}^{-1}\delta(gg_0x)\,dm(g)\\ &= \int_{\mathscr{G}} \tilde{g}_0\tilde{g}_0^{-1}\tilde{g}^{-1}\delta(gg_0x)\,dm(g)\\ &= \tilde{g}_0\int_{\mathscr{G}} \tilde{g}^{-1}\delta(gx)\,dm(g)\\ &= \tilde{g}_0\delta^*(x),\end{aligned}$$

par l'invariance à droite de m. Par ailleurs,

$$\sup_\theta R(\delta^*,\theta) \le \sup_\theta \int_{\mathscr{G}}\int_{\mathscr{X}} \mathrm{L}\left(\theta, \tilde{g}^{-1}\delta(gx)\right) f(x|\theta)\,dx\,dm(g) \tag{9.4}$$

par convexité de L. Il vient

$$\begin{aligned}\sup_\theta R(\delta^*,\theta) &\le \sup_\theta \int_{\mathscr{G}}\int_{\mathscr{X}} \mathrm{L}\left(\bar{g}\theta, \delta(gx)\right) f(x|\theta)\,dx\,dm(g)\\ &= \sup_\theta \int_{\mathscr{G}} R(\bar{g}\theta,\delta)\,dm(g)\\ &\le \sup_\theta R(\delta,\theta),\end{aligned}$$

ce qui entraîne[72] la domination de δ par δ^*.

Une conséquence du théorème de Hunt-Stein est que, dans le cas normal, l'estimateur du maximum de vraisemblance, $x \sim \mathscr{N}_p(\theta, I_p)$, est minimax pour tout p, bien qu'inadmissible pour $p \ge 3$. Le même résultat est vrai si $x \sim \mathscr{N}_p(\theta, \sigma^2 I_p)$ et la variance inconnue σ^2 est estimée par s^2/q, avec $s^2 \sim \sigma^2\chi^2_q$.

9.6 L'invariance en Statistique bayésienne

Pour conclure ce chapitre, nous mentionnons ici les réserves émises par Berger (1985b) quant aux conséquences des exigences d'invariance dans l'approche bayésienne. Elles concernent en particulier le processus de détermination de distributions non informatives, même s'il a l'avantage de présenter en le justifiant un choix alternatif à l'a priori de Jeffreys (Exemple 9.18).

Une critique qu'on peut adresser à la notion d'invariance est que, bien qu'intuitivement attractive, elle n'est pas dénuée d'ambiguïté et, puisqu'il est

[72] Insistons sur le fait que ces indications n'ont pas valeur de preuve rigoureuse, puisque l'application du théorème de Fubini à (9.12) n'est pas toujours justifiée. Il se trouve que cette opération de moyenne ne peut être effectuée que sous des conditions précises. Sinon, on obtiendrait de même un résultat d'admissibilité pour le meilleur estimateur équivariant sous coût convexe, résultat contesté par l'effet Stein.

parfois possible de considérer plusieurs groupes globalement invariants, les meilleurs estimateurs équivariants en résultant peuvent être distincts, ce qui contredit le principe de vraisemblance.

Un inconvénient plus direct de la méthode est que les structures naturelles d'invariance d'un modèle statistique peuvent être trop faibles et donc sans intérêt pour déterminer un estimateur, ou trop fortes et donc trop contraignantes. Une illustration extrême du premier écueil est obtenue avec la distribution de Poisson, pour laquelle il n'existe aucune structure d'invariance. L'exemple suivant se place dans le cas opposé (voir aussi l'Exemple 9.25).

Exemple 9.30. Soit une famille de distributions symétriques par rapport à un paramètre de position θ, c'est-à-dire telles que $x \sim f(|x-\theta|)$. La fonction de coût est $\varrho(|d-\theta|)$. Si on prend en compte l'invariance par symétrie, c'est-à-dire le fait que la distribution de $y = -x$ appartienne à la même famille, les estimateurs correspondant à $\pi(\theta) = 1$ et satisfaisant

$$\delta(x+c) = \delta(x) + c \qquad \text{et} \qquad \delta(-x) = -\delta(x)$$

se réduisent à $\delta(x) = x$, qui n'est pas nécessairement un choix judicieux. ‖

Un excès d'invariance peut évidemment être modéré en ignorant certaines structures d'invariance, c'est-à-dire en ne considérant qu'un sous-groupe $\mathscr{G}_0$ de $\mathscr{G}$ qui induise une action transitive sur Θ, tout en étant aussi petit que possible. Cependant, même lorsqu'il est envisageable, le choix d'un tel sous-groupe peut se révéler crucial dans la suite du processus inférentiel.

Une dernière critique importante est que la modélisation de problèmes statistiques par des structures d'invariance peut être néfaste d'un point de vue *subjectif*, puisqu'elle impose la compatibilité des structures de décision avec l'invariance–et donc, en particulier, le choix d'un coût invariant–ce qui peut contredire l'information a priori–la seule distribution a priori compatible étant la mesure de Haar. La méthode peut également être peu *efficace*, puisque les estimateurs équivariants sont parfois fortement inadmissibles, comme le montrent l'effet Stein et l'Exemple 9.25 (voir aussi les Exemples 4.4-4.9 de Lehmann, 1983, Section 4.4). Par ailleurs, l'invariance ne conduit pas nécessairement à une distribution non informative satisfaisante comme on le voit dans l'Exemple 9.30. Enfin, en pratique, le calcul des mesures de Haar à droite peut se révéler fastidieux.

9.7 Exercices

Section 9.2

9.1 (Blackwell et Girshick, 1954) On considère la distribution f avec les poids $f(k) = 1/k(k+1)$ pour $k = 1, 2, \ldots$ et $x \sim f(x-\theta)$, avec $\theta \in \mathbb{R}$. Pour la fonction de coût

$$\mathrm{L}(\theta, d) = \begin{cases} d - \theta & \text{si } d > \theta, \\ 0 & \text{sinon,} \end{cases}$$

montrer que les estimateurs équivariants sont de la forme $x - c$ et que tout estimateur équivariant a un risque infini. Comparer à l'estimateur constant $\delta_0(x) = c$.

9.2 Soit x une observation issue d'une loi de Cauchy $\mathscr{C}(\theta, 1)$. Pour un coût quadratique, montrer que tous les estimateurs équivariants sont de risque infini. Proposer un estimateur à risque fini différent de l'estimateur constant.

9.3 (Berger, 1985b) Soit

$$x = (x_1, \ldots, x_n) \sim f(x_1 - \theta, \ldots, x_n - \theta),$$

avec θ inconnu. On veut tester l'hypothèse $H_0 : f = f_0$ contre $H_1 : f = f_1$ sous le coût $0 - 1$.

a. Montrer que $T(x) = (x_1 - x_n, \ldots, x_{n-1} - x_n)$ est une statistique invariante maximale pour le groupe de transformations

$$\mathscr{G} = \{g_c;\ g_c(x_1, \ldots, x_n) = (x_1 + c, \ldots, x_n + c),\ c \in \mathbb{R}\}\,.$$

b. En déduire qu'un test invariant ne dépend que de $y = T(x)$ et que les tests optimaux ont la région de rejet suivante :

$$W = \{f_1^*(y) \geq K f_0^*(y)\},$$

où f_i^* est la densité de y sous H_i.

9.4 (Berger, 1985b) Soit x distribué selon

$$P_\theta(x = \theta - 1) = P_\theta(x = \theta + 1) = 1/2.$$

La fonction de coût associée est

$$\mathrm{L}(\theta, d) = \begin{cases} |\theta - d| & \text{si } |\theta - d| \leq 1, \\ 1 & \text{sinon.} \end{cases}$$

Déterminer les meilleurs estimateurs équivariants pour le groupe de translation et montrer qu'ils sont dominés par

$$\delta^*(x) = \begin{cases} x + 1 & \text{si } x \leq 0, \\ x - 1 & \text{sinon.} \end{cases}$$

9.5 (Berger, 1985b) Soit $x_1, \ldots, x_n$ un échantillon de la distribution normale tronquée à $\mathbb{R}_+$ de densité

$$f(x|\theta) = \left(\frac{2}{\pi}\right)^{1/2} e^{-(x-\theta)^2/2} \mathbb{I}_{[\theta, +\infty)}(x).$$

Montrer que le meilleur estimateur équivariant de θ sous coût quadratique est

$$\delta^*(x) = \bar{x} - \frac{\exp\{-n(x_{(1)} - \bar{x})^2/2\}}{\sqrt{2n\pi}\Phi(\sqrt{n}(x_{(1)} - \bar{x}))}.$$

Section 9.3

9.6 Soit $x \sim \mathcal{N}(\theta, a\theta^2)$, avec $\theta \in \mathbb{R}$ et $a > 0$ connu. Le paramètre θ est estimé sous le coût $\mathrm{L}(\theta, d) = (\frac{d}{\theta} - 1)^2$.

a. Montrer que le problème est invariant sous le groupe de transformations

$$\mathcal{G} = \{g_c;\ g_c(x) = cx,\ c > 0\}.$$

L'action du groupe est-elle transitive ?

b. Donner les meilleurs estimateurs équivariants et du maximum de vraisemblance de θ.

c. Les comparer aux estimateurs obtenus dans l'Exercice 3.33 et dans l'Exemple 9.24.

d. Montrer, à l'aide de l'Exercice 3.33, que le meilleur estimateur équivariant δ_0 est un estimateur de Bayes généralisé.

9.7 (Lehmann, 1983) On cherche à estimer un paramètre d'échelle σ, sous le coût

$$\mathrm{L}(\sigma, \delta) = \left(\frac{\delta}{\sigma} - 1\right)^2, \tag{9.5}$$

pour n observations

$$x_1, \ldots, x_n \sim \frac{1}{\sigma^n} f\left(\frac{x_1}{\sigma}, \ldots, \frac{x_n}{\sigma}\right).$$

a. Si $z = (x_1/x_n, \ldots, x_{n-1}/x_n, x_n/|x_n|)$, montrer que tout estimateur de σ équivariant sous transformation d'échelle peut s'écrire

$$\delta(x) = \delta_0(x)/\omega(z),$$

avec δ_0 un estimateur équivariant particulier et que z est une statistique invariante maximale.

b. Déterminer la fonction ω^* qui minimise

$$\mathbb{E}[\mathrm{L}(\sigma, \delta(x))|z]$$

sous (9.5) et en déduire le meilleur estimateur équivariant.

c. Transformer l'écriture de cet estimateur pour retrouver le résultat de la Section 9.4 avec la mesure de Haar correspondante.

d. Reprendre les questions précédentes pour le problème d'estimation de σ^r ($r \in \mathbb{R}_+^*$) sous le coût

$$\mathrm{L}(\sigma, \delta) = \left(\frac{\delta}{\sigma^r} - 1\right)^2.$$

9.8 Appliquer les résultats de l'Exercice 9.7 aux cas suivants :
(i) $x_1, \ldots, x_n$ i.i.d. $\mathcal{N}(0, \sigma^2)$;
(ii) $x_1, \ldots, x_n$ i.i.d. $\mathcal{G}(\alpha, \sigma)$; et
(iii) $x_1, \ldots, x_n$ i.i.d. $\mathcal{U}[0, \sigma]$.

9.9 Reprendre l'Exercice 9.7 sous les coûts suivants :

$$\mathrm{L}(\sigma, \delta) = \frac{|\delta - \sigma|}{\sigma}, \quad \mathrm{L}(\sigma, \delta) = \frac{\delta}{\sigma} - \log(\delta/\sigma) - 1, \quad \mathrm{L}(\sigma, \delta) = \left(\frac{\sigma}{\delta} - 1\right)^2.$$

9.10 (Lehmann, 1983) On considère le problème d'estimation de σ dans le cas où

$$x = (x_1, \ldots, x_n) \sim \frac{1}{\sigma^n} f\left(\frac{x_1 - \theta}{\sigma}, \ldots, \frac{x_n - \theta}{\sigma}\right),$$

sous l'action du *groupe affine*

$$\mathscr{G}_a = \{g_{a,b};\ g_{a,b}(x) = ax + b\mathbf{1},\ a > 0, b \in \mathbb{R}\},$$

avec $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^n$.

a. Déterminer le meilleur estimateur équivariant sous le coût (9.5) de même que dans l'Exercice 9.7. (*Indication :* Utiliser les transformations $y_i = x_i - x_n$ et poser $z_i = y_i/y_{n-1}$ $(i \neq n-1)$, $z_{n-1} = y_{n-1}/|y_{n-1}|$.)

b. Comparer à une formulation bayésienne avec la mesure de Haar à droite.

c. Reprendre les questions précédentes pour l'estimation de θ sous le coût

$$\mathrm{L}(\theta, \delta) = \frac{(\theta - \delta)^2}{\sigma^2}.$$

d. Appliquer au cas où $x_i - \theta \sim \mathscr{E}xp(\sigma)$ et montrer que le meilleur estimateur équivariant de θ est

$$\delta^*(x) = x_{(1)} - \frac{1}{n^2} \sum_{i=1}^{n} (x_i - x_{(1)}).$$

9.11 *(Eaton, 1989) On considère $\mathscr{G} \subset \mathbb{R}_+^* \times \mathbb{R}$ muni de l'opération de groupe

$$(a_1, b_1)(a_2, b_2) = (a_1 a_2, a_1 b_2 + b_1).$$

Si $D = \{x \in \mathbb{R}^n;\ x_1 = \cdots = x_n\}$, on considère $\mathscr{X} = \mathbb{R}^n - D$. On suppose que $\mathscr{G}$ agit sur $\mathscr{X}$ de la façon suivante

$$(a, b)x = ax + be_n,$$

avec $e_n = (1, \ldots, 1)^t$. Montrer que la statistique invariante maximale est

$$f(x) = \frac{x - \bar{x}e_n}{s(x)},$$

avec $\bar{x} = \sum x_i/n$, $s^2(x) = \sum (x_i - \bar{x})^2$.

9.12 *(Eaton, 1989) Vérifier que, s'il existe un *multiplicateur* ξ sur $\mathscr{G}$, c'est-à-dire une fonction à valeurs réelles telle que $\xi(g_1 g_2) = \xi(g_1)\xi(g_2)$, qui satisfasse

$$f(x|\theta) = f(gx|\bar{g}\theta)\xi(g)$$

uniformément sur $\mathscr{X}$, Θ, $\mathscr{G}$, la famille

$$\mathscr{P} = \{f(x|\theta); \theta \in \Theta\}$$

est $\mathscr{G}$-invariante. En déduire que, dans ce cas, l'estimateur du maximum de vraisemblance est équivariant, comme tout estimateur de Bayes associé à une mesure a priori *relativement invariante*, c'est-à-dire telle qu'il existe un multiplicateur ξ_1 avec $\pi(gB) = \xi_1(g)\pi(B)$ uniformément en B et g.

9.13 *(Delampady, 1989) Soit $x \sim \mathcal{N}_p(\theta, I_p)$. On teste l'hypothèse $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$. Ce problème est invariant sous l'action du groupe orthogonal $\mathcal{G}_o$ et on considère uniquement les distributions a priori de la classe invariante

$$I = \{\pi;\ \pi(gA) = \pi(A), \forall A \in \mathcal{B}(\mathbb{R}^p), \forall g \in \mathcal{G}_o\}.$$

a. Montrer que $t(x) = ||x||^2$ est une statistique invariante maximale, distribuée selon une loi du χ^2_p décentré, de paramètre de non-centralité $\eta(\theta) = ||\theta||^2$ (la statistique invariante maximale correspondante sur $\bar{\mathcal{G}}_0$), et que sa densité peut s'écrire

$$q(t(x)|\eta(\theta)) = \int_{\mathcal{G}_o} f(gx|\theta)\, d\mu(g),$$

avec μ mesure de Haar sur $\mathcal{G}_0$.

b. En déduire que si B^π est le facteur de Bayes, il vérifie

$$\inf_{\pi \in I} B^\pi(x) = \frac{q(t(x)|\theta_0)}{q(t(x)|\hat{\eta})},$$

avec $\hat{\eta}$ estimateur du maximum de vraisemblance de η.

c. Comparer avec la *p-value* pour différentes valeurs de $t(x)$.

9.14 Montrer que les coûts intrinsèques définis en Section 2.5.4 sont naturellement invariants.

9.15 On considère $x \sim \mathcal{N}(\theta, \sigma^2)$. Le paramètre d'intérêt est e^θ et σ^2 est connu.

a. Montrer que

$$\mathbb{E}_\theta[e^{ax}] = e^{a\theta + a^2\sigma^2/2}.$$

b. Parmi les estimateurs de la forme $\delta_c(x) = e^{x+c\sigma^2}$, déterminer le meilleur estimateur (en c) pour le coût quadratique L_2, δ^*. Montrer que δ^* est un estimateur de Bayes et déterminer l'a priori correspondant π^*. (*Indication :* Considérer d'abord la mesure de Lebesgue et le coût quadratique pondéré

$$L_0(\theta, \delta) = e^{-2\theta}(e^\theta - \delta)^2.$$

Quel est l'estimateur de Bayes pour l'a priori de Lebesgue sous L_2 ?)

c. Reprendre la question précédente pour le coût d'erreur absolu

$$L_1(\theta, \delta) = |e^\theta - \delta|.$$

Montrer que le meilleur estimateur est associé à $\pi(\theta) = e^{-\theta}$. Cette réponse est-elle surprenante du point de vue de l'invariance ?

d. Étant donné l'estimateur δ^*, nous souhaitons évaluer les performances de δ^* sous L_0 et L_2, c'est-à-dire estimer $L_0(\theta, \delta^*(x))$ et $L_2(\theta, \delta^*(x))$ sous le coût quadratique

$$\left(L_0(\theta, \delta^*(x)) - \gamma\right)^2. \tag{9.6}$$

Montrer que, pour $\pi(\theta) = 1$, le coût a posteriori $\mathbb{E}^\pi[L_0(\theta, \delta^*)|x]$ est constant et égal au risque constant de δ^*.

e. Montrer que, pour $\pi^*(\theta) = \exp(-2\theta)$, la variance a posteriori de δ^* est

$$\gamma^\pi(x) = e^{2x - 2\sigma^2}\left(1 - e^{-\sigma^2}\right).$$

Montrer que γ^π est un estimateur non biaisé du risque, $\mathbb{E}_\theta[L_2(\theta, \delta^*(x))]$, et qu'il est dominé par l'estimateur de Bayes de $L_2(\theta, \delta^*(x))$ sous $\pi(\theta) = e^{-4\theta}$. Peut-on justifier l'utilisation de cet a priori par des considérations d'invariance ?

Section 9.4

9.16 *(Eaton, 1989) Montrer que, pour un groupe topologique $\mathcal{G}$, deux intégrales invariantes à gauche, c'est-à-dire deux fonctionnelles telles que

$$\int_{\mathcal{G}} f(gx)\mu(dx) = \int_{\mathcal{G}} f(x)\mu(dx)$$

pour tous $f \in \mathscr{L}_1(\mu)$ et $g \in \mathcal{G}$, sont nécessairement proportionnelles.

9.17 *(Eaton, 1989) On considère ν_ℓ une mesure de Haar à gauche, $f \in K(\mathcal{G})$ et

$$J_1(f) = \int_{\mathcal{G}} f(xg^{-1})\nu_\ell(dx).$$

a. Montrer que J_1 est invariant à gauche. En déduire qu'il existe une fonction Δ sur $\mathcal{G}$ telle que

$$J_1(f) = \Delta(g)\int_{\mathcal{G}} f(x)\nu_\ell(dx) = \Delta(g)J(f).$$

La fonction Δ est appelée le *module de* $\mathcal{G}$.

b. Montrer que Δ ne dépend pas du choix de J_1 et que $\Delta(g_1g_2) = \Delta(g_1)\Delta(g_2)$ (c'est-à-dire que Δ est un *multiplicateur*).

c. Soit J_2 tel que

$$J_2(f) = \int_{\mathcal{G}} f(x)\Delta(x^{-1})\nu_\ell(dx).$$

Montrer que J_2 est invariante à droite et satisfait

$$J_2(f) = \int_{\mathcal{G}} f(x^{-1})\nu_\ell(dx).$$

En déduire que, si ν_ℓ est une mesure de Haar à gauche,

$$\nu_r(dx) = \Delta(x^{-1})\nu_\ell(dx)$$

est une mesure de Haar à droite.

d. Si $\mathcal{G}$ est compact, montrer que Δ est identiquement égal à 1. (*Indication :* Utiliser la continuité de Δ et le fait que $\Delta(\mathcal{G})$ soit compact.)

e. On note $\mathcal{G} = GL_n$, le groupe linéaire de $\mathbb{R}^n$ et dx la mesure de Lebesgue sur $\mathscr{L}_{n,n}$, l'espace vectoriel des matrices $n \times n$. Montrer que

$$J(f) = \int_{\mathcal{G}} f(x)\frac{dx}{|\det(x)|^n}$$

est à la fois invariante à droite et à gauche. En déduire que $\Delta = 1$. $\mathcal{G}$ est-il compact ?

9.18 *(Eaton, 1989) Soient $\mathcal{G}$ un groupe compact agissant sur $\mathcal{X}$ et ν l'unique distribution de probabilité de Haar sur $\mathcal{G}$. On définit U, variable aléatoire uniforme sur $\mathcal{G}$, par

$$P(U \in B) = \nu(B).$$

a. Soit $x \in \mathcal{X}$. Montrer que μ_x, définie par

$$\mu_x(B) = P(Ux \in B)$$

est l'unique probabilité $\mathcal{G}$-invariante sur l'orbite de x, O_x.

b. Si P est une distribution $\mathscr{G}$-invariante sur $\mathscr{X}$, montrer que

$$P = \int_{\mathscr{X}} \mu_x P(dx).$$

c. Une *section mesurable* $\mathscr{Y} \subset \mathscr{X}$ est définie par
(i) $\mathscr{Y}$ est mesurable ;
(ii) $\forall x \in \mathscr{X}$, $\mathscr{Y} \cap O_x = \{y(x)\}$; et
(iii) la fonction $t(x) = y(x)$ est mesurable pour la σ-algèbre induite par $\mathscr{X}$ sur $\mathscr{Y}$.
Montrer que, pour toute distribution de probabilité Q sur $\mathscr{Y}$,

$$P = \int_{\mathscr{Y}} \mu_y Q(dy)$$

est $\mathscr{G}$-invariant sur $\mathscr{X}$ et que, réciproquement, toute probabilité $\mathscr{G}$-invariante peut s'écrire de cette façon.

d. Soient U une variable aléatoire uniforme sur $\mathscr{G}$, $\mathscr{Y}$ une section mesurable de $\mathscr{X}$ et X une variable aléatoire sur $\mathscr{X}$. Déduire de c. l'équivalence entre les propriétés suivantes :
(i) la distribution de gX est indépendante de $g \in \mathscr{G}$; et
(ii) il existe Y, variable aléatoire sur $\mathscr{Y}$, indépendante de U, telle que UY a la même distribution que X.

e. Appliquer au cas $\mathscr{X} = \{0,1\}^n$.

9.19 On considère $x \sim \mathscr{N}(\theta, 1)$ et on s'intéresse plus particulièrement à la quantité $h(\theta) = e^{c\theta}$.

a. Déterminer le risque de l'estimateur de Bayes de $h(\theta)$ associé à $\pi(\theta) = 1$ et au coût quadratique, $R(\theta, \delta^\pi)$, et montrer que l'estimateur de Bayes de $h(\theta)$ associé à $\pi'(\theta) = R(\theta, \delta^\pi)^{-1}$ domine δ^π.

b. Remarquer que $R(\theta, \delta^\pi)^{-1}(e^{c\theta} - \delta)^2$ est un coût invariant et établir le résultat suivant : *Pour tout coût invariant* $\mathrm{L}(\theta, \delta)$*, si* δ^π *est l'estimateur associé à* L *et à la mesure de Haar* π*, et si* $\omega(\theta) = \mathbb{E}_\theta[\mathrm{L}(\theta, \delta^\pi(x))]$*, l'estimateur associé à* L *et* $\pi'(\theta) = \pi(\theta)/\omega(\theta)$ *est le meilleur estimateur équivariant.*

Section 9.5

9.20 *(Berger, 1985b) On considère le cas particulier où le groupe $\mathscr{G}$ est fini, c'est-à-dire

$$\mathscr{G} = \{g_1, \ldots, g_m\}.$$

On suppose que le coût $\mathrm{L}(\theta, a)$ est invariant, convexe en a, et, en outre, que l'action induite par le groupe $\mathscr{G}$ sur $\mathscr{D}$ satisfait

$$\tilde{g}\left(\frac{1}{m}\sum_{i=1}^m a_i\right) = \frac{1}{m}\sum_{i=1}^m \tilde{g}(a_i).$$

Démontrer le théorème de Hunt-Stein avec l'hypothèse supplémentaire que $\mathscr{D}$ est convexe. (*Indication :* Montrer que, pour tout estimateur δ, il existe un estimateur invariant associé δ^I qui domine δ.)

9.21 Dans le cadre de l'Exemple 9.25, déterminer le risque exact de l'estimateur δ_1. (*Indication :* Remarquer que z_1/z_2 est distribuée comme une variable aléatoire de Cauchy.)

9.22 Considérer l'estimation de ϱ dans l'Exemple 9.25 pour la structure d'invariance induite par le groupe multiplicatif.

9.23 Soient $(x_1, \ldots, x_p)$ et $(y_1, \ldots, y_p)$ de distributions normales $\mathcal{N}_p(0, \Sigma)$ et $\mathcal{N}_p(0, \Delta\Sigma)$. On teste l'hypothèse $H_0 : \Delta \leq \Delta_0$ contre $H_1 : \Delta > \Delta_0$.

a. Montrer que le problème est invariant sous $\mathcal{GL}_p$, groupe des transformations linéaires régulières.

b. Montrer que $\mathcal{GL}_p$ est transitif sur l'espace d'échantillonnage, à un ensemble de mesure nulle près. En déduire que les estimateurs équivariants sont constants, c'est-à-dire que les tests invariants de niveau α sont $\varphi_\alpha(x, y) = 1 - \alpha$.

c. Montrer que $\varphi_c(x, y) = \mathbb{I}_{y_1^2 \leq c x_1^2}$ domine φ_α sous le coût $0 - 1$ pour $\alpha = P_{\Delta_0}(y_1^2 > c x_1^2)$.

d. $\mathcal{GL}_p$ est-il moyennable ?

9.24 Soit l'échantillon $x_1, \ldots, x_n \sim \mathcal{C}(\mu, \sigma^2)$.

a. Montrer que la mesure de Haar est $\pi^H(\mu, \sigma) \propto 1/\sigma$.

b. On s'intéresse à la reparamétrisation $y_i = 1/x_i$. Montrer que $y_i \sim \mathcal{C}(\nu, \tau^2)$ et exprimer ν et τ en fonction de μ et σ.

c. Montrer que $\pi^H(\nu, \tau) \propto 1/\tau$ n'est pas la transformée de $\pi^H(\mu, \sigma)$ et conclure sur les limites de l'invariance pour la reparamétrisation.

Section 9.6

9.25 *(Villegas, 1990) On considère une famille de distributions de probabilité P_θ sur $\mathcal{X}$, avec $\theta \in \Theta$ et $T(x)$ à valeurs dans un espace affine euclidien E, telle que la fonction de vraisemblance soit

$$\ell(\theta|x) = c_1(x) c_2(\theta) \exp\{-||T(x) - \theta||^2/2\}.$$

Ce modèle est appelé *bayésien euclidien* si $\pi(\theta) = 1$.

a. En déduire que la distribution a priori euclidienne correspondante pour un modèle de Poisson $\mathcal{P}(\lambda)$ est $\pi(\lambda) = 1/\lambda$.

b. Montrer que la p-value $p(x) = P_{\lambda_0}(X \geq x)$ du test de $H_0 : \lambda \leq \lambda_0$ contre $H_1 : \lambda > \lambda_0$ est liée à la distribution a priori, mais que cela n'est pas vrai pour le test alternatif de $H_0 : \lambda \geq \lambda_0$ contre $H_1 : \lambda < \lambda_0$.

c. Montrer que la distribution a priori de Haldane

$$\pi(p) = \frac{1}{p(1-p)} \tag{9.7}$$

est également un modèle euclidien lorsque $x \sim \mathcal{B}(n, p)$. La loi (9.7) est-elle encore l'a priori euclidien pour la distribution binomiale négative $\mathcal{N}eg(n, p)$?

d. Si $0 < x < n$, montrer que, dans le cas binomial, les p-values $P_{p_0}(X \leq x)$ et $P_{p_0}(X \geq x)$ associées aux hypothèses $H_0 : p \geq p_0$ et $H_0 : p \leq p_0$ ne correspondent pas à la distribution euclidienne (9.7).

e. Dans le cas normal $\mathcal{N}(\mu, \sigma^2)$, montrer que les distributions euclidiennes a priori sont les suivantes :
(i) $\pi(\theta) = 1$ si $\theta = \mu$;
(ii) $\pi(\theta) = 1$ si $\theta = \sigma^{-2}$; et
(iii) $\pi(\theta) = \theta_2$ si $(\theta_1, \theta_2) = (\mu, \sigma^{-2})$.

9.26 Étudier les problèmes de compatibilité entre les exigences d'invariance et le principe de vraisemblance. Déterminer en particulier si l'estimateur du maximum de vraisemblance est toujours un estimateur invariant.

9.27 *Pour une fonction de coût arbitraire $\mathrm{L}(\theta,\delta)$ et une distribution a priori donnée π, on suppose que l'estimateur de Bayes δ^π est tel que $0 < R(\theta,\delta^\pi) < \infty$ quel que soit θ.

a. Si on définit $\mathrm{L}^\pi(\theta,\delta) = \mathrm{L}(\theta,\delta)/R(\theta,\delta^\pi)$, montrer que δ^π a un risque constant de 1. Cela implique-t-il que δ^π soit minimax ? (*Indication :* δ^π n'est pas nécessairement l'estimateur de Bayes sous L^π.)

b. On considère le cas particulier où $x \sim \mathcal{N}(\theta,1)$ et π est $\mathcal{N}(\theta_0,\tau^2)$. Calculer $R(\theta,\delta^\pi)$ et étudier le comportement de l'estimateur de Bayes associé à π et L^π, comparativement à δ^π (numériquement, si nécessaire).

c. Si δ_1^π, associé à π et $\mathrm{L}_1^\pi = \mathrm{L}^\pi$, est différent de δ^π, une suite d'estimateurs δ_n^π peut être définie récursivement avec $\mathrm{L}_n^\pi = \mathrm{L}_{n-1}^\pi/R(\theta,\delta_{n-1}^\pi)$. Que peut-on dire de la limite de la suite (δ_n^π) ?

d. Étudier la suite ci-dessus pour $x \sim \mathcal{P}(\lambda)$, $\pi(\lambda) = 1$ et $\mathrm{L}(\lambda,\delta) = (1-\lambda/\delta)^2$.

9.8 Notes

9.8.1 Invariance et paradoxes de marginalisation

En plus d'apporter un nouveau point de vue sur l'estimation équivariante, Helland (1999) considère que l'utilisation de mesures de Haar à droite est un moyen d'échapper aux *paradoxes de marginalisation*, comme l'ont observé Dawid *et al.* (1973). (Voir les Exercices 3.45-3.51.)

Plus précisément, étant donné un modèle $(\mathcal{X},\Theta,f(x|\theta))$ muni d'un groupe $\mathcal{G}$ agissant sur $\mathcal{X}$ et le groupe correspondant $\bar{\mathcal{G}}$ agissant sur Θ, une fonction $h(\theta)$ est dite *estimable invariablement* si $h(\theta_1) = h(\theta_2)$ entraîne $h(\bar{g}\theta_1) = h(\bar{g}\theta_2)$ pour tout $\bar{g} \in \bar{\mathcal{G}}$ (Hora and Buehler, 1966). Helland (1999) estime que la perspective d'invariance et l'utilisation de la mesure de Haar correspondante devraient être limitées à l'estimation de fonctions des paramètres estimables invariablement. Par exemple, bien que la mesure de Lebesgue soit la mesure de Haar à droite pour le groupe de translation et s'applique donc à la distribution normale $\mathcal{N}_p(\theta,I_p)$, elle ne devrait pas être utilisée pour l'estimation de $||\theta||^2$, à cause de l'inefficacité mise en évidence à la Section 3.5.4. Ce point de vue est d'une certaine façon lié à la construction de lois a priori de référence : étant donné un paramètre d'intérêt, on devrait d'abord déterminer la structure d'invariance adéquate puis en déduire la mesure de Haar à droite en tant qu'a priori non informatif correspondant. Deux défauts de cette approche sont (a) qu'il existe des fonctions pour lesquelles il est impossible de trouver des groupes d'invariance non triviaux et (b) qu'il y a toujours une part d'arbitraire dans le choix de ces groupes, lorsqu'ils existent.

Exemple 9.31. (Helland, 1999) Si $x \sim \mathcal{N}_p(\theta,\sigma^2 I_p)$ et θ est sur la sphère de rayon c, le meilleur estimateur équivariant de θ sous le groupe de rotations est associé à la mesure uniforme sur la sphère et est donné par (8.10). (Voir l'Exercice 4.37.) Si $c = ||\theta||^2$ est inconnu, il peut être estimé à partir de $||x||$ plutôt que de x (voir les Exemples 3.34 et 3.36), en utilisant par exemple l'estimateur du maximum de vraisemblance ou l'amélioration de Saxena et Alam (1982) $(||x||^2 - p)^+$. Si on insère l'expression de l'estimateur $\kappa(x)$ de c dans δ_c, on obtient l'estimateur

$$\tilde{\delta} = \kappa(x)\frac{I_{p/2}(\kappa(x)||x||)}{I_{p/2-1}(\kappa(x)||x||)}\,\frac{x}{||x||},$$

avec I_ν *fonction de Bessel modifiée*. Il se comporte comme un estimateur à rétrécisseur (2.16) pour de grandes valeurs de $||x||$, comme le montrent Bock et Robert (1985) et Beran (1996). (Voir l'Exercice 10.37.) ‖

Pour une fonction estimable invariablement $h(\theta)$, Helland (1999) considère le sous-groupe de $\bar{\mathscr{G}}$ suivant :

$$\bar{K} = \{\bar{g} \in \bar{\mathscr{G}};\, h(\bar{g}\theta) = h(\theta) \text{ pour tout } \theta \in \Theta\}\,,$$

puisque $h(\theta)$ est un invariant maximal pour $\bar{K}$. Étant donné le sous-groupe correspondant de $\mathscr{G}$,

$$K = \{g \in \mathscr{G};\, \bar{g} \in \bar{K}\}\,,$$

soit z une variable maximalement invariante pour K. Si $\eta = h(\theta)$, Helland (1999) montre que le paradoxe de marginalisation ne se produit pas pour (η, z) lorsqu'on utilise la mesure de Haar à droite associée à $\bar{\mathscr{G}}$. Cela signifie que, sous cette mesure, si $\theta = (\eta, \xi)$ et $x = (z, y)$, et si la distribution a posteriori marginale de η dépend seulement de z, elle peut être obtenue comme distribution a posteriori sur z uniquement.

10

Extensions hiérarchique et empirique

> "Books and papers and scrolls covered nearly every flat surface, with all sorts of odd things interspeded among the piles, and sometimes on top of them. Strange shapes of glass or metal, spheres and tubes interlinked, and circles held inside circles, stood among bones and skulls of every shape and description."
>
> **Robert Jordan**, *The Dragon Reborn.*

10.1 Lois a priori incomplètes

Dans les chapitres précédents, nous avons mis en avant (parfois avec insistance !) l'ambivalence de l'analyse bayésienne : elle a un potentiel d'élimination suffisant pour conduire à une prise de décision pratique mais cette efficacité doit être maîtrisée. Ainsi, les choix subjectifs de l'analyse bayésienne peuvent toujours être réglés pour aboutir à une conclusion fixée à l'avance. De tels travers sont certes également possibles dans un cadre fréquentiste, par le choix des fonctions de coût ou d'estimation, et l'approche classique ne fait même pas de distinction entre les parties objective et subjective d'une analyse. Mais notre message essentiel ici, déjà présenté dans le Chapitre 3, est que le choix d'une loi a priori par le statisticien devrait toujours être *justifiable*, c'est-à-dire établi à partir d'arguments sensés (ou "reproductibles"). Par conséquent, le fait que les *outils bayésiens* puissent donner des inférences erronées ne saurait être vu comme un défaut du *paradigme bayésien*.

Une critique plus pertinente est en revanche que l'information a priori est rarement assez riche pour en déduire une loi a priori exacte. Il paraît

alors nécessaire d'incorporer cette incertitude au modèle bayésien, bien que la notion de lois a priori semble insuffisante pour rendre pleinement compte de l'ignorance. La modélisation de cette incertitude résiduelle a inspiré des variations autour du paradigme bayésien, comme les *probabilités hautes et basses* de Dempster (1968) ou les *probabilités imprécises* de Walley (1991).

L'hypothèse de l'*analyse bayésienne hiérarchique* est au contraire que ces considérations peuvent être incorporées au paradigme bayésien. Il s'agit de modéliser l'information a priori en la décomposant en plusieurs niveaux de distributions a priori conditionnelles. On peut par là même distinguer les caractères structurel et subjectif de l'information. Suivant le paradigme bayésien, l'incertitude à tous les niveaux est prise en compte au moyen de lois a priori additionnelles. Dans les cas les plus simples, la structure hiérarchique ne contient que deux niveaux, les paramètres du premier étant associés à une distribution a priori définie dans le second. La distribution de premier niveau est en général une loi a priori conjuguée, un choix qui se justifie par la facilité de calcul mais aussi parce que le niveau le plus haut peut d'une certaine manière compenser les erreurs de modélisation des plus bas niveaux. (Une autre justification pour la modélisation conjuguée se trouve dans Dalal et Hall, 1983, et Diaconis et Ylvisaker, 1985 ; voir la Section 3.4.) Nous avons déjà étudié des exemples d'une telle modélisation dans le Chapitre 6, comme dans l'Exemple 6.21.

Une caractéristique générale de la modélisation hiérarchique est qu'elle améliore la robustesse des estimateurs de Bayes obtenus : tout en intégrant l'information a priori, ces estimateurs sont également performants d'un point de vue fréquentiste (minimaxité et admissibilité), même si ces deux critères sont souvent difficiles à concilier.

On trouve d'autres justifications à la modélisation bayésienne hiérarchique dans les problèmes réels, puisqu'il existe des cadres en médecine, biologie, élevage animalier, économie, etc., dans lesquels la population à laquelle on s'intéresse peut être vue comme sous-population d'une population, voire comme sous-population d'une sous-population d'une population globale. C'est, par exemple, le cas en *méta analyse*, lorsqu'on cherche à rassembler les résultats de plusieurs expériences concernant le même phénomène mais réalisées sur différents lieux ou populations et suivant divers protocoles (voir, par exemple, Mosteller et Chalmers, 1992, Mengersen et Tweedie, 1995, ou Givens *et al.*, 1997).

Exemple 10.1. (Guihenneuc-Jouyaux *et al.*, 1998) Le *virus de l'immunodéficience humaine* (VIH) est le virus responsable du SIDA. Pour un patient donné, la transition de l'infection VIH vers le SIDA peut être représentée par sept étapes de gravité croissante, la dernière étant celle du SIDA. Les déplacements entre les états sont représentés par un modèle de Markov à temps continu avec *générateur infinitésimal* Λ. (Cela signifie que la distribu-

tion de l'état à l'instant T, connaissant la distribution $\omega_0 = (\omega_{01}, \ldots, \omega_{07})$ à l'instant 0, est donnée par le produit matriciel $\omega_0 \cdot \exp\{T\Lambda\}$[73].)

Les six premières étapes d'infection du VIH ne sont pas observables directement mais seulement par le biais de variables aléatoires ($1 \leq i \leq n\,, 1 \leq j \leq n_i$),

$$x_{ij} \sim \mathcal{N}\left(\mu_{S_{ij}}, \sigma^2\right)\,,$$

où i désigne l'individu et j le point d'évolution, $1 \leq S_{ij} \leq 6$ étant l'étape du VIH. Les x_{ij} représentent des marqueurs sanguins (taux de T4) sujets à une grande variabilité et à des erreurs de mesure. Il s'agit d'un cas particulier de *modèle de Markov caché* (Exercice 6.50), avec la difficulté supplémentaire que la chaîne de Markov cachée opère à temps continu. Mais l'algorithme forward-backward vu dans l'Exercice 6.51 s'applique aussi ici (Exercice 10.2). Un modèle similaire a été proposé par Kirby et Spiegelhalter (1994).

Les S_{ij} constituent donc le premier niveau d'un modèle hiérarchique, avec des hyperparamètres, comme la matrice génératrice Λ, correspondant au second niveau et commun à tous les individus. Un autre hyperparamètre est δ, la distribution a priori de l'étape du VIH à la première observation. Il est souvent pratique de représenter ces modèles hiérarchiques sous forme de graphes (plus précisément de *graphes acycliques orientés* ou DAG (pour Directed Acyclic Graph)). La Figure 10.1 donne cette représentation pour le modèle VIH, avec la convention usuelle que les *rectangles* correspondent à des quantités observées ou connues et les *cercles* aux quantités inconnues ; les *flèches* symbolisent la dépendance probabiliste. (Voir la Note 10.7.1 et Lauritzen, 1996, pour plus de détails sur les modèles graphiques.) ‖

L'analyse bayésienne empirique part de la même idée d'imprécision sur l'information a priori, mais la traite à un niveau plus pragmatique. On considère dans cette optique qu'il est illusoire d'espérer modéliser cette imprécision sur plusieurs niveaux de lois conditionnelles, alors même que le premier niveau est déjà très faiblement connu. De façon assez paradoxale, l'analyse bayésienne empirique repose elle aussi sur une modélisation a priori conjuguée, en estimant les hyperparamètres à partir des observations et en utilisant ensuite cet "a priori estimé" comme a priori normal pour l'inférence. Il va sans dire que le remplacement des hyperparamètres par des hyperparamètres estimés, qui constitue la base de l'analyse bayésienne empirique, l'exclut de fait du paradigme bayésien. Mais elle permet au statisticien de tirer parti de l'information a priori vague de manière simplifiée. En outre, il se trouve que les estimateurs ainsi construits ont souvent de bonnes propriétés fréquentistes même s'il y a trop d'arbitraire dans la détermination des hyperparamètres pour en faire une règle générale. Un avantage annexe de la modélisation bayésienne empirique est de fournir des justifications bayésiennes à l'effet de Stein (Note

[73] L'extension de la fonction exponentielle aux cas multivariés comme celui-ci est obtenue en utilisant le développement en série de $\exp(x)$. Voir l'Exercice 10.2.

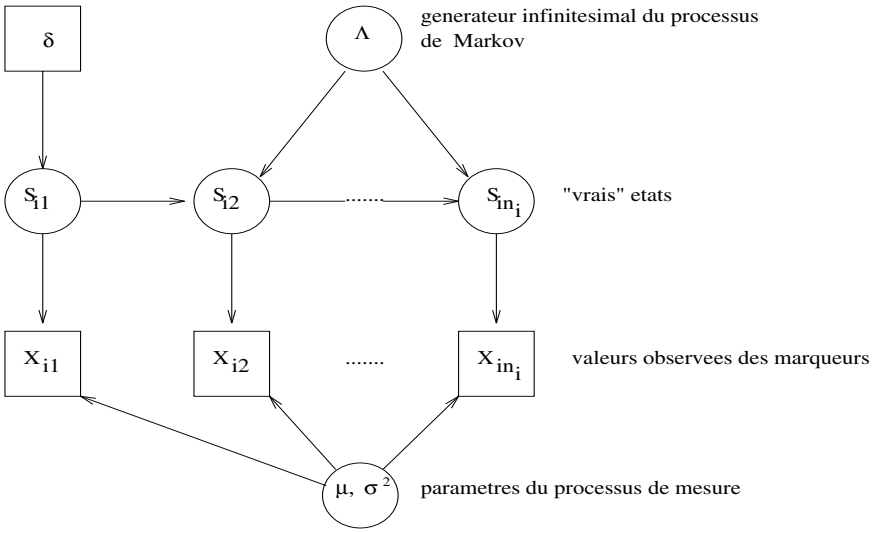

Fig. 10.1. Graphe acyclique orienté du modèle hiérarchique. (*Source :* Guihenneuc-Jouyaux *et al.*, 1998.)

2.8.2). L'analyse bayésienne empirique se présente enfin comme une alternative attrayante lorsque l'analyse bayésienne hiérarchique est trop compliquée à mettre en œuvre, même si cet argument est de moins en moins justifié avec l'efficacité grandissante des techniques de calcul (voir le Chapitre 6).

10.2 Analyse bayésienne hiérarchique

Cette section n'est qu'une courte introduction à l'analyse bayésienne hiérarchique et elle cible plus particulièrement quelques aspects intéressants de cette approche. Pour un traitement plus exhaustif, voir Berger (1985b), en lien avec la notion de robustesse, Deely et Lindley (1981), Dumouchel et Harris (1983), George (1986b), Angers et MacGibbon (1990), Gelman *et al.* (2003), Hobert (2000b) et Draper (1995). Pour les applications à l'élevage animalier, voir, par exemple, Fouley *et al.* (1992).

10.2.1 Modèles hiérarchiques

Pour des raisons liées à la modélisation des observations ou à la décomposition de l'information a priori, il peut arriver que le modèle statistique bayésien soit *hiérarchique*, c'est-à-dire mette en jeu plusieurs niveaux de distributions a priori conditionnelles.

Définition 10.2. *Un* modèle bayésien hiérarchique *est un modèle statistique bayésien* $(f(x|\theta),\ \pi(\theta))$*, dans lequel la loi a priori* $\pi(\theta)$ *est décomposée en plusieurs lois conditionnelles*

$$\pi_1(\theta|\theta_1),\ \pi_2(\theta_1|\theta_2),\ldots,\ \pi_n(\theta_{n-1}|\theta_n)$$

et une loi marginale $\pi_{n+1}(\theta_n)$ *telle que*

$$\pi(\theta)=\int_{\Theta_1\times\ldots\times\Theta_n}\pi_1(\theta|\theta_1)\pi_2(\theta_1|\theta_2)\cdots\pi_{n+1}(\theta_n)\,d\theta_1\cdots d\theta_{n+1}. \quad (10.1)$$

Les paramètres θ_i *sont appelés* hyperparamètres de niveau i $(1\le i\le n)$.

Avant d'insister sur l'utilité d'une telle décomposition, remarquons qu'on trouve également des structures hiérarchiques dans des modèles statistiques classiques.

Exemple 10.3. Un cas typique d'utilisation de modèles hiérarchiques est la prise en compte d'*effets aléatoires* au sein d'un modèle linéaire. Cette extension peut s'écrire sous la forme

$$\begin{aligned} y|\theta &\sim \mathcal{N}_p(\theta,\Sigma_1),\\ \theta|\beta &\sim \mathcal{N}_p(X\beta,\Sigma_2),\end{aligned}$$

sans lien avec une modélisation bayésienne. La moyenne de y, θ, est décomposée en *effets fixes*, $X\beta$, et en *effets aléatoires*, $Z\eta$, avec η normale de moyenne 0 (la matrice de covariance Σ_2 peut alors être singulière). Ces modèles sont souvent employés en biométrie, en particulier en *amélioration de races animales*, pour différencier l'influence des éléments fixes (par exemple, lignée, race, année, etc.) de celle des facteurs aléatoires (par exemple, femelles dans une lignée). ‖

Un autre exemple classique de structure hiérarchique non bayésienne est celui des *modèles à variables latentes*, comme les mélanges (Section 6.4) ou les mélanges cachés (Note 6.6.3). Le vecteur des variables latentes z constitue alors le premier niveau du modèle bayésien hiérarchique, la modélisation a priori en tant que telle ayant lieu à des niveaux plus élevés.

Ces exemples montrent bien que la frontière entre modèles hiérarchiques classiques et bayésiens est parfois floue et dépend essentiellement de l'interprétation des paramètres. Par exemple, dans le modèle à effets aléatoires, la procédure est classique si l'inférence porte sur les effets fixes (β) mais bayésienne si on considère l'effet global (θ). De même, si on s'intéresse aux variables latentes z_t, comme dans le modèle à volatilité stochastique (4.41) pour les y_t^*, il s'agit de *données* manquantes alors que, si elles sont utilisées pour représenter le modèle de façon plus pratique, comme dans le cas des

mélanges utilisés pour la modélisation non paramétrique, les z_t peuvent être considérés comme partie intégrante de la modélisation a priori.

Insistons sur le fait qu'un modèle bayésien hiérarchique n'est rien d'autre qu'un cas particulier de modèle bayésien. Ainsi, si

$$x \sim f(x|\theta),\ \theta \sim \pi_1(\theta|\theta_1), \ldots,\ \theta_n \sim \pi_{n+1}(\theta_n), \tag{10.2}$$

on retrouve le modèle bayésien usuel

$$x \sim f(x|\theta), \quad \theta \sim \pi(\theta),$$

pour l'a priori

$$\pi(\theta) = \int_{\Theta_1 \times \ldots \times \Theta_n} \pi_1(\theta|\theta_1) \ldots \pi_n(\theta_{n-1}|\theta_n)\pi_{n+1}(\theta_n)\, d\theta_1 \cdots d\theta_n.$$

Cela montre que les modèles hiérarchiques s'intègrent bien au paradigme bayésien et donc que cette approche bénéficie des propriétés générales d'optimalité de la perspective bayésienne avec quelques avantages additionnels liés à la décomposition de la loi a priori (voir la Section 10.3). Cela montre également pourquoi il est rarement nécessaire d'aller plus loin que deux niveaux de décomposition conditionnelle dans la hiérarchie. Si les hyperparamètres $\theta_1, \ldots, \theta_n$ ne sont d'aucun intérêt pour l'inférence (sur θ), il est équivalent de considérer le modèle hiérarchique plus simple

$$x|\theta \sim f(x|\theta), \qquad \theta|\theta_1 \sim \pi_1(\theta|\theta_1)\,,$$

avec

$$\theta_1 \sim \pi_2(\theta_1) = \int_{\Theta_2 \times \ldots \times \Theta_n} \pi_1(\theta_1|\theta_2) \cdots \pi_{n+1}(\theta_n)\ d\theta_2 \cdots d\theta_n\,,$$

qui élimine les étapes intermédiaires et les hyperparamètres supplémentaires. Néanmoins, une décomposition plus compliquée peut toujours se justifier pour la construction et le calcul pratique d'estimateurs de Bayes, comme nous l'avons vu dans le Chapitre 6.

Exemple 10.4. Robert et Reber (1998) étudient une expérience dans laquelle des rats sont intoxiqués par une substance, puis traités par un placebo ou un médicament. Le modèle associé à cette expérience est un *modèle linéaire à effets additifs* : étant donné x_{ij}, y_{ij} et z_{ij}, j-ièmes réponses du rat i aux étapes respectivement de contrôle, d'intoxication et de traitement, on suppose que $(1 \le i \le I)$

$$\begin{array}{ll} x_{ij} \sim \mathcal{N}(\theta_i, \sigma_c^2), & 1 \le j \le J_i^c\,, \\ y_{ij} \sim \mathcal{N}(\theta_i + \delta_i, \sigma_a^2), & 1 \le j \le J_i^a\,, \\ z_{ij} \sim \mathcal{N}(\theta_i + \delta_i + \xi_i, \sigma_t^2), & 1 \le j \le J_i^t\,, \end{array}$$

où θ_i est la mesure de contrôle moyenne, δ_i l'effet moyen d'intoxication et ξ_i l'effet moyen de traitement pour le rat i, les variances de ces mesures étant

constantes pour les effets de contrôle, d'intoxication et de traitement. Une variable (observée) supplémentaire est w_i, qui est égale à 1 si le rat est traité avec le médicament, et 0 sinon.

Puisque le but de l'expérience est de déterminer l'effet global du médicament testé, les différentes moyennes individuelles sont mises en relation par une loi a priori commune (conjuguée) $(1 \leq i \leq I)$,

$$\theta_i \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2), \qquad \delta_i \sim \mathcal{N}(\mu_\delta, \sigma_\delta^2),$$

et

$$\xi_i \sim \mathcal{N}(\mu_P, \sigma_P^2) \qquad \text{ou} \qquad \xi_i \sim \mathcal{N}(\mu_D, \sigma_D^2),$$

suivant que le rat i soit traité avec un placebo ou avec un médicament. Les hyperparamètres du modèle,

$$\mu_\theta, \mu_\delta, \mu_P, \mu_D, \sigma_c, \sigma_a, \sigma_t, \sigma_\theta, \sigma_\delta, \sigma_P, \sigma_D\,,$$

sont alors associés aux lois a priori non informatives de Jeffreys. Cet a priori permet de déduire une loi a posteriori bien définie pourvu qu'il y ait au moins deux observations pour chaque étape de l'expérience. ‖

10.2.2 Justifications

L'analyse bayésienne hiérarchique s'appuie en partie sur les travaux de Good (voir Good, 1980, 1983, pour plus de détails). Lindley et Smith (1972) traitent le cas particulier des modèles linéaires, en jouant sur la dualité entre l'analyse bayésienne classique d'un modèle à effets aléatoires et l'analyse bayésienne hiérarchique d'un modèle de régression standard. Bien qu'un modèle bayésien hiérarchique ne soit qu'un cas particulier de modèle bayésien, comme en témoigne l'équation (10.1), la décomposition

$$\pi(\theta) = \int_{\Theta_1} \pi_1(\theta|\theta_1)\pi_2(\theta_1)\,d\theta_1$$

ou sa généralisation (10.1) peuvent être privilégiées pour un certain nombre de raisons :

(i) Les deux premiers niveaux de la hiérarchie peuvent être suggérés par des raisons *objectives* liées à la modélisation du phénomène observé comme cas particulier d'une métapopulation sur laquelle on dispose de connaissances a priori, ce qui justifie le recours à l'approche bayésienne. C'est le cas des Exemples 10.3 et 10.4. Plus généralement, comme nous l'avons vu en Section 10.1, les modèles bayésiens hiérarchiques interviennent naturellement en méta analyse lorsqu'on doit regrouper les résultats de différentes études.

Exemple 10.5. (Berger, 1985b) On considère $x_i \sim \mathcal{N}(\beta_i, 10)$ $(i = 1, \ldots, 7)$, correspondant à des mesures annuelles indépendantes du quotient intellectuel (QI) d'un enfant, sur sept années consécutives. Dans la mesure où les tests de QI intègrent une correction tenant compte de l'âge, il est raisonnable de considérer que les β_i ont la même moyenne θ, qui est la "vraie" valeur du QI. On peut alors poser la loi a priori de premier niveau suivante :

$$\beta_i|\theta \sim \mathcal{N}(\theta, \sigma_\pi^2) \qquad (i = 1, \ldots, 7).$$

De plus, si l'enfant fait partie d'un ensemble bien identifié, on peut disposer d'une information sur cette population comme, par exemple,

$$\theta \sim \mathcal{N}(\xi, \tau^2),$$

avec ξ et τ connus. Nous obtenons ainsi le deuxième niveau d'analyse. Au contraire, une alternative non informative serait de prendre $\pi_2(\theta) = 1$. ‖

(ii) En poussant plus loin la justification ci-dessus, un chercheur peut vouloir diviser la modélisation a priori en deux parties, la première correspondant à l'information *structurelle* concernant le modèle et la seconde correspondant à une information plus *subjective*. Par exemple, l'information peut être liée à des restrictions linéaires imprécises sur les paramètres d'un modèle de régression et la loi des hyperparamètres $\pi_2(\theta_1)$ peut prendre en compte le caractère incertain de ces restrictions.

Exemple 10.6. Albert (1988) s'intéresse aux incertitudes sur les modèles linéaires généralisés (McCullagh et Nelder, 1989) $(i = 1, \ldots, n)$

$$y_i|x_i \sim \exp\{\theta_i \cdot y_i - \psi(\theta_i)\}\,, \qquad \nabla\psi(\theta_i) = \mathbb{E}[y_i|x_i] = h(x_i^t\beta)\,,$$

où h est la fonction de *lien* et $x_i \in \mathbb{R}^q$ un vecteur de covariables, en reportant la contrainte de linéarité $\nabla\psi(\theta_i) = h(x_i^t\beta)$ à un niveau plus haut de hiérarchie, c'est-à-dire en introduisant l'a priori conjugué

$$\theta_i \sim \exp\{\lambda\,[\theta_i \cdot \xi_i - \psi(\theta_i)]\}$$

tel que $\mathbb{E}[\nabla\psi(\theta_i)] = h(x_i^t\beta)$. Le paramètre de régression β est alors transféré à un second niveau avec, éventuellement, un a priori normal $\beta \sim \mathcal{N}_q(0, \tau^2 I_q)$, qui admet comme cas limite $\tau = \infty$ l'a priori constant. La variance a posteriori de $\psi(\theta_i)$ est alors un indicateur de la précision du modèle linéaire généralisé et permet donc de mesurer la pertinence de l'hypothèse de linéarité. ‖

Exemple 10.7. (Suite de l'Exemple 10.4) Une alternative également considérée dans Robert et Reber (1998) est de choisir comme loi a priori

$$\delta_i \sim p\mathcal{N}(\mu_{\delta 1}, \sigma^2_{\delta 1}) + (1-p)\mathcal{N}(\mu_{\delta 2}, \sigma^2_{\delta 2}), \tag{10.3}$$

qui introduit deux niveaux différents d'intoxication, c'est-à-dire deux réactions à l'intoxication au sein de la population de rats. Comme l'expliquent Robert et Reber (1998), il existe des raisons liées au métabolisme justifiant cette modification de la loi a priori. Même si la structure de mélange se répercute sur les lois marginales des y_{ij}, elle est différente d'un modèle de mélange habituel puisqu'elle exige que les y_{ij} pour $1 \leq j \leq J_i^a$ appartiennent à la *même* composante de mélange. ‖

(iii) À l'inverse, dans un *cadre non informatif*, un modèle bayésien hiérarchique est un compromis entre les lois non informatives de Jeffreys, qui sont diffuses mais parfois difficiles à utiliser et à expliquer, et les lois conjuguées, qui sont subjectivement peu justifiables mais numériquement pratiques. Lorsque les hyperparamètres ont une *hyperdistribution* a priori (ou *hyper a priori*), on fait un pas vers le non informatif, tout en étant généralement capable d'établir la loi a posteriori de θ. Une option est d'itérer cet argument par l'introduction d'une loi conjuguée sur θ_1, $\pi_2(\theta_1|\theta_2)$ et une loi non informative sur θ_2. Néanmoins, l'introduction d'une loi conjuguée sur θ_1 ne permet plus forcément de garantir que l'estimateur de Bayes soit calculable analytiquement et, pire encore, ne semble pas améliorer la robustesse du modèle. Quel que soit le nombre de niveaux dans la distribution, intégrer sur les paramètres inconnus ne peut que renforcer la robustesse de la loi a priori par rapport à une approche conjuguée classique. Les lecteurs pourront lire Berger (1985b) pour comprendre en quoi la modélisation hiérarchique est intéressante du point de vue de la robustesse.

Exemple 10.8. On considère le modèle de régression classique, $y = X\beta + \epsilon$, c'est-à-dire $y \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$, avec $\beta \in \mathbb{R}^p$. Pour des raisons structurelles, les coefficients de régression sont presque les mêmes. Par exemple, les β_i peuvent décrire les taux d'investissement de plusieurs constructeurs automobiles européens qui sont généralement assez proches. On suppose alors que $\beta_i \sim \mathcal{N}(\xi, \sigma^2_\pi)$, ξ étant la valeur usuelle. Un tel modèle est dit *échangeable* (voir la Note 3.8.2, Bernardo et Smith, 1994, et Gelman *et al.*, 2003). Si on dispose de plus d'informations sur la valeur habituelle, on peut prendre $\xi = \xi_0$ ou $\xi \sim \mathcal{N}(\xi_0, \tau^2)$. Sinon, le second niveau peut être non informatif : $\pi_2(\xi) = 1$. ‖

Exemple 10.9. (Suite de l'Exemple 10.3) Dans le cadre du modèle linéaire à effets aléatoires,

$$\begin{aligned} y|\theta &\sim \mathcal{N}_p(\theta, \Sigma_1), \\ \theta|\beta &\sim \mathcal{N}_p(X\beta, \Sigma_2), \end{aligned}$$

Lindley et Smith (1972) et Smith (1973) supposent que β vérifie également une relation linéaire et utilisent l'a priori suivant :

$$\beta \sim \mathcal{N}_n(Z\xi, \Sigma_3).$$

Un a priori alternatif assurant plus de robustesse est

$$\beta \sim \mathcal{T}_n(\alpha, Z\xi, \Sigma_3),$$

mais cette distribution met en jeu un niveau de hiérarchie supplémentaire par rapport à la distribution normale originale, comme le montre Dickey (1968). En effet, on a

$$\beta|z \sim \mathcal{N}_p(Z\xi, \Sigma_3/z), \qquad z \sim \mathcal{G}(\alpha/2, \alpha/2),$$

dans ce cas. Si on considère $\beta|z \sim \mathcal{N}_p(\mu, z\Sigma_3)$ (loi conjuguée) et $\pi(z) = 1/z$ (loi non informative), la loi marginale

$$\beta \sim \mathcal{T}_p(p/2, \mu, \Sigma_3),$$

est propre, contrairement au cas de la loi non informative $\pi(\beta) = 1$. ‖

(iv) Un autre aspect positif de l'analyse bayésienne hiérarchique est qu'elle augmente également la robustesse de l'analyse bayésienne classique d'un point de vue fréquentiste, puisqu'elle réduit l'arbitraire sur le choix de l'hyperparamètre (parfois reporté à un niveau plus élevé) et établit une moyenne des réponses bayésiennes conjuguées. La Section 10.3 montre que, dans le cas normal, de nombreuses lois a priori sur les hyperparamètres donnent des estimateurs de Bayes généralisés minimax.

(v) Un dernier avantage de l'approche bayésienne hiérarchique est sa capacité à souvent simplifier les *calculs bayésiens*. La décomposition d'une loi a priori π en plusieurs composantes $\pi_1, \ldots, \pi_n$ (qui peuvent être, par exemple, des lois conjuguées) permet parfois d'obtenir des approximations plus aisées de certaines quantités a posteriori par simulation, comme nous l'avons déjà mentionné en Section 6.3.5 au sujet de l'échantillonnage de Gibbs.

10.2.3 Décompositions conditionnelles

Une caractéristique particulièrement intéressante des modèles hiérarchiques est que le conditionnement est possible à tous les niveaux et cette liberté dans la décomposition de la loi a posteriori compense l'augmentation apparente de complexité de la structure. Par exemple, si

$$\theta|\theta_1 \sim \pi_1(\theta|\theta_1), \qquad \theta_1 \sim \pi_2(\theta_1),$$

nous avons le résultat suivant.

Lemme 10.10. *La loi a posteriori de θ est*

$$\pi(\theta|x) = \int_{\Theta_1} \pi(\theta|\theta_1, x)\pi(\theta_1|x)\, d\theta_1,$$

avec

$$\begin{aligned}
\pi(\theta|\theta_1, x) &= \frac{f(x|\theta)\pi_1(\theta|\theta_1)}{m_1(x|\theta_1)},\\
m_1(x|\theta_1) &= \int_{\Theta} f(x|\theta)\pi_1(\theta|\theta_1)\, d\theta,\\
\pi(\theta_1|x) &= \frac{m_1(x|\theta_1)\pi_2(\theta_1)}{m(x)},\\
m(x) &= \int_{\Theta_1} m_1(x|\theta_1)\pi_2(\theta_1)\, d\theta_1.
\end{aligned}$$

En outre, cette décomposition est valide pour les moments a posteriori, c'est-à-dire pour toute fonction h, on a

$$\mathbb{E}^{\pi}[h(\theta)|x] = \mathbb{E}^{\pi(\theta_1|x)}\left[\mathbb{E}^{\pi_1}\left[h(\theta)|\theta_1, x\right]\right],$$

où

$$\mathbb{E}^{\pi_1}[h(\theta)|\theta_1, x] = \int_{\Theta} h(\theta)\pi(\theta|\theta_1, x)\, d\theta.$$

Ce résultat découle naturellement du théorème de Bayes, la dernière égalité provenant du théorème de Fubini. Il n'en a pas moins des conséquences importantes sur le calcul des estimateurs de Bayes puisqu'il montre qu'on peut simuler $\pi(\theta|x)$ en générant d'abord θ_1 selon $\pi(\theta_1|x)$ puis θ selon $\pi(\theta|\theta_1, x)$, dans le cas où ces deux lois conditionnelles sont plus accessibles.

Exemple 10.11. (Suite de l'Exemple 10.4) La loi a posteriori du vecteur de paramètres complet s'écrit

$$\begin{aligned}
&\pi((\theta_i, \delta_i, \xi_i)_i, \mu_\theta, \ldots, \sigma_c, \ldots|\mathscr{D}) \propto\\
&\prod_{i=1}^{I} \left\{\exp -\{(\theta_i - \mu_\theta)^2/2\sigma_\theta^2 + (\delta_i - \mu_\delta)^2/2\sigma_\delta^2\}\right.\\
&\prod_{j=1}^{J_i^c} \exp -\{(x_{ij} - \theta_i)^2/2\sigma_c^2\} \prod_{j=1}^{J_i^a} \exp -\{(y_{ij} - \theta_i - \delta_i)^2/2\sigma_a^2\}\\
&\left.\prod_{j=1}^{J_i^t} \exp -\{(z_{ij} - \theta_i - \delta_i - \xi_i)^2/2\sigma_t^2\}\right\}\\
&\prod_{\ell_i=0} \exp -\{(\xi_i - \mu_P)^2/2\sigma_P^2\} \prod_{\ell_i=1} \exp -\{(\xi_i - \mu_D)^2/2\sigma_D^2\} \qquad (10.4)\\
&\sigma_c^{-\sum_i J_i^c - 1}\sigma_a^{-\sum_i J_i^a - 1}\sigma_t^{-\sum_i J_i^t - 1}(\sigma_\theta\sigma_\delta)^{-I-1}\sigma_D^{-I_D-1}\sigma_P^{-I_P-1},
\end{aligned}$$

où $\mathscr{D}$ désigne l'échantillon. Les lois marginales a posteriori des paramètres d'intérêt ne s'intègrent donc pas analytiquement et ne permettent pas d'obtenir des formules explicites pour les espérances a posteriori de ces paramètres. Néanmoins, on peut obtenir les lois conditionnelles complètes, comme le montre l'Exercice 10.14. ‖

Naturellement, le Lemme 10.10 n'est valable que lorsque les différentes intégrales sont bien définies. Mais ce n'est pas toujours le cas puisque les lois de second niveau sont généralement impropres. Le lemme suivant donne une condition suffisante d'existence des moments a posteriori pour $x|\theta \sim \mathscr{N}_p(\theta, \Sigma)$ (voir Berger et Robert, 1990, pour une démonstration).

Lemme 10.12. *Si la loi marginale*

$$m(x) = \int_\Theta f(x|\theta)\pi(\theta)\, d\theta$$

est finie pour tout $x \in \mathbb{R}^k$*, alors la moyenne et la variance de la loi a posteriori* $\pi(\theta|x)$ *existent toujours.*

Le résultat suivant porte sur un autre avantage des modèles hiérarchiques, à savoir leur influence dans le calcul des estimateurs bayésiens hiérarchiques :

Lemme 10.13. *Pour le modèle hiérarchique (10.2), la densité conditionnelle complète de* θ_i *sachant* x *et les* θ_j $(j \neq i)$ *vérifie*

$$\pi(\theta_i|x, \theta, \theta_1, \ldots, \theta_n) = \pi(\theta_i|\theta_{i-1}, \theta_{i+1})$$

avec la convention $\theta_0 = \theta$ *et* $\theta_{n+1} = 0$*.*

Preuve. Puisque

$$\begin{aligned}\pi(\theta_i|x, \theta, \theta_1, \ldots, \theta_n) &\propto f(x|\theta)\pi_1(\theta|\theta_1)\cdots\pi_{n+1}(\theta_{n+1}) \\ &\propto \pi_{i-1}(\theta_{i-1}|\theta_i)\pi_i(\theta_i|\theta_{i+1})\,,\end{aligned}$$

la distribution a posteriori ne dépend que des deux niveaux adjacents de la hiérarchie. □

L'importance de ce résultat pourtant simple réside dans le fait que seuls des hyperparamètres locaux interviennent dans les lois conditionnelles d'un modèle hiérarchique. Dans des cadres comme les modèles graphiques ou spatiaux, où la densité jointe est définie localement sur un groupe d'hyperparamètres (appelé *clique* dans les modèles graphiques), le Lemme 10.13 montre que les techniques numériques telles que l'échantillonneur de Gibbs (Section 6.3.3) sont les seules envisageables pour traiter ces modèles complexes.

10.2.4 Problèmes numériques

Un inconvénient des modèles hiérarchiques est qu'ils ne permettent en général pas un calcul explicite des estimateurs de Bayes, même lorsque les niveaux successifs sont conjugués, et il faut donc avoir recours à des techniques numériques d'approximation.

Exemple 10.14. On considère $x \sim \mathscr{B}(n,p)$ et $p|m \sim \mathscr{B}e(m,m)$ avec $m \in \mathbb{N}^*$. Alors,

$$\begin{aligned}\pi_1(p|m) &= \frac{\Gamma(2m)}{\Gamma(m)^2}\,[p(1-p)]^{m-1}\\ &= (2m-1)\binom{2m-2}{m-1}\,[p(1-p)]^{m-1}.\end{aligned}$$

Si la loi a priori de second niveau est $\pi_2(m) = 1/(2m-1)$, la loi a priori sur p est

$$\begin{aligned}\pi(p) &= \int_{\mathbb{N}^*} \pi_1(p|m)\pi_2(m)\,dm\\ &= \sum_{n=0}^{+\infty}\binom{2n}{n}[p(1-p)]^n.\end{aligned}$$

La loi a posteriori

$$\pi(p|x) = \int \pi_1(p|m,x)\pi_2(m|x)\,dm$$

ne peut être obtenue analytiquement puisque même si $\pi(p|m,x)$ est une loi bêta $\mathscr{B}e(m+x, m+n-x)$, $\pi(m|x)$ est la loi bêta-binomiale

$$\frac{(m+x-1)\dots m\,(m+n-x-1)\dots m}{(2m+n-1)\dots(2m)(2m-1)}$$

à un facteur de normalisation près. Les quantités a posteriori comme $\mathbb{E}^\pi[p|x]$ ne sont pas calculables analytiquement. ‖

La solution la plus naturelle en analyse hiérarchique est de faire appel à des outils de simulation. En effet, comme nous l'avons vu ci-dessus, la décomposition issue des Lemmes 10.10 et 10.13 est particulièrement pertinente dans ce contexte, puisqu'elle permet de simuler naturellement par l'échantillonneur de Gibbs ou d'autres techniques MCMC (Sections 6.3.2 et 6.3.3). Cela était déjà manifeste dans les exemples de la Section 6.3.5 et ceux qui suivent ne font que confirmer l'adéquation entre modèles hiérarchiques et méthodes MCMC (voir aussi Gelman *et al.*, 2003 , et Robert et Casella, 1999, Section 7.1.6).

Exemple 10.15. Soient

$$x \sim \mathcal{N}_p(\theta, \Sigma) \qquad \text{et} \qquad \theta|\mu, \xi \sim \mathcal{N}_p(\mu, B(\xi))\,,$$

avec $B(\xi) = \xi C - \Sigma$. La matrice définie positive C est fixée et ξ varie sur la demi-droite $[\lambda_{\max}(C^{-1}\Sigma), +\infty)$, où $\lambda_{\max}(A)$ désigne la plus grande valeur propre de A. Cette représentation de la matrice de covariance a posteriori simplifie les calculs tout en garantissant la robustesse des estimateurs. Par exemple, une modélisation de second niveau sur (μ, ξ) peut impliquer des lois non informatives. Cependant, une hypothèse commune est de supposer que $\mu = Y\beta$ pour $\beta \in \mathbb{R}^k$ et pour un régresseur donné Y tel que Y^tCY soit de rang plein, avec une loi non informative sur β. On peut alors montrer que $m(x) < +\infty$ si $p > 2 + k$ (Exercice 10.12). ‖

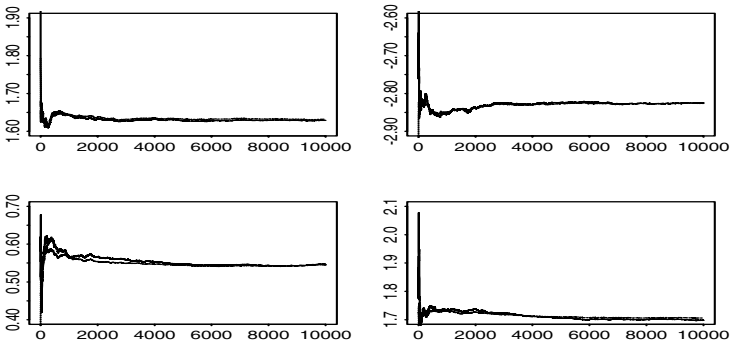

Fig. 10.2. Courbes de convergence pour μ_θ *(en haut à gauche)*, μ_δ *(en haut à droite)*, μ_P *(en bas à gauche)* et μ_D *(en bas à droite)* dans l'expérience de l'Exemple 10.4. Les courbes en pointillés qui représentent les moyennes Rao–Blackwellisées partielles sont presque indiscernables des moyennes standard. (*Source :* Robert et Reber, 1998.)

Exemple 10.16. (Suite de l'Exemple 10.4) Puisque les distributions conditionnelles complètes correspondent aux distributions standard (Exercice 10.14), l'échantillonneur de Gibbs est applicable. La Figure 10.2 montre la convergence des espérances a posteriori des quatre moyennes, en fonction du nombre d'itérations k, à la fois pour la moyenne partielle et l'estimateur de *Rao-Blackwell* (voir la Section 6.3.4). Étant donné que les deux quantités convergent vers le même estimateur de Bayes, la forte ressemblance des deux courbes est un indicateur partiel de convergence, ce qui suggère que dix mille itérations d'un échantillonneur de Gibbs devraient être suffisantes pour assurer la stabilité.

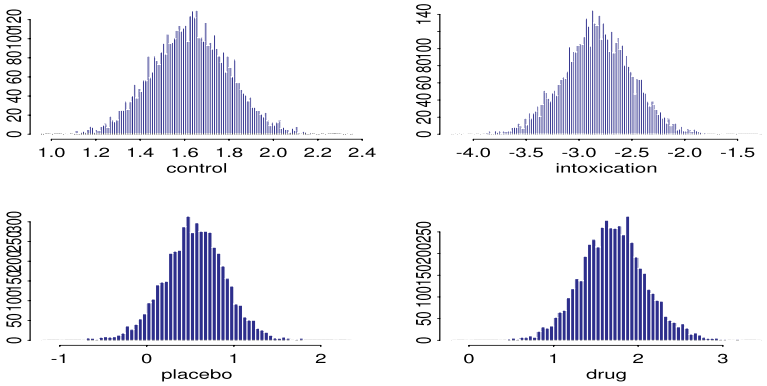

Fig. 10.3. Histogrammes des échantillons de Gibbs pour μ_θ, μ_δ, μ_P et μ_D dans l'expérience de l'Exemple 10.4. (*Source :* Robert et Reber, 1998.)

Puisque nous souhaitons évaluer les effets de l'intoxication et des deux traitements, nous nous intéressons aux comparaisons de μ_δ, μ_D, μ_P et de $\mu_D - \mu_P$ à 0. Le Tableau 10.1 donne les probabilités a posteriori que les effets soient significatifs, c'est-à-dire qu'on ait $0 > \mu_\delta$, $\mu_D > 0$, etc., ainsi que les intervalles de confiance, les uns comme les autres étant des approximations obtenues par les échantillons de Gibbs de la Figure 10.3. Cela nous permet de conclure que les effets de l'intoxication, des médicaments et du placebo sont significatifs, bien qu'à un degré moindre pour le placebo. On peut également constater que l'effet du médicament est significativement différent de celui du placebo. ‖

Tab. 10.1. Probabilités a posteriori de fiabilité et intervalles de confiance à 95% pour les effets de moyennes.

	μ_δ	μ_D	μ_P	$\mu_D - \mu_P$
Probabilité	1.00	0.9998	0.94	0.985
Confiance	[-3.48,-2.17]	[0.94,2.50]	[-0.17,1.24]	[0.14,2.20]

10.2.5 Extensions hiérarchiques du modèle normal

Dans cette section, comme dans la Section 10.3, nous considérons le cas particulier de la loi normale,

$$x \sim \mathcal{N}_p(\theta, \Sigma)$$

parce qu'il donne des expressions partiellement exprimables analytiquement. Comme dans Lindley et Smith (1972), Smith (1973) et Berger (1985b), nous

faisons appel à une loi conjuguée de premier niveau $\theta \sim \mathcal{N}_p(\mu, \Sigma_\pi)$, pour une décomposition plus facile des estimateurs.

Lemme 10.17. *Dans le modèle normal conjugué, l'estimateur de Bayes hiérarchique est*

$$\delta^\pi(x) = \mathbb{E}^{\pi_2(\mu,\Sigma_\pi|x)}[\delta(x|\mu, \Sigma_\pi)],$$

avec

$$\begin{aligned}\delta(x|\mu, \Sigma_\pi) &= x - \Sigma W(x - \mu),\\ W &= (\Sigma + \Sigma_\pi)^{-1},\\ \pi_2(\mu, \Sigma_\pi|x) &\propto (\det W)^{1/2} \exp\{-(x - \mu)^t W(x - \mu)/2\}\pi_2(\mu, \Sigma_\pi).\end{aligned}$$

La preuve est une conséquence directe du Lemme 10.10 et du fait que la loi marginale $m_1(x|\mu, \Sigma_\pi)$ est normale $\mathcal{N}_p(\mu, W^{-1})$.

Exemple 10.18. (Suite de l'Exemple 10.15) Le choix d'une loi a priori constante sur β donne une expression analytique de $\delta^\pi(x)$. Il existe alors une fonction h_k (Exercice 10.19) telle que

$$\delta^\pi(x) = x - h_{p-k-2}(||x||_*^2)\Sigma C^{-1}(x - Px),$$

avec

$$\begin{aligned}P &= Y(Y^t C^{-1} Y)^{-1} Y^t C^{-1},\\ ||x||_*^2 &= xC^{-1}(I_p - P)x.\end{aligned}$$

Remarquons que Px est la projection orthogonale de x sur le sous-espace $H = \{\mu = Y\beta,\, \beta \in \mathbb{R}^k\}$ selon la métrique définie par C^{-1}. L'estimateur δ^π est donc une somme pondérée de x et de cette projection. Par conséquent, δ^π prend en compte l'information a priori de façon adaptative, en fonction de la distance $||x||_*$ de x à H. ‖

Exemple 10.19. On considère le modèle hiérarchique *échangeable* :

$$\begin{aligned}x|\theta &\sim \mathcal{N}_p(\theta, \sigma_1^2 I_p),\\ \theta|\xi &\sim \mathcal{N}_p(\xi\mathbf{1}, \sigma_\pi^2 I_p),\\ \xi &\sim \mathcal{N}(\xi_0, \tau^2),\end{aligned}$$

avec $\mathbf{1} = (1, \ldots, 1)^t \in \mathbb{R}^p$. Dans ce cas,

$$\delta(x|\xi, \sigma_\pi) = x - \frac{\sigma_1^2}{\sigma_1^2 + \sigma_\pi^2}(x - \xi\mathbf{1}),$$

$$\pi_2(\xi,\sigma_\pi^2|x) \propto (\sigma_1^2+\sigma_\pi^2)^{-p/2}\exp\{-\frac{\|x-\xi\mathbf{1}\|^2}{2(\sigma_1^2+\sigma_\pi^2)}\}e^{-(\xi-\xi_0)^2/2\tau^2}\pi_2(\sigma_\pi^2)$$
$$\propto \frac{\pi_2(\sigma_\pi^2)}{(\sigma_1^2+\sigma_\pi^2)^{p/2}}\exp\left\{-\frac{p(\bar{x}-\xi)^2}{2(\sigma_1^2+\sigma_\pi^2)}-\frac{s^2}{2(\sigma_1^2+\sigma_\pi^2)}-\frac{(\xi-\xi_0)^2}{2\tau^2}\right\}$$

avec $s^2=\sum_i(x_i-\bar{x})^2$. Alors, $\pi_2(\xi|\sigma_\pi^2,x)$ suit une loi normale $\mathcal{N}(\mu(x,\sigma_\pi^2), V_\pi(\sigma_\pi^2))$, où

$$\mu(x,\sigma_\pi^2)=\bar{x}-\frac{\sigma_1^2+\sigma_\pi^2}{\sigma_1^2+\sigma_\pi^2+p\tau^2}(\bar{x}-\xi_0),\qquad V_\pi(\sigma_\pi^2)=\frac{\tau^2(\sigma_1^2+\sigma_\pi^2)}{\sigma_1^2+\sigma_\pi^2+p\tau^2}.$$

Alors

$$\delta^\pi(x)=\mathbb{E}^{\pi_2(\sigma_\pi^2|x)}\left[x-\frac{\sigma_1^2}{\sigma_1^2+\sigma_\pi^2}(x-\bar{x}\mathbf{1})-\frac{\sigma_1^2+\sigma_\pi^2}{\sigma_1^2+\sigma_\pi^2+p\tau^2}(\bar{x}-\xi_0)\mathbf{1}\right]$$

et

$$\pi_2(\sigma_\pi^2|x)\propto\frac{\tau\exp-\frac{1}{2}\left[\frac{s^2}{\sigma_1^2+\sigma_\pi^2}+\frac{p(\bar{x}-\xi_0)^2}{p\tau^2+\sigma_1^2+\sigma_\pi^2}\right]}{(\sigma_1^2+\sigma_\pi^2)^{(p-1)/2}(\sigma_1^2+\sigma_\pi^2+p\tau^2)^{1/2}}\ \pi_2(\sigma_\pi^2). \tag{10.5}$$

Berger (1985b, p. 184-185) donne une démonstration détaillée de ce résultat, ainsi que l'expression correspondante de la variance a posteriori de θ.

Noter que l'estimateur bayésien hiérarchique a une forme particulière

$$\delta^\pi(x)=x-\mathbb{E}^{\pi_2(\sigma_\pi^2|x)}\left[\frac{\sigma_1^2}{\sigma_1^2+\sigma_\pi^2}\right](x-\bar{x}\mathbf{1})$$
$$-\mathbb{E}^{\pi_2(\sigma_\pi^2|x)}\left[\frac{\sigma_1^2+\sigma_\pi^2}{\sigma_1^2+\sigma_\pi^2+p\tau^2}\right](\bar{x}-\xi_0)\mathbf{1}. \tag{10.6}$$

Cela signifie que les deux niveaux hiérarchiques induisent deux types différents de rétrécissement pour l'estimateur de Bayes. L'hypothèse d'échangeabilité explique le second terme, $(x-\bar{x}\mathbf{1})$, qui réduit l'observation vers la moyenne commune $\bar{x}$; ce serait l'estimateur à utiliser dans le cas d'une relation *exacte* entre les paramètres du modèle. De même, le troisième terme découle de l'hypothèse que la moyenne commune varie autour de ξ_0.

Dans le cas où l'information concernant ξ_0 n'est pas fiable, une loi non informative peut être utilisée pour le deuxième niveau, soit, $\pi_2(\sigma_\pi^2)=1$ et $\tau^2=+\infty$. Alors, pour $p\geq 4$,

$$\delta^\pi(x)=x-\mathbb{E}^{\pi_2(\sigma_\pi^2|x)}\left[\frac{\sigma_1^2}{\sigma_1^2+\sigma_\pi^2}\right](x-\bar{x}\mathbf{1})$$
$$=x-h_{p-2}(\|x-\bar{x}\mathbf{1}\|^2)(x-\bar{x}\mathbf{1}) \tag{10.7}$$

et

$$\pi_2(\sigma_\pi^2|x) \propto (\sigma_1^2+\sigma_\pi^2)^{-(p-1)/2} \exp\left\{-\frac{s^2}{2(\sigma_1^2+\sigma_\pi^2)}\right\}, \tag{10.8}$$

la fonction h_k étant celle de l'Exemple 10.8 (voir aussi l'Exercice 10.19). On peut vérifier que (10.7) et (10.8) viennent de (10.5) et (10.6) lorsque τ^2 tend vers $+\infty$, et que (10.8) ne définit une loi propre que lorsque $p \geq 4$. L'utilité de l'hypothèse d'échangeabilité en dimension 3 est subordonnée à l'existence d'une information supplémentaire, à savoir une information a priori sur la position de la moyenne commune ξ. Cette contrainte recoupe des résultats fréquentistes sur la minimaxité de (10.7), qui n'est vraie que pour $p \geq 4$ (Brown, 1988).

Il faut noter que, si σ_1 est également inconnu, avec une loi a priori (éventuellement non informative) π_0, les égalités (10.6) et (10.7) sont toujours vraies, à condition que les espérances soient prises par rapport à la loi a posteriori $\pi(\sigma_1^2, \sigma_\pi^2|x)$. De façon analogue, si ξ est distribué selon une loi de Student à α degrés de liberté $\mathscr{T}(\alpha, \xi_0, \tau^2)$ plutôt que selon une loi normale, nous avons montré dans l'Exemple 10.3 que cette loi se décompose en un mélange de lois gaussiennes $\mathscr{N}(\xi_0, \tau^2/z)$ par une loi gamma $\mathscr{G}(\alpha/2, \alpha/2)$ sur z. Par conséquent, δ^π peut être obtenu à partir de (10.6) et (10.7) en intégrant par rapport à z. Voir Angers (1987, 1992) pour une étude plus détaillée sur la modélisation a priori par des lois de Student. ‖

Exemple 10.20. (Suite de l'Exemple 10.8) Dans le cadre du modèle de régression classique, une hypothèse d'échangeabilité sur les paramètres β_i $(1 \leq i \leq p)$ conduit à des estimateurs similaires à ceux que nous venons de voir. Lorsque

$$\beta_i \sim \mathscr{N}(\xi, \sigma_\pi^2) \qquad \text{et} \qquad \pi(\xi) = 1,$$

Lindley et Smith (1972), par une analyse similaire à l'Exemple 10.19, obtiennent l'estimateur

$$\delta^\pi(y) = \left\{I_p + \frac{\sigma^2}{\sigma_\pi^2}(X^tX)^{-1}(I_p - p^{-1}J_p)\right\}^{-1} \hat{\beta},$$

avec $\hat{\beta}$ estimateur des moindres carrés $\hat{\beta} = (X^tX)^{-1}X^ty$ et J_p matrice $(p \times p)$ ne contenant que des 1. L'analogie avec l'exemple ci-dessus est plus marquante si on écrit δ^π sous la forme

$$\delta^\pi(y) = \bar{\beta}\mathbf{1} + \left\{I_p + \frac{\sigma^2}{\sigma_\pi^2}(X^tX)^{-1}(I_p - p^{-1}J_p)\right\}^{-1} (\hat{\beta} - \bar{\beta}\mathbf{1})$$

(car $(I_p - p^{-1}J_p)\bar{\beta}\mathbf{1} = 0$) puisque l'estimateur de Bayes est rétréci vers la moyenne commune $\bar{\beta}$ (au sens matriciel). Remarquons qu'il s'exprime aussi

$$\delta^\pi(y) = \left\{X^tX + \frac{\sigma^2}{\sigma_\pi^2}(I_p - p^{-1}J_p)\right\}^{-1} X^ty.$$

On voit bien alors comment l'échangeabilité atténue les problèmes numériques et statistiques dus à la quasi-colinéarité des colonnes de X. En effet, la matrice

$$\frac{\sigma^2}{\sigma_\pi^2}(I_p - p^{-1}J_p)$$

joue un rôle de stabilisation pour cet estimateur. Si, dans l'a priori de second niveau, nous considérons plutôt $\xi = 0$, l'estimateur de Bayes est alors (Exercice 10.23)

$$\begin{aligned}\delta^\pi(y) &= \left\{I_p + \frac{\sigma^2}{\sigma_\pi^2}(X^tX)^{-1}\right\}^{-1}\hat{\beta} \\ &= \left(X^tX + \frac{\sigma^2}{\sigma_\pi^2}I_p\right)^{-1}X^ty.\end{aligned}$$

On appelle ces estimateurs *estimateurs ridges*. Ils ont été proposés par Hoerl et Kennard (1970) comme un remède aux problèmes de *multicolinéarité* dans la matrice X^tX, qui interviennent lorsque deux régresseurs (ou plus) sont presque colinéaires. Le facteur matriciel

$$\left[I_p + k(X^tX)^{-1}\right]^{-1}$$

stabilise l'estimateur des moindres carrés lorsque certaines valeurs propres de X^tX sont proches de 0 (voir également Lindley et Smith, 1972, et Goldstein et Smith, 1974). Ces estimateurs ont été ensuite généralisés en considérant un facteur matriciel de la forme

$$\left[I_p + h(y)(X^tX)^{-1}\right]^{-1},$$

qui peut correspondre au cas σ_π^2 inconnu, avec une distribution a priori $\pi_2(\sigma_\pi^2)$, puisque l'estimateur de Bayes est alors

$$\delta^\pi(y) = \mathbb{E}^{\pi_2(\sigma_\pi^2|y)}\left[I_p + \frac{\sigma^2}{\sigma_\pi^2}(X^tX)^{-1}\right]^{-1}\hat{\beta}.$$

Le point de vue classique donne l'impression que les impératifs de réduction de la multicolinéarité et de minimaxité sont contradictoires, puisque Casella (1980, 1985a) montre que des conditions nécessaires de minimaxité pour les estimateurs ridges ne sont pas compatibles avec l'influence stabilisatrice de ces estimateurs. Robert (1998) observe le même phénomène pour d'autres classes d'estimateurs à rétrécisseur et montre que cet antagonisme est dû à la *monodimensionnalité* du problème de multicolinéarité, ce qui explique pourquoi une amélioration uniforme de $\hat{\beta}$ est impossible, du fait de l'effet Stein. ‖

10.3 Optimalité des estimateurs bayésiens hiérarchiques

D'une[74] façon générale, les estimateurs bayésiens hiérarchiques étant similaires aux estimateurs de Bayes habituels, ils ne sont ni plus ni moins admissibles que les estimateurs de Bayes décrits dans les chapitres précédents. Par exemple, les conditions nécessaires et suffisantes du Chapitre 8 restent vraies pour les estimateurs bayésiens hiérarchiques. De même, les propriétés liées à l'invariance dans le Chapitre 9 ne sont en aucun cas liées à la structure, hiérarchique ou non, des lois à priori.

Mais nous verrons également dans un cas particulier qu'il est effectivement possible de tirer parti de la spécificité des estimateurs bayésiens hiérarchiques pour obtenir une condition générale de minimaxité, en utilisant les lois a priori de second niveau. De tels résultats montrent ce que l'approche bayésienne hiérarchique permet de gagner en robustesse, en incluant l'information a priori la plus subjective aux niveaux les plus élevés. On conçoit alors cette démarche comme un compromis entre une analyse bayésienne directe et une réponse aux exigences fréquentistes.

Considérons de nouveau le modèle normal, $x \sim \mathcal{N}_p(\theta, \Sigma)$ avec Σ connu. Comme dans la Section 10.2.5, la loi a priori de premier niveau sur θ est conjuguée, $\theta \sim \mathcal{N}_p(\mu, \Sigma_\pi)$. La loi a priori π_2 sur les hyperparamètres μ, Σ_π se décompose ainsi :

$$\pi_2(\mu, \Sigma_\pi) = \pi_2^1(\Sigma_\pi|\mu)\pi_2^2(\mu).$$

Dans ce cas,

$$m(x) = \int_{\mathbb{R}^p} m(x|\mu)\pi_2^2(\mu)\, d\mu,$$

avec

$$m(x|\mu) = \int f(x|\theta)\pi_1(\theta|\mu, \Sigma_\pi)\pi_2^1(\Sigma_\pi|\mu)\, d\theta\, d\Sigma_\pi.$$

En outre, l'estimateur de Bayes

$$\delta^\pi(x) = x + \Sigma\nabla \log m(x) \tag{10.9}$$

peut s'écrire

$$\delta^\pi(x) = \int \delta(x|\mu)\pi_2^2(\mu|x)\, d\mu,$$

où

$$\delta(x|\mu) = x + \Sigma\nabla \log m(x|\mu),$$
$$\pi_2^2(\mu|x) = \frac{m(x|\mu)\pi_2^2(\mu)}{m(x)}.$$

[74] Cette section peut être omise en première lecture puisqu'elle ne traite que de la minimaxité d'une classe particulière d'estimateurs bayésiens hiérarchiques dans le cas gaussien. Son but est d'illustrer le gain en robustesse obtenu grâce à la modélisation hiérarchique.

Ces décompositions conditionnelles seront utiles ci-dessous.

Soit Q matrice $(p \times p)$ symétrique définie positive associée au coût quadratique

$$\mathrm{L}_Q(\theta, \delta) = (\theta - \delta)^t Q(\theta - \delta). \tag{10.10}$$

Un estimateur δ est minimax pour le coût (10.10) s'il satisfait

$$R(\theta, \delta) = \mathbb{E}_\theta[\mathrm{L}_Q(\theta, \delta(x))] \leq \mathrm{tr}(\Sigma Q),$$

puisque $\mathrm{tr}(\Sigma Q)$ est le risque minimax de $\delta_0(x) = x$. La *méthode de l'estimateur sans biais du risque* a été suggérée par Stein (1973, 1981) pour déterminer des conditions suffisantes de minimaxité. (Voir Brown, 1988, et Rukhin, 1995, pour des analyses détaillées de cette méthode.) Il s'agit d'obtenir un opérateur différentiel $\mathscr{D}$, indépendant de θ, tel que

$$R(\theta, \delta) = \mathbb{E}_\theta[\mathscr{D}\delta(x)],$$

pour tout paramètre θ et tout estimateur δ. Cette technique donne effectivement une condition suffisante de minimaxité sous la forme $\mathscr{D}\delta(x) \leq \mathrm{tr}(Q\Sigma)$ (Exercice 2.56). Dans le cas particulier (10.9), l'opérateur différentiel est obtenu grâce au résultat suivant (Berger et Robert, 1990).

Lemme 10.21. *Si $m(x)$ vérifie les trois conditions*

$$(1)\ \mathbb{E}_\theta \|\nabla \log m(x)\|^2 < +\infty; \qquad (2)\ \mathbb{E}_\theta \left| \frac{\partial^2 m(x)}{\partial x_i \partial x_j} \Big/ m(x) \right| < +\infty;$$

et $(1 \leq i \leq p)$

$$(3)\ \lim_{|x_i| \to +\infty} \left| \nabla \log m(x) \right| \exp\{-(1/2)(x-\theta)^t \Sigma^{-1}(x-\theta)\} = 0,$$

l'estimateur sans biais du risque de δ^π s'écrit

$$\mathscr{D}\delta^\pi(x) = \mathrm{tr}(Q\Sigma) + \frac{2}{m(x)} \mathrm{tr}(H_m(x)\tilde{Q}) - (\nabla \log m(x))^t \tilde{Q} (\nabla \log m(x)),$$

avec

$$\tilde{Q} = \Sigma Q \Sigma, \qquad H_m(x) = \left(\frac{\partial^2 m(x)}{\partial x_i \partial x_j} \right).$$

Cet estimateur sans biais du risque conduit alors à une condition suffisante de minimaxité :

$$\frac{2}{m(x)} \mathrm{tr}(H_m(x)\tilde{Q}) - (\nabla \log m(x))^t \tilde{Q} (\nabla \log m(x)) \leq 0.$$

On note div l'opérateur *divergence*, c'est-à-dire

$$\mathrm{div} f(x) = \sum_{i=1}^n \frac{\partial f_i}{\partial x_i}(x),$$

pour toute fonction différentiable f de $\mathbb{R}^n$ dans $\mathbb{R}^n$.

Corollaire 10.22. *Si m satisfait les conditions du Lemme* 10.21 *et si*

$$\operatorname{div}\left(\tilde{Q}\nabla\sqrt{m(x)}\right) \leq 0, \tag{10.11}$$

δ^π est minimax.

Preuve. Il suffit de considérer le développement de $\operatorname{div}(\tilde{Q}\nabla\sqrt{m(x)})$ pour obtenir

$$\begin{aligned}\operatorname{div}(\tilde{Q}\nabla\sqrt{m(x)}) &= \frac{1}{2}\operatorname{div}\left(\tilde{Q}\frac{\nabla m(x)}{\sqrt{m(x)}}\right)\\ &= \frac{1}{2\sqrt{m(x)}}\operatorname{div}(\tilde{Q}\nabla m(x)) - \frac{1}{4}\left(\frac{\nabla m(x)}{m(x)\sqrt{m(x)}}\right)^t\tilde{Q}\nabla m(x)\\ &= \frac{\sqrt{m(x)}}{4}\left[\frac{2}{m(x)}\operatorname{tr}(H_m(x)\tilde{Q}) - \nabla\log m(x)^t\tilde{Q}\nabla\log m(x)\right]\end{aligned}$$

et calculer le terme additionnel en $\mathscr{D}\delta^\pi(x)$. □

Dans le cas particulier où $\Sigma = Q = I_p$, la condition du Corollaire 10.22 peut se simplifier en une condition sur le *laplacien* de $m(x)^{1/2}$:

$$\Delta\sqrt{m(x)} = \sum_{i=1}^{n}\frac{\partial^2}{\partial x_i^2}(\sqrt{m(x)}) \leq 0$$

(on dit alors que $\sqrt{m(x)}$ est *surharmonique*). Comme la vérification pratique de cette inégalité est souvent difficile, on utilise parfois une condition de minimaxité plus explicite, découlant du Corollaire 10.22 en conditionnant par rapport à μ.

Lemme 10.23. *L'estimateur δ^π est minimax si*

$$\operatorname{div}\left(\tilde{Q}\nabla m(x|\mu)\right) \leq 0. \tag{10.12}$$

Preuve. On a

$$\operatorname{div}(\tilde{Q}\nabla m(x)) = \int \operatorname{div}\left(\tilde{Q}\nabla m(x|\mu)\right)\pi_2^2(\mu)\,d\mu$$

et (10.12) entraîne (10.11). □

Par conséquent, si $\tilde{Q} = I_p$ et si $m(x|\mu)$ est surharmonique, l'estimateur bayésien hiérarchique correspondant est minimax. Ce résultat peut sembler évident dans sa formulation et sa démonstration mais il est en réalité assez général. Il donne en effet *une condition nécessaire et suffisante de minimaxité qui ne dépend pas de $\pi_2^2(\mu)$* et autorise donc une modélisation quelconque

sur l'hyperparamètre μ. D'un point de vue subjectif, il semble beaucoup plus important d'avoir une liberté totale pour la loi a priori de μ que pour le choix complémentaire sur Σ_π, puisqu'il est souvent plus facile d'obtenir de l'information sur μ que sur Σ_π. L'exemple suivant montre de plus que la condition (10.12) est vérifiée par un ensemble important de lois π_2^1.

Exemple 10.24. (Suite de l'Exemple 10.15) On considère de nouveau le cas où $\Sigma_\pi = \xi C - \Sigma$ et $Q = \Sigma^{-1} C \Sigma^{-1}$ (alors $\tilde{Q} = C$). Il vient du Lemme 10.12 que

$$m(x|\mu) \propto \int_0^\infty \xi^{-p/2} \exp\left\{ -\frac{(x-\mu)^t C^{-1}(x-\mu)}{2\xi} \right\} \pi_2^1(\xi|\mu)\, d\xi.$$

Donc

$$\operatorname{div}\left(\tilde{Q}\nabla m(x|\mu)\right) \propto \int_0^\infty \left(-\frac{p}{\xi} + \frac{(x-\mu)^t C^{-1}(x-\mu)}{\xi^2} \right) \times e^{-(x-\mu)^t C^{-1}(x-\mu)/2\xi} \xi^{-p/2} \pi_2^1(\xi|\mu)\, d\xi\,,$$

et (10.12) est équivalent à

$$\psi(a) = \int_0^\infty (2a - p\xi) \xi^{-(p+4)/2} e^{-a/\xi} \pi_2^1(\xi|\mu)\, d\xi \le 0, \qquad \forall a \ge 0.$$

Si π_2^1 est presque partout différentiable, on obtient, par une intégration par parties,

$$\psi(a) = -2e^{-a/\xi_0} \xi_0^{-p/2} \pi_2'(\xi_0|\mu) - \int_{\xi_0}^{+\infty} \xi^{-p/2} e^{-a/\xi} \pi_2^1(\xi|\mu)\, d\xi,$$

avec $\xi_0 = \inf(\operatorname{supp}(\pi_2^1))$ et π_2' dérivée de π_2^1. Cette expression implique que :

Proposition 10.25. *Si $\pi_2^1(\xi|\mu)$ est croissante quel que soit $\mu \in \mathbb{R}^p$, δ^π est minimax pour toute loi a priori π_2^2.*

Par conséquent, si $\pi_2^1(\xi|\mu) = 1$ pour $\xi_0 \le \xi$ avec $\lambda_{\max}(C^{-1}\Sigma) \le \xi_0$, l'estimateur de Bayes correspondant est minimax. ‖

Cet exemple peut être étendu au cas où $\theta \sim \mathcal{N}_p(\mu, \sigma_\pi^2 \Sigma)$ et où $\pi_2^1(\sigma_\pi^2|\mu)$ est strictement croissante ($C = \Sigma$ et $\xi = \sigma_\pi^2 - 1$). Cette classe n'inclut évidemment pas tous les estimateurs hiérarchiques et tous les estimateurs minimax, mais elle contient tout de même tous les estimateurs minimax proposés par Strawderman (1971) et Berger (1975a, 1980b), certains étant de plus admissibles (voir également Kubokawa, 1991, et l'Exercice 10.36).

Remarquons que les lois a priori suggérées par la Proposition 10.25 sont contre-intuitives : il semble en effet difficile, des points de vue subjectif et

non informatif, de trouver des avantages à une loi croissante en terme de variance. Les lois a priori sont au contraire souvent décroissantes pour de grandes valeurs de σ_π^2. C'est par exemple le cas de la loi non informative de Jeffreys $\pi(\sigma_\pi^2) = 1/\sigma_\pi^2$. Ce résultat est donc une indication implicite de l'*aspect artificiel de la notion de minimaxité* : donner le même poids *a posteriori* à toutes les valeurs possibles du paramètre revient à favoriser *a priori* les plus invraisemblables (ou les *moins favorables*).

L'exemple ci-dessous illustre les points forts de l'approche bayésienne hiérarchique d'un point de vue minimax, même lorsque la loi de premier niveau est plus rudimentaire. Il présente également une propriété de robustesse de minimaxité, au sens où la minimaxité ne dépend pas tant de la normalité de la loi a priori que de sa symétrie sphérique. Ce résultat est donc un équivalent bayésien aux résultats fréquentistes de Cellier *et al.* (1989).

Exemple 10.26. Soit $x \sim \mathcal{N}_p(\theta, I_p)$. La moyenne θ est estimée sous coût quadratique. Au lieu d'utiliser une loi de premier niveau conjuguée, on choisit la loi uniforme sur la sphère de rayon c,

$$\pi_1(\theta|c) \propto \mathbb{I}_{\{||\theta||^2=c\}},$$

formulant ainsi seulement une hypothèse de symétrie sphérique pour la loi a priori globale. La loi a priori de second niveau $\pi_2(c)$ est une loi gamma $\mathcal{G}(\alpha, \beta)$. L'estimateur de Bayes est alors (voir Robert *et al.*, 1990)

$$\delta^\pi(x) = \frac{2\alpha}{p}\frac{1}{1+2\beta}\,\frac{{}_1F_1(\alpha+1;(p+2)/2;||x||^2/(2+4\beta))}{{}_1F_1(\alpha;p/2;||x||^2/(2+4\beta))}\,x,$$

où on note ${}_1F_1$ la fonction confluente hypergéométrique. Avec $\alpha < 1$ et $\beta = 0$, on obtient

$$\delta^\pi(x) = \frac{2\alpha}{p}\,\frac{{}_1F_1(\alpha+1;(p+2)/2;||x||^2/2)}{{}_1F_1(\alpha;p/2;||x||^2/2)}\,x,$$

qui est un estimateur minimax et admissible (voir Alam, 1973). ‖

10.4 L'alternative bayésienne empirique

La méthodologie que nous étudions à présent jusqu'à la fin de ce chapitre ne découle pas des principes bayésiens[75], puisqu'elle consiste à approcher la loi a priori par des méthodes *fréquentistes* lorsque l'information a priori est trop limitée. Nous l'incluons tout de même dans ce livre pour plusieurs raisons :

[75] Le nom de *bayésien empirique* est doublement trompeur puisque, d'une part, la méthode n'est pas bayésienne et, d'autre part, les véritables méthodes bayésiennes sont tout autant empiriques puisque liées aux données ! À moins, bien sûr, de prendre *empirique* dans son sens péjoratif...

(i) elle peut être considérée comme une méthode duale de l'analyse bayésienne hiérarchique présentée ci-dessus ;
(ii) elle est *asymptotiquement* équivalente à l'approche bayésienne ;
(iii) elle est souvent étiquetée "bayésienne" par les fréquentistes et les praticiens ; et
(iv) elle constitue dans certains cas une approximation acceptable lorsque la modélisation bayésienne réelle est trop compliquée ou trop chère.

Nous verrons que l'analyse bayésienne empirique se situe entre les approches classique et bayésienne et nous montrerons que l'alternative hiérarchique est souvent préférable. Cette section n'est qu'une courte introduction à l'approche bayésienne empirique. Les lecteurs intéressés pourront se reporter à Morris (1983b), Berger (1985b), Maritz et Lwin (1989) ou bien Carlin et Louis (2000a,b) pour des études plus complètes sur le sujet.

10.4.1 Le principe bayésien empirique non paramétrique

Robbins (1951, 1955, 1964, 1983) décrit le *point de vue bayésien empirique* de la façon suivante. Soient $(n+1)$ observations indépendantes $x_1, \ldots, x_{n+1}$ de densités $f(x_i|\theta_i)$; le problème porte sur l'inférence sur θ_{n+1}, avec l'hypothèse supplémentaire que les θ_i ont tous été tirés selon le même a priori *inconnu* g. D'un point de vue bayésien, cela revient à dire que la loi d'échantillonnage est connue, mais que la loi a priori ne l'est pas. La loi marginale,

$$f_g(x) = \int f(x|\theta)g(\theta)\,d\theta, \tag{10.13}$$

peut alors être utilisée pour retrouver la distribution g à partir des observations, puisque $x_1, \ldots, x_n$ peut être vu comme un échantillon i.i.d. de loi f_g. En résolvant ce problème inverse, on obtient ainsi une approximation $\hat{g}_n$ qu'on peut substituer à la vraie loi a priori pour obtenir l'expression suivante de la loi a posteriori

$$\tilde{\pi}(\theta_{n+1}|x_{n+1}) \propto f(x_{n+1}|\theta_{n+1})\hat{g}_n(\theta_{n+1}). \tag{10.14}$$

Naturellement, cette technique n'est pas bayésienne, bien qu'elle repose sur la formule de Bayes (10.14) et rejoigne parfois la modélisation classique, puisqu'elle *utilise les données deux fois.* Confronté à la méconnaissance de g, le réflexe bayésien serait d'indexer cette loi par un hyperparamètre λ et de modéliser cette ignorance par une loi a priori de second niveau $\pi_2(\lambda)$[76]. Deely et Lindley (1981) comparent les deux approches dans le cas d'une loi de Poisson.

Le cadre initial de Robbins (1955) est principalement *non paramétrique* et fait usage des observations $x_1, \ldots, x_{n+1}$ pour estimer f_g. (Dans le cas général,

[76] L'indexation par λ n'est formellement pas restrictive, comme le montre l'Exercice 1.2.

la densité marginale f_g peut être estimée par une méthode à noyau ; voir par exemple Devroye et Györfi, 1985.)

Exemple 10.27. On considère les x_i distribués selon une loi $\mathscr{P}(\theta_i)$ $(i = 1, \ldots, n)$. Si $p_k(x_1, \ldots, x_n)$ est le nombre d'observations égales à k, $k \in \mathbb{N}$, $p_k(x_1, \ldots, x_n)$ donne une estimation de la loi marginale,

$$f_g(k) = \int_0^{+\infty} e^{-\theta} \frac{\theta^k}{k!} g(\theta)\, d\theta.$$

Si $x_{n+1} \sim \mathscr{P}(\theta_{n+1})$ et si θ_{n+1} est estimé sous coût quadratique, l'estimateur de Bayes est

$$\begin{aligned}\delta^g(x_{n+1}) = \mathbb{E}^g[\theta|x_{n+1}] &= \frac{\int_0^{+\infty} e^{-\theta}\theta^{x_{n+1}+1} g(\theta)\, d\theta}{\int_0^{+\infty} e^{-\theta}\theta^{x_{n+1}} g(\theta)\, d\theta} \\ &= \frac{f_g(x_{n+1}+1)}{f_g(x_{n+1})}(x_{n+1}+1).\end{aligned}$$

Donc l'approximation bayésienne empirique de δ^g est

$$\delta^{\mathrm{EB}}(x_{n+1}) = \frac{p_{x_{n+1}+1}(x_1, \ldots, x_n)}{p_{x_{n+1}}(x_1, \ldots, x_n) + 1}(x_{n+1}+1), \tag{10.15}$$

où on a remplacé f_g par son approximation. ‖

Cette méthode souffre de plusieurs inconvénients :

(a) L'utilisation d'estimations non paramétriques, par exemple pour la densité a priori, comme préliminaire à une procédure d'estimation paramétrique semble sous-optimale, car il est toujours plus difficile d'évaluer les erreurs commises dans une étape non paramétrique. Ainsi, dans l'exemple ci-dessus, si le numérateur de (10.15) est nul, l'estimateur est nul.

(b) Plus généralement, les estimations non paramétriques de la densité de mélange g dans (10.13) par des techniques de maximum de vraisemblance sont souvent basiques, puisqu'elles correspondent à des lois à support fini. (Voir Bohning, 1999, ou Carlin et Louis, 2000a[77]). De telles lois a priori sont rarement acceptables du point de vue bayésien.

(c) Il est assez rare d'avoir des relations fonctionnelles entre la moyenne (ou toute autre fonction d'intérêt) et la loi marginale, comme dans l'Exemple 10.27. Lorsqu'une telle relation n'existe pas, le calcul de l'estimateur de g est généralement trop compliqué pour garantir que les estimateurs en résultant soient de bonnes approximations des vrais estimateurs de Bayes.

[77] Cela est également vrai pour des approches a priori utilisant des noyaux ou des processus de Dirichlet.

(d) L'approximation n'est effectivement justifiée que pour des échantillons de taille importante, c'est-à-dire lorsque l'estimateur de la loi marginale $\hat{f}_g^n$ est acceptable. Dans le cas contraire, comme le montre l'Exemple 10.27, $\hat{f}_g^n$ subit des variations trop importantes et doit être lissé pour être d'une quelconque utilité (voir Maritz et Lwin, 1989).

(e) L'hypothèse selon laquelle on dispose de nombreux problèmes identiques et indépendants concernant la même loi a priori est forte et peut ne pas être vérifiée en pratique. Ainsi, un échantillon unique, même très grand, ne permet pas d'estimer f_g, car il ne correspond qu'à une seule observation de θ. Cette critique reste valable avec l'approche paramétrique (voir, notamment, la Proposition 10.31).

Pour toutes ces raisons, nous ne pousserons pas plus loin l'étude de l'analyse bayésienne empirique non paramétrique ; nous allons à présent nous restreindre au *principe bayésien empirique paramétrique* de Morris (1983b).

10.4.2 Principe bayésien empirique paramétrique

Un intérêt pratique des techniques bayésiennes empiriques est de déterminer des approximations dans des contextes non informatifs. Nous avons montré dans les chapitres précédents que l'approche bayésienne était un outil efficace pour obtenir des procédures optimales d'un point de vue fréquentiste, sous la forme d'un cadre unifié d'inférence statistique. L'analyse bayésienne empirique peut alors être vue comme une approximation pratique de cet outil.

Pour les familles exponentielles, lorsque la loi a priori n'est pas disponible, le plus simple est de considérer l'a priori conjugué associé à $f(x|\theta)$, $\pi(\theta|\lambda)$. Tandis que l'approche hiérarchique introduit une loi supplémentaire sur les hyperparamètres λ, l'analyse bayésienne empirique propose d'estimer ces hyperparamètres à partir de la loi marginale

$$m(x|\lambda) = \int_\Theta f(x|\theta)\pi(\theta|\lambda)\, d\theta$$

pour obtenir $\hat{\lambda}(x)$ et d'utiliser $\pi(\theta|\hat{\lambda}(x), x)$ en tant que pseudo-a posteriori. Cette méthode est donc une version paramétrique de l'idée originale de Robbins (1955).

Un inconvénient de l'analyse bayésienne empirique est qu'elle repose sur des méthodes fréquentistes pour l'estimation des hyperparamètres de $m(x|\lambda)$, alors qu'on pourrait tout aussi bien employer des techniques bayésiennes, comme le montre la Note 10.7.2. Une conséquence de ce choix est qu'un grand nombre de méthodes est utilisable : par exemple, l'estimateur de λ peut être choisi par la méthode des moments ou par la méthode du maximum de vraisemblance. Il en résulte un aspect arbitraire de l'analyse bayésienne empirique, qui est le défaut principal de la démarche, puisqu'il exclut l'emploi de la Théorie de la Décision. L'analyse bayésienne empirique est alors souvent

employée comme un outil pour justifier a posteriori des estimateurs *déjà existants*, comme nous le verrons à la Section 10.5. L'approche la plus répandue est d'utiliser des estimateurs du maximum de vraisemblance pour des raisons à la fois pratiques et théoriques, en particulier à cause de la ressemblance entre l'estimation par maximum de vraisemblance et le paradigme bayésien. Une justification supplémentaire de ce choix est donnée ci-dessous dans le cas particulier de l'estimation d'un paramètre naturel d'une famille exponentielle sous coût quadratique.

Lemme 10.28. *On considère*

$$x \sim f(x|\theta) = e^{\theta \cdot x - \psi(\theta)} h(x), \qquad x \in \mathbb{R}^k.$$

Si θ est distribué selon $\pi(\theta|\lambda)$, $\lambda \in \mathbb{R}^p$, et $\hat{\lambda}(x)$ est la solution des équations de vraisemblance associées à $m(x|\lambda)$, l'estimateur de Bayes empirique de θ vérifie

$$\begin{aligned}\delta^{\mathrm{EB}}(x) &= (\nabla \log m(x|\lambda))\,\big|_{\lambda=\hat{\lambda}(x)} - \nabla \log h(x)\\ &= \nabla[\log m(x|\hat{\lambda}(x))] - \nabla \log h(x).\end{aligned}$$

Preuve. On a

$$\nabla \log m(x|\hat{\lambda}(x)) = (\nabla \log m(x|\lambda))\,\big|_{\lambda=\hat{\lambda}(x)} + \nabla_x \hat{\lambda}(x) \nabla_\lambda m(x|\lambda)\big|_{\lambda=\hat{\lambda}(x),}$$

où $\nabla_\lambda m(x|\lambda)$ est le vecteur de composantes

$$\frac{\partial m(x|\lambda)}{\partial \lambda_i} \qquad (1 \le i \le p)\,,$$

et $\nabla_x \hat{\lambda}(x)$ est la matrice $(k \times p)$ de composantes

$$\frac{\partial \hat{\lambda}_i(x)}{\partial x_j} \qquad (1 \le i \le p, 1 \le j \le k)\,.$$

Par définition de $\hat{\lambda}(x)$, le second terme est nul. □

Par conséquent, un calcul bayésien habituel à partir de la loi a posteriori approchée $\pi(\theta|\hat{\lambda}(x))$ conduit au même résultat que l'approche bayésienne empirique, où on remplace λ par $\hat{\lambda}(x)$. Mais cette justification n'est manifestement pas d'une grande généralité puisqu'elle n'est vraie que pour la moyenne a posteriori du paramètre naturel d'une famille exponentielle.

Exemple 10.29. (Suite de l'Exemple 10.27) On suppose que $\pi(\theta|\lambda)$ est une loi exponentielle $\mathscr{E}xp(\lambda)$. Alors

$$\begin{aligned}m(x_i|\lambda) &= \int_0^{+\infty} e^{-\theta} \frac{\theta^{x_i}}{x_i!} \lambda e^{-\theta\lambda} d\theta\\ &= \frac{\lambda}{(\lambda+1)^{x_i+1}} = \left(\frac{1}{\lambda+1}\right)^{x_i} \frac{\lambda}{\lambda+1},\end{aligned}$$

et $x_i|\lambda \sim \mathscr{G}eo(\lambda/\lambda+1)$. L'estimateur du maximum de vraisemblance de λ est $\hat{\lambda}(x) = 1/\bar{x}$ et l'estimateur de Bayes empirique de θ_{n+1} est

$$\delta^{\mathrm{EB}}(x_{n+1}) = \frac{x_{n+1}+1}{\hat{\lambda}+1} = \frac{\bar{x}}{\bar{x}+1}(x_{n+1}+1),$$

la moyenne $\bar{x}$ étant établie sur les n premières observations. ‖

Exemple 10.30. Soient $x_1, \ldots, x_n$, n observations indépendantes de $\mathscr{B}(m, p_i)$. Casella (1985b) (voir aussi Morisson, 1979) utilise ce modèle pour représenter la décision d'acheter une nouvelle voiture dans l'année à venir. On suppose que les paramètres p_i $(1 \leq i \leq n)$ sont distribués selon la même loi a priori conjuguée

$$p_i \sim \mathscr{B}(\alpha, \beta).$$

L'estimateur de Bayes correspondant de p_i est

$$\delta_i^{\pi}(x_i) = \frac{\alpha+\beta}{\alpha+\beta+1}\frac{\alpha}{\alpha+\beta} + \left(1 - \frac{\alpha+\beta}{\alpha+\beta+1}\right)\frac{x_i}{m}$$

et la loi marginale de x_i est appelée *bêta-binomiale*,

$$P(x_i = k|\alpha, \beta) = \frac{B(k+\alpha, m-k+\beta)}{B(\alpha, \beta)}.$$

comme dans l'Exemple 10.14. Kendall et Stuart (1979) montrent que, pour cette loi marginale,

$$\mathbb{E}(x_i|m) = \frac{\alpha}{\alpha+\beta}, \qquad \mathrm{var}(x_i|m) = \frac{1}{m}\frac{\alpha\beta}{(\alpha+\beta)^2}\frac{\alpha+\beta+m}{\alpha+\beta+1}.$$

Lorsque α et β sont estimés par la méthode des moments, l'estimateur bayésien empirique de p_i est

$$\gamma_i^{\mathrm{EB}}(x_1, \ldots, x_n) = \frac{\hat{\alpha} + (x_i/m)}{\hat{\alpha}+\hat{\beta}+1}.$$

(L'Exercice 10.29 porte sur les données utilisées par Morisson, 1979.) ‖

La Section 10.5 présente les parallèles remarquables existant entre les manifestations de l'effet de Stein et l'approche bayésienne empirique et déduit grâce à cette dernière de bons estimateurs pour l'estimation ponctuelle, ainsi que pour les tests et régions de confiance. Le résultat qui suit explique, au contraire, pourquoi les tests de Bayes empiriques sont d'une utilité limitée pour un unique échantillon.

Proposition 10.31. *On considère le test de* $H_0 : \theta = \theta_0$ *contre* $H_1 : \theta = \theta_1$ *à partir d'un échantillon* $x_1, \ldots, x_n$, *i.i.d.* $f(x|\theta)$. *Une approche bayésienne empirique donne la procédure de test de rapport de vraisemblances*

$$\varphi(x) = \begin{cases} 1 & \text{si } \prod_{i=1}^n f(x_i|\theta_0) > \prod_{i=1}^n f(x_i|\theta_1)\,, \\ 0 & \text{sinon,} \end{cases} \tag{10.16}$$

quel que soit le niveau de confiance.

Preuve. Dans ce cadre, l'ensemble des paramètres inconnus est réduit à π_0, la probabilité a priori de H_0. La loi marginale de x est alors

$$m(x|\pi_0) = \pi_0 \prod_{i=1}^n f(x_i|\theta_0) + (1-\pi_0) \prod_{i=1}^n f(x_i|\theta_1)$$

et correspond à l'estimateur du maximum de vraisemblance de π_0 suivant :

$$\hat{\pi}_0(x_1, \ldots, x_n) = \begin{cases} 1 & \text{si } \prod_{i=1}^n f(x_i|\theta_0) > \prod_{i=1}^n f(x_i|\theta_1)\,, \\ 0 & \text{sinon.} \end{cases}$$

La réponse bayésienne étant

$$\varphi^\pi(x_1, \ldots, x_n) = \begin{cases} 1 & \text{si } P(\theta = \theta_0|x_1, \ldots, x_n, \pi_0) > \alpha\,, \\ 0 & \text{sinon,} \end{cases}$$

la probabilité a posteriori de H_0 est

$$P(\theta = \theta_0|x_1, \ldots, x_n, \hat{\pi}_0) = \frac{\hat{\pi}_0 \prod_{i=1}^n f(x_i|\theta_0)}{\hat{\pi}_0 \prod_{i=1}^n f(x_i|\theta_0) + (1-\hat{\pi}_0) \prod_{i=1}^n f(x_i|\theta_1)}$$

d'où (10.16). □

Lorsque l'on considère plusieurs problèmes de test simultanément, ce comportement extrême des tests de Bayes empiriques disparaît (Maritz et Lwin, 1989). Cependant, il est plutôt rare d'avoir à tester *simultanément* plusieurs hypothèses sur des paramètres de la *même loi* et l'intérêt pratique de l'approche bayésienne empirique pour les tests s'en trouve d'autant limité. On considère l'estimation des régions de confiance dans la Section 10.5, en relation avec l'effet Stein. Pour des études alternatives, voir Laird et Louis (1987) ou Carlin et Gelfand (1991).

Présentons en guise de conclusion une légère modification de l'approche bayésienne empirique consistant à utiliser des *mélanges* de lois conjuguées, puisqu'ils constituent également une famille conjuguée (voir le Lemme 3.23). Si $x_i \sim f(x_i|\theta_i)$ et

$$\theta_i \sim \sum_{j=1}^n p_j \pi(\theta_i|\lambda_j),$$

la loi marginale de x_i est

$$x_i|p, \lambda \sim \sum_{j=1}^{n} p_j \int_{\Theta} f(x_i|\theta)\pi(\theta|\lambda_j)\, d\theta.$$

(Voir la Section 6.4 et la Note 6.6.6 pour de plus amples détails sur l'analyse bayésienne de ce problème.) Maritz et Lwin (1989) étudient plus particulièrement l'application à l'analyse bayésienne empirique. Un inconvénient de cette extension est bien sûr qu'elle nécessite un plus grand nombre d'hyperparamètres et donc un plus grand nombre d'échantillons indépendants, tout en présentant toujours certains des défauts énumérés ci-dessus.

Insistons de nouveau sur le fait que la légitimité des méthodes bayésiennes empiriques n'est qu'*asymptotique* (Deely et Lindley, 1981). Leur popularité est liée aux bonnes propriétés fréquentistes de certains estimateurs ainsi obtenus et aux simplifications importantes qu'elles apportent dans la résolution de problèmes complexes, en comparaison de l'analyse bayésienne hiérarchique. (Consulter, par exemple, Carter et Rolph, 1974, ou Hui et Berger, 1983.) Pour des problèmes portant sur des échantillons de tailles finies, les méthodes bayésiennes empiriques ne sont que des approximations des méthodes bayésiennes exactes et ne peuvent donc se prévaloir de la même cohérence. En particulier, il est impossible de mener une inférence bayésienne complète en utilisant $\pi(\theta|x, \lambda(x))$, car ce n'est pas une loi a posteriori. Enfin, l'accroissement permanent de la puissance et des méthodes de calcul disponibles (Chapitre 6) réduit le besoin en approximations empiriques d'analyses hiérarchiques plus complexes. (Voir Berger, 1985b, Berger et Berliner, 1986, et Berger et Robert, 1990.)

10.5 Justifications bayésiennes empiriques de l'effet Stein

L'analyse bayésienne empirique de l'effet Stein, décrit en Note 2.8.2, fournit un cadre d'unification des différentes apparitions de ce paradoxe, selon lequel l'estimation jointe de paramètres indépendants peut être améliorable globalement en termes de qualité de l'estimation, sans qu'aucune composante ne puisse être améliorée uniformément. En outre, cette analyse explique la forme originelle des estimateurs de James-Stein et montre qu'ils correspondent à l'information a priori vague que θ est proche de 0.

10.5.1 Estimation ponctuelle

Nous commençons par un exemple qui illustre naturellement le fondement bayésien empirique de l'effet Stein.

Exemple 10.32. Soient $x \sim \mathcal{N}_p(\theta, I_p)$ et $\theta_i \sim \mathcal{N}(0, \tau^2)$. La loi marginale de x est alors

$$x|\tau^2 \sim \mathcal{N}_p(0, (1+\tau^2)I_p)$$

et conduit à l'estimateur du maximum de vraisemblance de τ^2 suivant,

$$\hat{\tau}^2 = \begin{cases} (||x||^2/p) - 1 & \text{si } ||x||^2 > p\,, \\ 0 & \text{sinon.} \end{cases}$$

L'estimateur bayésien empirique correspondant de θ_i sous coût quadratique est obtenu en remplaçant τ^2 par $\hat{\tau}^2$ dans l'estimateur de Bayes,

$$\begin{aligned} \delta^{\mathrm{EB}}(x) &= \frac{\hat{\tau}^2 x}{1+\hat{\tau}^2} \\ &= \left(1 - \frac{p}{||x||^2}\right)^+ x. \end{aligned} \tag{10.17}$$

L'estimateur (10.17) est en fait un estimateur tronqué de James-Stein. Par conséquent, ces estimateurs peuvent être interprétés en tant qu'estimateurs bayésiens empiriques liés à l'information que les espérances des observations sont proches de 0. L'estimateur originel de James-Stein peut également s'écrire comme un estimateur bayésien empirique, avec une méthode d'estimation fréquentiste alternative. En réalité, étant donné la loi marginale de x, le meilleur estimateur sans biais de $1/(1+\tau^2)$ est $(p-2)/||x||^2$, ce qui conduit à

$$\delta^{\mathrm{EB}}(x) = \left(1 - \frac{p-2}{||x||^2}\right) x. \tag{10.18}$$

‖

Cet exemple illustre aussi les lacunes dans les justifications de l'approche bayésienne empirique, qui ne sait pas comparer les différentes méthodes d'estimation des hyperparamètres. Ce problème de défaut de classement est plus largement caractéristique de l'approche fréquentiste dans son ensemble. La comparaison entre les estimateurs (10.17) et (10.18) doit être fondée sur d'autres considérations.

Exemple 10.33. Soient deux vecteurs indépendants, $x \sim \mathcal{N}_p(\theta, \sigma^2 I_p)$ et $y \sim \mathcal{N}_q(0, \sigma^2 I_q)$, comme dans la régression linéaire. Le paramètre d'intérêt est le facteur de variance σ^2, évalué sous coût entropique,

$$\mathrm{L}(\sigma^2, d) = \frac{d}{\sigma^2} - \log(d/\sigma^2) - 1.$$

Outre les considérations intrinsèques (Section 2.5.4), une raison pour laquelle on peut préférer ce coût au coût quadratique est qu'il induit l'estimateur

du maximum de vraisemblance $||y||^2/p+q$ en tant que meilleur estimateur équivariant[78] de σ^2. Sous ce coût, l'estimateur de Bayes de σ^2 est

$$\delta^\pi(x) = \left(\mathbb{E}^\pi[\sigma^{-2}|x]\right)^{-1}. \tag{10.19}$$

Pour la loi conjuguée gamma-normale sur (θ, σ^2),

$$\theta|\sigma^2 \sim \mathcal{N}_p(0, \tau\sigma^2 I_p), \qquad \sigma^{-2} \sim \mathcal{G}(\nu/2, \beta/2)\,,$$

l'estimateur (10.19) est alors

$$\delta^\pi(x,y) = \frac{1}{p+q+\nu}\left(\frac{||x||^2}{1+\tau} + ||y||^2 + \beta\right)$$

et la maximisation de la vraisemblance marginale (en (τ, ν, β)) conduit à l'estimateur bayésien empirique suivant (voir Kubokawa *et al.*, 1993b) :

$$\delta^{\mathrm{EB}}(x,y) = \min\left(\frac{||y||^2}{q}, \frac{||x||^2+||y||^2}{p+q}\right). \tag{10.20}$$

Mettons en évidence l'aspect intuitif de cet estimateur : il n'utilise l'information additionnelle dans x concernant σ^2 que si $||x||^2$ n'est pas trop grand, c'est-à-dire si θ est proche de 0, puisque

$$\frac{||x||^2+||y||^2}{p+q}$$

est le meilleur estimateur équivariant d'échelle de σ^2 lorsque $\theta = 0$.

L'intérêt réel de ce résultat est montré par Brewster et Zidek (1974) : l'estimateur (10.20) améliore uniformément le meilleur estimateur équivariant $\delta^\star(x,y) = ||y||^2/q$ sous coût entropique. (Voir Maatta et Casella, 1990, pour une étude détaillée des différents aspects de l'estimation de variance.) ‖

Morris (1983a) considère l'effet Stein en plus grande généralité que dans le cadre de l'Exemple 10.32. Il étudie en fait le modèle bayésien de régression

$$\begin{aligned} x|\theta &\sim \mathcal{N}_p(\theta, \Lambda), \\ \theta|\beta, \sigma_\pi^2 &\sim \mathcal{N}_p(Z\beta, \sigma_\pi^2 I_p), \end{aligned}$$

avec $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ et Z matrice $(p \times q)$ de rang plein. La loi marginale de x est alors

$$x_i|\beta, \sigma_\pi^2 \sim \mathcal{N}(z_i'\beta, \sigma_\pi^2 + \lambda_i)$$

et la loi a posteriori de θ est

[78] Cet argument ne saurait justifier l'utilisation du coût entropique, puisqu'il légitime a posteriori un estimateur donné, au lieu d'utiliser des considérations utilitaires conduisant à une détermination pratique de l'estimateur.

$$\theta_i|x_i,\beta,\sigma_\pi^2 \sim \mathcal{N}\left((1-b_i)x_i + b_i z_i'\beta, \lambda_i(1-b_i)\right),$$

avec $b_i = \lambda_i/(\lambda_i+\sigma_\pi^2)$. Si toutes les variances λ_i sont identiques et égales à σ^2, les meilleurs estimateurs équivariants de β et b sont donnés par

$$\hat{\beta} = (Z^tZ)^{-1}Z^tx \qquad \text{et} \qquad \hat{b} = \frac{(p-q-2)\sigma^2}{s^2},$$

avec $s^2 = \sum_{i=1}^p (x_i - z_i'\hat{\beta})^2$. On déduit de ces estimateurs des hyperparamètres l'estimateur bayésien empirique de θ suivant :

$$\delta^{\mathrm{EB}}(x) = Z\hat{\beta} + \left(1 - \frac{(p-q-2)\sigma^2}{||x - Z\hat{\beta}||^2}\right)(x - Z\hat{\beta}), \tag{10.21}$$

qui est de la forme des estimateurs de Stein généraux.

Dans le cas particulier où on suppose que les moyennes sont identiques (hypothèse d'échangeabilité), la matrice Z est réduite au vecteur $\mathbf{1}$ et β est un nombre réel ; l'estimateur bayésien empirique est alors

$$\delta^{\mathrm{EB}}(x) = \bar{x}\mathbf{1} + \left(1 - \frac{(p-3)\sigma^2}{||x - \bar{x}\mathbf{1}||^2}\right)(x - \bar{x}\mathbf{1}).$$

Il s'agit donc de l'estimateur de Stein qui rétrécit vers la moyenne commune, présenté dans Efron et Morris (1975). Voir Morris (1983b) pour le cas où les variances λ_i ne sont pas identiques.

10.5.2 Évaluation de la variance

Comme nous l'avons décrit ci-dessus, l'estimation des hyperparamètres β et σ_π^2 modifie considérablement le comportement des procédures résultantes. Bien que nous venions de voir que les estimateurs ponctuels obtenus sont en général efficaces, cette approche sous-estime la variance a posteriori de $\pi(\theta|x,\beta,b)$ en utilisant la variance empirique $\mathrm{var}(\theta_i|x,\hat{\beta},\hat{b})$. Il est donc trompeur d'utiliser l'analyse bayésienne empirique pour étudier les performances de δ^{EB} en estimant son coût quadratique $(\theta_i - \delta_\pi^{\mathrm{EB}})^2$ par $\mathrm{var}(\theta_i|x,\hat{\beta},\hat{b})$, car on obtient une sous-estimation de l'erreur résultant de l'utilisation de δ^{EB}.

Morris (1982) considère la variabilité supplémentaire issue de l'estimation des hyperparamètres en modifiant les estimateurs. Dans le cas échangeable, les procédures obtenues sont

$$\begin{aligned}\delta^{\mathrm{EB}}(x) &= x - \tilde{B}(x - \bar{x}\mathbf{1}),\\ V_i^{\mathrm{EB}}(x) &= \left(\sigma^2 - \frac{p-1}{p}\tilde{B}\right) + \frac{2}{p-3}\hat{b}(x_i - \bar{x})^2,\end{aligned}$$

avec

$$\hat{b} = \frac{p-3}{p-1}\ \frac{\sigma^2}{\sigma^2+\hat{\sigma}_\pi^2}, \qquad \hat{\sigma}_\pi^2 = \max\left(0, \frac{||x-\bar{x}\mathbf{1}||^2}{p-1} - \sigma_\pi^2\right)$$

et

$$\tilde{B} = \frac{p-3}{p-1}\ \min\left(1, \frac{\sigma^2(p-1)}{||x-\bar{x}\mathbf{1}||^2}\right).$$

Cette dernière quantité estime le rapport $\sigma^2/(\sigma^2+\sigma_\pi^2)$. Cependant, cette modification, bien que plus satisfaisante, souffre toujours du défaut général de l'inférence bayésienne empirique, à savoir que les procédures sont souvent justifiées par des raisons sur mesure (ou *empiriques*!) qui ne peuvent être généralisées en un principe (même si Kass et Steffey, 1989, présentent une généralisation partielle).

Remarquons l'analogie entre la variance empirique modifiée V_i^{EB} et la variance hiérarchique pour le même modèle,

$$V_i^{\mathrm{HB}}(x) = \sigma^2\left(1 - \frac{p-1}{p}\ \mathbb{E}^\pi\left[\frac{\sigma^2}{\sigma^2+\sigma_\pi^2}\middle| x\right]\right) + \mathrm{var}\left[\frac{\sigma^2}{\sigma^2+\sigma_\pi^2}\middle| x\right](x_i-\bar{x})^2$$

(Berger, 1985b). Cette ressemblance n'est pas une coïncidence, puisque cette modification améliore l'approche bayésienne empirique originelle en empruntant un peu plus à l'analyse bayésienne réelle. Ghosh et Saleh (1989) et Blattberg et George (1991) présentent des exemples d'utilisation de l'analyse bayésienne empirique en Économétrie et établissent un lien avec les estimateurs de Stein dans les modèles de régression.

10.5.3 Régions de confiance

Il existe une autre caractéristique de l'effet Stein interprétable dans un contexte bayésien empirique. Dans le cas des *régions de confiance recentrées* (Section 5.5), Hwang et Casella (1982) montrent que, à volume égal, certaines régions ont une probabilité de couverture plus grande que la moyenne. Ces ensembles correspondent alors à des régions HPD empiriques.

Exemple 10.34. Hwang et Casella (1982) comparent la région de confiance habituelle

$$C_0(x) = \{\theta;\ \|\theta - x\|^2 \le c_\alpha\},$$

où $x \sim \mathcal{N}_p(\theta, I_p)$ à

$$C_a(x) = \{\theta;\ \|\theta - \delta_a(x)\|^2 \le c_\alpha\},$$

avec $\delta_a(x) = [1-(a/||x||^2)]^+x$. Ils montrent que, pour a suffisamment petit et $p \ge 4$, l'ensemble C_a vérifie, quel que soit θ,

$$P_\theta(\theta \in C_a(x)) > P_\theta(\theta \in C_0(x)) = 1-\alpha.$$

Casella et Hwang (1983) considèrent également les régions recentrées de *volume variable*

$$C_\delta^v(x) = \{\theta;\ \|\theta - \delta(x)\|^2 \le v(x)\}$$

et ils déterminent δ et v par une analyse bayésienne empirique à partir d'une région HPD α-crédible. Le centre de la région est l'estimateur de James-Stein

$$\delta(x) = \left(1 - \frac{p-2}{||x||^2}\right)^+ x$$

et le rayon est donné par

$$v(x) = \begin{cases} \left(1 - \frac{p-2}{c_\alpha}\right)\left[c_\alpha - p\log\left(1 - \frac{p-2}{c_\alpha}\right)\right] & \text{si } ||x||^2 < c_\alpha\,, \\ \left(1 - \frac{p-2}{||x||2}\right)\left[c_\alpha - p\log\left(1 - \frac{p-2}{||x||2}\right)\right] & \text{sinon.} \end{cases}$$

La forme du rayon variable est justifiée par un coût linéaire

$$\mathrm{L}(\theta, C) = k\ \mathrm{vol}(C) - \mathbb{I}_C(\theta),$$

déjà vu à la Section 5.5 (Exercice 10.29). Cette région de confiance bayésienne empirique a alors un niveau de confiance d'au moins $1-\alpha$ (au sens fréquentiste), sauf pour les plus petites valeurs de p. ‖

Exemple 10.35. Une objection classique à l'encontre des régions de confiance recentrées repose sur leur inutilité pratique dans la mesure où le niveau de confiance rapporté est toujours

$$\inf_\theta P_\theta(\theta \in C_a(x)) = 1 - \alpha = P_\theta(\theta \in C_0(x)).$$

En ce sens, on peut dire que les régions standard sont plus précises puisqu'elles coïncident exactement avec le niveau de confiance rapporté. Nous avons déjà parlé de la valeur intrinsèque de ces niveaux de confiance à la Section 5.5 et les lecteurs pourront se référer au Chapitre 5 pour des commentaires sur le côté artificiel de la notion de niveau de confiance. On propose aussi une voie alternative à la fin du Chapitre 5 : il s'agit d'un niveau de confiance conditionnel, $\gamma(x)$, plus adapté à la région recentrée $C_a(x)$, qu'on évalue sous coût quadratique

$$(\gamma(x) - \mathbb{I}_{C_a(x)})^2. \tag{10.22}$$

Pour le modèle présenté dans l'Exemple 10.33, George et Casella (1994) proposent une solution bayésienne empirique à ce problème d'évaluation avec une région recentrée de la forme

$$C^{\mathrm{EB}}(x) = \{\theta;\ \|\theta - (1-\hat{b})x\|^2 \le c\}$$

et un rapport de confiance

$$\gamma^{\mathrm{EB}}(x) = P(\chi_p^2 \le c/(1-\hat{b})).$$

Si $\theta \sim \mathcal{N}_p(0, \tau^2 I_p)$, la réponse bayésienne serait

$$\begin{aligned}\gamma^\pi(x) &= P^\pi(\theta \in C_B(x)|x) \\ &= P^\pi(||\theta - (1-b)x||^2 \leq c|x) \\ &= P(\chi^2_p \leq c/(1-b)),\end{aligned}$$

puisque $\theta|x \sim \mathcal{N}_p((1-b)x, (1-b))$ avec $1-b = \tau^2/(\sigma^2+\tau^2)$. Les estimateurs bayésiens empiriques établis par George et Casella (1994) dans γ^{EB} sont

$$1 - \hat{b}(x) = \max\left(d, 1 - \frac{a}{||x||^2}\right) = u_{a,d}(||x||^2),$$

et C^{EB} est centré sur l'estimateur de Stein tronqué associé à a et $d \leq 1$. George et Casella (1994) montrent de plus que l'estimateur bayésien empirique ainsi obtenu

$$\gamma^{\mathrm{EB}}(x) = P\left[\chi^2_p \leq \frac{c}{\max\{d, (||x||^2 - a)/||x||^2\}}\right],$$

domine le rapport constant $1-\alpha$ sous coût quadratique (10.22), pour $d \leq 1$ et a suffisamment petit. Une valeur possible pour d est

$$d = \frac{2c}{c + 2a + \sqrt{c(c+4a)}}.$$

Voir Lu et Berger (1989b) pour une autre solution. ‖

10.5.4 Commentaires

Pour conclure cette étude des méthodes bayésiennes empiriques, revenons sur leur nature duale : ces procédures inférentielles s'inspirent à la fois de méthodes fréquentistes et bayésiennes et on peut penser que les améliorations qu'elles apportent aux estimateurs fréquentistes classiques proviennent de l'approche bayésienne. Mais leur sous-optimalité (notamment en termes d'admissibilité) peut être attribuée au refus d'adopter un point de vue complètement bayésien et à l'arbitraire en résultant dans les choix de méthodes. Il est finalement assez logique qu'une méthode qui repose sur des estimateurs classiques mais sous-optimaux (tels que l'estimateur du maximum de vraisemblance de la moyenne dans le cas normal multidimensionnel) et sur des techniques de circonstance non légitimées par la Théorie de la Décision (comme l'estimation sans biais ou la méthode des moments) ne puisse déboucher sur des procédures optimales. Le fait que ces estimateurs soient dominés par de vrais estimateurs de Bayes (Brown, 1988) est un argument supplémentaire pour l'adoption sans restriction du paradigme bayésien, même s'il nécessite une modélisation hiérarchique. Le développement de nouvelles techniques numériques (Chapitre 6), permettant aujourd'hui de traiter des modèles bien plus complexes qu'avant, apparaît comme un coup de grâce porté à ces méthodes empiriques qui présentaient auparavant l'avantage d'alléger la lourdeur des calculs des analyses bayésiennes complètes.

10.6 Exercices

Section 10.1

10.1 Dans le cas d'un modèle représenté par un graphe acyclique orienté, comme celui de la Figure 10.1, montrer que la densité conditionnelle complète d'une variable (ou *nœud*) sachant les autres variables du modèle est la même que la loi de ce nœud sachant uniquement les nœuds auxquels il est connecté.

10.2 Dans le cadre de l'Exemple 10.1,

a. Montrer que, si le générateur Λ peut se décomposer en $P\tilde{\Lambda}P^t$, où P est la matrice orthogonale des vecteurs propres de Λ et $\tilde{\Lambda}$ est la matrice diagonale des valeurs propres de Λ, λ_i $(i = 1, \ldots, 7)$, alors

$$\exp\{\Lambda\} = P \begin{pmatrix} e^{\lambda_1} & 0 & \ldots & 0 \\ 0 & e^{\lambda_2} & \ldots & 0 \\ & & \ddots & \\ 0 & 0 & \ldots & e^{\lambda_7} \end{pmatrix} P^t \,,$$

et $\exp\{T\Lambda\} = \exp\{\Lambda\}^T$.

b. En déduire que les formules de récurrence forward-backward (6.31) de l'Exercice 6.51 s'appliquent ici en remplaçant p_{ij} par $p_{ij}^{(T)}$, élément (i, j) de la matrice $\exp\{\Lambda\}^T$.

c. Déterminer la complexité numérique de ces formules et comparer avec la représentation alternative suivante : introduire des valeurs manquantes x_{ij}^* telles que les individus soient observés à intervalles réguliers avec un temps interobservations de η, ajouter ces valeurs manquantes à l'échantillon et calculer les formules de forward-backward sur l'échantillon complet. (Remarquer que cette opération fait que $\exp\{\Lambda\}^\eta$ n'est calculée qu'une fois.)

Section 10.2.1

10.3 Montrer que l'hyper a priori choisi dans l'Exemple 10.4 donne une loi a posteriori bien définie s'il y a au moins deux observations pour chaque étape de l'expérience.

10.4 Représenter le modèle hiérarchique de l'Exemple 10.4 sous la forme d'un graphe acyclique orienté, comme celui de la Figure 10.1.

10.5 Soit $J \sim \mathscr{M}_k(N; p_1, \ldots, p_k)$ une variable aléatoire multinomiale. On suppose que N est simulé selon une loi de Poisson de paramètre λ. Déterminer la loi marginale de J. Donner, en particulier, la matrice de covariance. Généraliser au cas $p = (p_1, \ldots, p_k) \sim \mathscr{D}(\alpha_1, \ldots, \alpha_k)$, loi de Dirichlet.

10.6 Une mouche pond N œufs selon une loi de Poisson $\mathscr{P}(\lambda)$ et chaque œuf survit avec probabilité p.

a. Montrer que la loi du nombre d'œufs survivants x est alors hiérarchique

$$x|N \sim \mathscr{B}(N, p), \qquad N \sim \mathscr{P}(\lambda).$$

b. Calculer la distribution marginale de x et la distribution a posteriori de N.

10.7 Dans le cadre de l'Exemple 10.6, avec p connu, donner la loi a posteriori de N si $\pi_2(\lambda) = 1/\lambda$. Étudier la généralisation au cas où p est inconnu et $\pi_1(p) = 1$.

Section 10.2.2

10.8 Si $y|\theta \sim \mathcal{N}_p(\theta, \Sigma_1)$, $\theta|\beta \sim \mathcal{N}_p(X\beta, \Sigma_2)$ et $\beta \sim \mathcal{N}_q(\mu, \Sigma_3)$, établir les lois a priori et a posteriori de θ.

10.9 On se place dans un cadre de *régression logistique*, c'est-à-dire qu'on considère des observations $(x_1, y_1), \ldots, (x_n, y_n)$ telles que $x_i \in \mathbb{R}^k$ et $y_i \in \{0, 1\}$ avec

$$P(y_i = 1|x_i) = \exp(x_i^t\beta)/(1 + \exp(x_i^t\beta))\,.$$

Déterminer une condition suffisante sur $\pi(\tau)$ pour que la distribution a posteriori de β soit définie lorsque $\beta|\tau \sim \mathcal{N}_q(0, \tau^2 I_p)$. (Les x_i sont supposés fixés.)

10.10 Reprendre l'Exercice 10.9 avec un modèle *probit*, c'est-à-dire avec

$$P(y_i = 1|x_i) = \Phi(x_i^t\beta)$$

et Φ fonction de répartition de la loi normale standard.

Section 10.2.3

10.11 Établir les Lemmes 10.10 et 10.12.

10.12 *(Berger et Robert, 1990) Dans le cadre de l'Exemple 10.15, on suppose que $\mu \in H = \{\mu = Y\beta;\ \beta \in \mathbb{R}^\ell\}$ et $\pi_2(\beta, \sigma_\pi^2) = 1$. Montrer que $m(x) < +\infty$ si $p > 2 + \ell$.

10.13 Spiegelhalter et Lauritzen (1990) présentent le modèle suivant : les poids y_{ij} de soixante rats sont relevés chaque semaine, les trente premières observations constituant le groupe de contrôle ($1 \le i \le 60, 1 \le j \le 5$). Le modèle associé est

$$y_{ij} \sim \mathcal{N}(\alpha_i + \beta_i j, \sigma_i)\,,$$

avec $\sigma_i = \sigma_c$ pour $i \le 30$, $\sigma_i = \sigma_t$ pour $31 \le i \le 60$ et

$$\begin{aligned}(\alpha_i, \beta_i) &\sim \mathcal{N}_2\left((\alpha_c, \beta_c), \Sigma_c\right) \qquad (i = 1, \ldots, 30)\,,\\ (\alpha_i, \beta_i) &\sim \mathcal{N}_2\left((\alpha_t, \beta_t), \Sigma_t\right) \qquad (i = 31, \ldots, 60)\,.\end{aligned}$$

Compléter le modèle avec des lois a priori non informatives sur les hyperparamètres et étudier si la distribution a posteriori est bien définie.

10.14 Dans le cadre de l'Exemple 10.11, on définit les moyennes

$$\overline{x}_i = \frac{1}{J_i^c}\sum_{i=1}^{J_i^c} x_{ij}, \quad \overline{y}_i = \frac{1}{J_i^a}\sum_{i=1}^{J_i^a} y_{ij}, \quad \overline{z}_i = \frac{1}{J_i^t}\sum_{i=1}^{J_i^t} z_{ij}$$

et

$$\overline{\theta} = \frac{1}{I}\sum_{i=1}^{I}\theta_i, \qquad \overline{\delta} = \frac{1}{I}\sum_{i=1}^{I}\delta_i.$$

Montrer que les densités conditionnelles complètes s'écrivent ($1 \le i \le I$)

$$\begin{aligned}&\mu_\theta \sim \mathcal{N}(\overline{\theta}, \sigma_\theta^2/I), && \mu_\delta \sim \mathcal{N}(\overline{\delta}, \sigma_\delta^2/I),\\ &\mu_P \sim \mathcal{N}\left(\sum_{\ell_i=0}\xi_i/I_P, \sigma_P^2/I_P\right), && \mu_D \sim \mathcal{N}\left(\sum_{\ell_i=1}\xi_i/I_D, \sigma_D^2/I_D\right),\end{aligned}$$

$$\theta_i \sim \mathcal{N}\left(\frac{\sigma_\theta^{-2}\mu_\theta + J_i^c\sigma_c^{-2}\overline{x}_i + J_i^a\sigma_a^{-2}(\overline{y}_i-\delta_i) + J_i^t\sigma_t^{-2}(\overline{z}_i-\delta_i-\xi_i)}{\sigma_\theta^{-2}+J_i^c\sigma_c^{-2}+J_i^a\sigma_a^{-2}+J_i^t\sigma_t^{-2}}, (\sigma_\theta^{-2}+J_i^c\sigma_c^{-2}+J_i^a\sigma_a^{-2}+J_i^t\sigma_t^{-2})^{-1}\right)$$

$$\delta_i \sim \mathcal{N}\left(\frac{\sigma_\delta^{-2}\mu_\delta + J_i^a\sigma_a^{-2}(\overline{y}_i-\theta_i) + J_i^t\sigma_t^{-2}(\overline{z}_i-\theta_i-\xi_i)}{\sigma_\delta^{-2}+J_i^a\sigma_a^{-2}+J_i^t\sigma_t^{-2}}, (\sigma_\delta^{-2}+J_i^a\sigma_a^{-2}+J_i^t\sigma_t^{-2})^{-1}\right)$$

$$\xi_i \sim \mathcal{N}\left(\frac{\sigma_D^{-2\ell_i}\sigma_P^{-2(1-\ell_i)}\mu_D^{\ell_i}\mu_P^{1-\ell_i} + J_i^t\sigma_t^{-2}(\overline{z}_i-\theta_i-\delta_i)}{\sigma_D^{-2\ell_i}\sigma_P^{-2(1-\ell_i)}+J_i^t\sigma_t^{-2}}, (\sigma_D^{-2\ell_i}\sigma_P^{-2(1-\ell_i)}+J_i^t\sigma_t^{-2})^{-1}\right)$$

$$\sigma_c^{-2} \sim \mathcal{G}a\left(\sum_i \frac{J_i^c}{2}, \sum_{i,j}\frac{(x_{ij}-\theta_i)^2}{2}\right),$$

$$\sigma_a^{-2} \sim \mathcal{G}a\left(\sum_i \frac{J_i^a}{2}, \sum_{i,j}\frac{(y_{ij}-\theta_i-\delta_i)^2}{2}\right),$$

$$\sigma_t^{-2} \sim \mathcal{G}a\left(\sum_i \frac{J_i^t}{2}, \sum_{i,j}\frac{(z_{ij}-\theta_i-\delta_i-\xi_i)^2}{2}\right),$$

$$\sigma_\theta^{-2} \sim \mathcal{G}a\left(\tfrac{I}{2}, \sum_i\frac{(\theta_i-\mu_\theta)^2}{2}\right), \quad \sigma_\delta^{-2} \sim \mathcal{G}a\left(\tfrac{I}{2}, \sum_i\frac{(\delta_i-\mu_\delta)^2}{2}\right),$$

$$\sigma_P^{-2} \sim \mathcal{G}a\left(\tfrac{I_P}{2}, \sum_{\ell_i=0}\frac{(\xi_i-\mu_P)^2}{2}\right), \ \sigma_D^{-2} \sim \mathcal{G}a\left(\tfrac{I_D}{2}, \sum_{\ell_i=1}\frac{(\xi_i-\mu_D)^2}{2}\right).$$

10.15 (Suite de l'Exercice 10.14) Lorsque les δ_i sont distribués selon (10.3), donner les densités conditionnelles complètes correspondantes.

Section 10.2.4

10.16 *(Berger et Robert, 1990) Soient $x \sim \mathcal{N}_p(\theta, \Sigma)$, $\theta \sim \mathcal{N}_p(y\beta, \sigma_\pi^2 I_p)$, et $\beta \sim \mathcal{N}_\ell(\beta_0, A)$, avec $\text{rang}(A) = m$.

a. Montrer que si, pour $K > 0$, les deux intégrales

$$\int_0^K \pi_2(\sigma_\pi^2)d\sigma_\pi^2 \quad \text{et} \quad \int_K^{+\infty} \frac{1}{(\sigma_\pi^2)^{(p-\ell+m)/2}}\pi_2(\sigma_\pi^2)d\sigma_\pi^2$$

sont finies, alors $m(x) < +\infty$ pour tout $x \in \mathbb{R}^p$.

b. Montrer que la condition a. est satisfaite si, pour $\epsilon > 0$, $K_1 > 0$, $K_2 > 0$,

$$\pi_2(\sigma_\pi^2) < \frac{K_1}{K_2 + (\sigma_\pi^2)^{(2+\epsilon-p+\ell-m)/2}},$$

c'est-à-dire si $\pi_2(\sigma_\pi^2) = 1$ et $p - l + m > 2$.

10.17 *(Berger, 1985b) Dans le cadre de l'Exemple 10.19, calculer la variance a posteriori. Étudier également le cas non informatif.

10.18 (Lindley et Smith, 1972) Élargir l'Exemple 10.19 au modèle général

$$x|\theta \sim \mathcal{N}_p(A_1\theta, \Sigma_1), \qquad \theta|\beta \sim \mathcal{N}_\ell(A_2\beta, \Sigma_2), \qquad \beta|\xi \sim \mathcal{N}_q(A_3\xi, \Sigma_3),$$

et vérifier les résultats de l'Exemple 10.8.

10.19 (Berger, 1985b) Montrer que, pour le modèle de l'Exemple 10.19 avec des lois non informatives sur ξ et σ_π^2, l'estimateur bayésien hiérarchique est

$$\delta^\pi(x) = x - h_{p-2}(||x - \bar{x}\mathbf{1}||^2)(x - \bar{x}\mathbf{1})$$

avec

$$h_p(t) = \frac{p}{2t}(1 - H_p(t)),$$

$$H_p(t) = \begin{cases} \dfrac{t^{p/2}}{(p/2)!\left\{e^t - \sum_{i=1}^{(p-2)/2} t^i/i!\right\}} & \text{si } p \text{ est pair,} \\ \dfrac{t^{p/2}}{\Gamma(p/2)\left\{e^t[2\Phi(\sqrt{2t}) - 1] - \sum_{i=1}^{(p-3)/2} \frac{t^{(i+3)/2}}{\Gamma(i+3/2)}\right\}} & \text{si } p \text{ est impair.} \end{cases}$$

10.20 Dans le cadre de l'Exemple 10.14, calculer la moyenne a posteriori de p quand $x = 3,\ n = 5$.

Section 10.2.5

10.21 Comparer les modèles

$$x \sim \mathscr{N}_p(\theta, I_p), \quad \theta|\mu \sim \mathscr{N}_p(\mu, \tau^2 I_p), \quad \pi_2(\mu, \tau^2) = 1/\tau^2,$$

et

$$x \sim \mathscr{N}_p(\theta, I_p), \quad \theta|\mu \sim \mathscr{N}_p(\mu, I_p), \quad \mu|\xi \sim \mathscr{N}_p(\xi, \tau^2 I_p), \quad \pi_2(\xi, \tau^2) = 1/\tau^2,$$

selon les estimateurs de θ.

10.22 Soient $x_i \sim \mathscr{N}(\mu_i, \sigma^2)$ et $\mu_i|\mu, \tau \sim \mathscr{N}(\mu, \tau^2)$ $(i = 1, \ldots, n)$.

a. Montrer que $\pi(\mu, \tau) = 1/\tau$ conduit à une loi a posteriori indéfinie.

b. Montrer que $\pi(\mu, \tau) = 1$ permet de contourner le problème ci-dessus.

10.23 Dans le cadre de l'Exemple 10.8, montrer que l'estimateur de Bayes

$$\delta^\pi(y) = \mathbb{E}^{\pi_2(\sigma_\pi^2|y)}\left[I_p + \frac{\sigma^2}{\sigma_\pi^2}(X^tX)^{-1}\right]^{-1}\hat{\beta}$$

peut s'écrire sous la forme

$$\left[I_p + h(y)(X^tX)^{-1}\right]^{-1}\hat{\beta}.$$

(*Indication :* Utiliser une diagonalisation conjointe de I_p et X^tX.) Expliquer comment cet estimateur peut aider à réduire la multicolinéarité.

Section 10.3

10.24 *(Stein, 1981) Démontrer le Lemme 10.21 à l'aide d'une intégration par parties et mettre ce résultat en relation avec l'Exercice 2.56.

10.25 *Si H est la matrice hessienne définie au Lemme 10.21, montrer que l'équivalent de (10.10) pour la matrice de covariance est

$$V^{\mathrm{EB}}(x) = \Sigma + \Sigma \frac{H(x)}{m(x)} \Sigma - \Sigma (\nabla \log m(x))(\nabla \log m(x))^t \Sigma.$$

En utilisant une technique analogue à l'Exercice 10.24, montrer qu'un estimateur sans biais de l'erreur matricielle moyenne

$$\mathbb{E}_\theta[(\theta - \delta(x))(\theta - \delta(x))^t]$$

peut s'écrire sous la forme différentielle

$$\hat{V}_{\delta^{\mathrm{HB}}}(x) = \Sigma + 2\Sigma \frac{H(x)}{m(x)} \Sigma - \Sigma (\nabla \log m(x))(\nabla \log m(x))^t \Sigma.$$

Déduire de cette expression l'estimateur sans biais du risque quadratique.

10.26 *En utilisant l'approximation suivante de ${}_1F_1(a;b;z)$:

$${}_1F_1(a;b;z) \simeq \frac{\Gamma(b)}{\Gamma(a)} e^{z/2} (z/2)^{a-b} \left(1 + \frac{(1-a)(b-a)}{(z/2)}\right),$$

donner une approximation de l'estimateur δ^π de l'Exemple 10.26 et la comparer à l'estimateur de James-Stein.

10.27 On considère $x \sim \mathcal{N}_p(\theta, I_p)$, $\theta \sim \mathcal{N}_p(0, \tau^2 I_p)$ et, si $\eta = 1/(1+\tau^2)$, on suppose que $\pi_2(\eta) = \eta^{2-(p/2)}$. Montrer que l'estimateur bayésien hiérarchique correspondant peut s'écrire explicitement

$$\delta^{\mathrm{HB}}(x) = \left(\frac{1}{1 - e^{-||x||^2/2}} - \frac{2}{||y||^2}\right) x\,,$$

et déterminer s'il est minimax et admissible.

10.28 *(Hartigan, 1983) Soit une observation $x \sim \mathcal{N}_p(\theta, I_p)$.

a. Si f est une fonction positive croissante majorée par $2(p-2)$, montrer que

$$\delta_f(x) = \left(1 - \frac{f(||x||^2)}{||x||^2}\right) x$$

domine $\delta_0(x) = x$ pour le coût quadratique usuel. (*Indication :* Utiliser l'estimateur sans biais du risque obtenu dans l'Exercice 2.56.)

b. Soit π un a priori sur θ tel que, conditionnellement à τ^2, $\theta \sim \mathcal{N}_p(0, \tau^2)$ et $\tau^2 \sim \pi_1$. On suppose que l'hyper a priori π_1 est une fonction log-concave de $\log(\tau^2+1)$ et que $(\tau^2+1)^{1-\alpha}\pi_1(\tau^2)$ est strictement croissante en τ^2. À l'aide du résultat général de a., montrer que l'estimateur bayésien hiérarchique associé à π domine δ_0 si $4 - 2\alpha \leq p$. (*Indication :* Montrer que $\delta^\pi(x) = (1 - \mathbb{E}[(\tau^2+1)^{-1}|x])x$ et que $\mathbb{E}[(\tau^2+1)^{-1}|x])$ est strictement croissante en $||x||^2$ tout en étant clairement majorée par $2(p-2)$.)

c. Montrer que de telles lois a priori ne peuvent être propres que pour $\alpha < 0$ et donc qu'on ne peut garantir l'admissibilité de ces estimateurs de Bayes minimax que pour $p \geq 5$.

d. Montrer que le risque de Bayes est en fait fini pour $\alpha < 2$ et en déduire que les estimateurs bayésiens hiérarchiques sont admissibles quel que soit p. [*Note :* Strawderman, 1971, a montré dans le cas particulier $\pi_1(\tau^2) = (1+\tau^2)^{\alpha-1}$ que la dimension limite pour l'existence d'estimateurs de Bayes minimax propres est précisément $p = 5$.]

Section 10.4.2

Tab. 10.2. Intentions d'achat de voiture par foyer.

Intentions	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Réponses	293	26	21	21	10	9	12	13	11	10	21

10.29 (Casella, 1985b) Dans un sondage sur les intentions d'achat de voiture, quatre cent quarante-sept foyers évaluent leur probabilité de se procurer un nouveau véhicule au cours de l'année à venir. Les résultats de ce sondage sont donnés dans le Tableau 10.2.

Les réponses x_i ($1 \leq i \leq 447$) sont modélisées par une loi binomiale renormalisée $\mathcal{B}(10, p_i)$, c'est-à-dire que $10x_i \sim \mathcal{B}(10, p_i)$, et les p_i sont distribués suivant une loi $\mathcal{B}e(\alpha, \beta)$.

a. Utiliser la distribution marginale pour donner des estimateurs de α et β par la méthode des moments.

b. Calculer un estimateur bayésien empirique des p_i sous coût quadratique.

c. Les vraies intentions p_i étant connues à la fin de l'année, on peut rapporter dans le Tableau 10.3 les différences avec les déclarations initiales. Comparer les coûts quadratiques de l'estimateur classique (c'est-à-dire $\hat{p}_i = x_i$), de l'estimateur bayésien empirique et d'un estimateur de Bayes de votre choix.

Tab. 10.3. Proportions d'achats de voitures en fonction des probabilités d'intention.

Intentions	0	0.1—0.3	0.4—0.6	0.7—0.9	1
Déclarations	0	0.19	0.51	0.79	1
Réalisations	0.07	0.19	0.41	0.48	0.583

10.30 Déterminer l'équivalent de la Proposition 10.31 pour le test $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$ pour deux problèmes indépendants d'échantillons $x_1, \ldots, x_n \sim f(x|\theta)$, $y_1, \ldots, y_m \sim f(y|\theta')$ et $P(\theta = \theta_0) = P(\theta' = \theta_0) = \pi_0$.

Généraliser à p échantillons et appliquer au cas du test de $\theta_i = 0$ contre $\theta_i = 1$ pour $x_i \sim \mathcal{N}(\theta_i, 1)$ ($1 \leq i \leq p$).

10.31 *(Hartigan, 1983) On considère $x \sim \mathcal{N}_p(\theta, \sigma^2 I_p)$ et $\theta \sim \mathcal{N}_p(0, \tau^2 I_p)$, avec σ^2 inconnu et $s^2 \sim \sigma^2 \chi^2_k$.

a. Donner un estimateur bayésien empirique de θ à partir des estimateurs de maximum de vraisemblance de τ^2 et σ^2 et déterminer si l'estimateur obtenu est minimax. (*Indication :* Utiliser l'Exercice 10.28.)

b. Comparer aux estimateurs bayésiens empiriques utilisant les estimateurs des moments de σ^2 et τ^2.

c. Si $\pi(\sigma^2,\tau^2) \propto (\sigma^2+\sigma_0^2)^{\alpha-1}(\sigma^2)^{\beta-1}$, montrer que la distribution a posteriori de $(\sigma^{-2},(\sigma^2+\tau^2)^{-1})$ est

$$\chi^2_{k-2\beta}/s \times \chi^2_{p-2\alpha}/||x||^2 \mathbb{I}_{\sigma^2\le\sigma^2+\tau^2}.$$

Montrer que l'estimateur obtenu est minimax si

$$\frac{p-\alpha}{k-\beta-2} \le \frac{2(p-2)}{k+1}.$$

(*Indication :* Utiliser le Théorème 2.52.)

10.32 (Hartigan, 1983) Soient un modèle multinomial $\mathcal{M}_k(n;p_1,\ldots,p_k)$ et une observation $(n_1,\ldots,n_k)$. Une loi conjuguée a priori possible est la loi de Dirichlet $\mathcal{D}(\alpha_1,\ldots,\alpha_k)$.

a. Montrer que

$$\mathbb{E}\left[\sum_{i=1}^k n_i^2\right] = n + (n-1)\frac{\alpha+1}{k\alpha+1},$$

et déterminer quand l'équation des moments obtenue à partir de cette égalité a une solution positive. Donner un estimateur bayésien empirique de $(p_1,\ldots,p_k)$ dans ce cas.

b. Calculer un estimateur bayésien empirique en utilisant les estimateurs du maximum de vraisemblance des α_i. [*Note :* Consulter Good, 1975, pour plus de détails sur ce modèle.]

10.33 *(Morris, 1983a) Une famille exponentielle de densité

$$f(x|\theta) = h(x)e^{\theta x - \psi(\theta)}$$

a une *variance quadratique* si la variance peut s'écrire

$$V(\mu) = \psi''(\theta) = v_0 + v_1\mu + v_2\mu^2,$$

où $\mu = \psi'(\theta)$ est l'espérance de $f(x|\theta)$. Morris (1982) donne une caractérisation des six familles de variance quadratique (Exercice 3.9). Ces lois sont notées $NEF(\mu,V(\mu))$.

a. Montrer que la loi conjuguée de μ peut s'écrire

$$g(\mu) = K(m,\mu_0)e^{m\mu_0\theta(\mu)-m\psi(\theta(\mu))}V^{-1}(\mu) \tag{10.23}$$

et que

$$\mathbb{E}^\pi[\mu] = \mu_0, \qquad V^\pi(\mu) = \tau_0^2 = \frac{V(\mu_0)}{m-v_2},$$

Par conséquent, la loi conjuguée est aussi une famille exponentielle de variance quadratique. Établir une table de correspondances entre loi de l'échantillon et loi a priori conjuguée pour les six familles obtenues dans l'Exercice 3.24.

b. Montrer que l'estimateur de Bayes associé à (10.23) pour n observations indépendantes $x_1,\ldots,x_n$ et le coût quadratique est

$$\delta^\pi(x_1,\ldots,x_n) = (1-B)\bar{x} + B\mu_0,$$

où

$$B = \frac{V(\mu_0)+v_2\tau_0^2}{V(\mu_0)+(n+v_2)\tau_0^2}.$$

c. Montrer que, pour la loi conjuguée (10.23), les moments marginaux de $\bar{x}$ sont

$$\mathbb{E}[\bar{x}] = \mu_0, \qquad \text{var}(\bar{x}) = \frac{V(\mu_0)}{n}\frac{m+n}{m-v_2}.$$

d. Soient k observations indépendantes

$$x_i|\mu_i \sim NEF(\mu_i, V(\mu_i)/n) \qquad (1 \le i \le k),$$

de paramètres indépendants μ_i selon la loi conjuguée (10.23). Si $\bar{x} = \sum_i x_i/k$ et $s = \sum_i (x_i - \bar{x})^2$ et si

$$\frac{\mathbb{E}[V(\bar{x})(k-1)]}{\mathbb{E}[s]} = \mathbb{E}\left[\frac{V(\bar{x})(k-3)}{s}\right]$$

(les espérances étant prises sous la loi marginale), montrer qu'un estimateur bayésien empirique pour μ_i est

$$\delta_i^{\text{EB}}(x_1, \ldots, x_k) = (1 - \hat{B})x_i + \hat{B}\bar{x},$$

avec

$$\hat{B} = \min\left(\frac{v_2}{n+v_2}\frac{k-1}{k} + \frac{n}{n+v_2}\frac{(k-3)V(\bar{x})}{ns}, 1\right).$$

Section 10.5.1

10.34 Montrer que, pour la loi marginale de l'Exemple 10.32, $(p-2)/||x||^2$ est effectivement un estimateur sans biais de $1/(1+\tau^2)$.

10.35 Démontrer la formule (10.20) de l'Exemple 10.33.

10.36 *(Kubokawa, 1991) Soient $\delta^{\text{JS}}(x) = [1 - (p-2)/||x||^2]x$, l'estimateur de James-Stein, et $x \sim \mathcal{N}_p(\theta, I_p)$. On pose $\lambda = ||\theta||^2/2$; $f_p(t;\lambda)$ est la densité du khi deux décentrée avec paramètre de non-centralité λ.

a. Pour la troncature de δ^{JS}

$$\delta_1(x; c, r) = \begin{cases} \left(1 - \dfrac{c}{||x||^2}\right)x & \text{si } ||x||^2 < r, \\ \delta^{\text{JS}}(x) & \text{sinon,} \end{cases}$$

montrer que le risque quadratique de $\delta_1(x; c, r)$ est minimal pour

$$c_1(r, \lambda) = p - 2 - \frac{2f_p(r;\lambda)}{\int_0^r (1/t) f_p(t;\lambda)\, dt}.$$

b. On pose

$$c_1(r) = p - 2 - \frac{2}{\int_0^1 t^{p/2-2} e^{(1-t)r/2} dt}.$$

Montrer que $\delta_1(x; c_1(r), r)$ domine δ^{JS} pour tout r.

c. En utilisant un argument de limite, montrer que

$$\delta_1^*(x) = \left(1 - \frac{c_1(||x||^2)}{||x||^2}\right)x$$

domine δ^{JS}. [*Note :* Cet estimateur vient de Strawderman, 1971, et Berger, 1975a. Voir l'Exercice 10.28.]

d. Montrer que δ_1^* est admissible. (*Indication :* On pourra utiliser la condition suffisante du Théorème 8.13.)

10.37 *(Bock et Robert, 1985) On considère $x \sim \mathcal{N}_p(\theta, I_p)$ et $\theta \sim \mathcal{U}_{\{\|\theta\|^2=c\}}$, loi uniforme sur la sphère de rayon c. Proposer un estimateur bayésien empirique de θ en fonction de $||x||^2$ et montrer que, si cet estimateur découle de l'estimateur du maximum de vraisemblance de c, alors $\delta^{\mathrm{EB}}(x) = h(x)x$ avec

$$\left(1 - \frac{p}{||x||^2}\right)^+ \leq h(x) \leq \left(1 - \frac{p-1}{||x||^2}\right)^+.$$

Discuter la robustesse de l'effet Stein en termes de symétrie sphérique.

10.38 *(George, 1986a) Soit $y \sim \mathcal{N}_p(\theta, I_p)$. Cet exercice établit un estimateur qui choisit entre plusieurs partitions de y en vecteurs plus petits avant de rétrécir l'observation vers chacun de ces vecteurs. Pour $k = 1, \ldots, K$, notons

$$y = (y_{k1}, \ldots, y_{kJ_k})C_k \qquad \text{et} \qquad \theta = (\theta_{k1}, \ldots, \theta_{kJ_k})C_k$$

les partitions de y et θ en vecteurs y_{kj} et θ_{kj} de dimensions p_{kj} $(1 \leq j \leq J_k)$, avec C_k matrice de permutation contenant des 0 et des 1 avec un seul 1 pour chaque ligne et chaque colonne. Pour $k = 1, \ldots, K$, soit $\delta_k = (\delta_{k1}, \ldots, \delta_{kJ_k})C_k$ un estimateur de composantes

$$\delta_{kj}(y_{kj}) = y_{kj} + \nabla \log m_{kj}(y_{kj}),$$

où les fonctions m_{kj} de $\mathbb{R}^{p_{kj}}$ dans $\mathbb{R}$ sont deux fois différentiables. On pose également

$$m_k(y) = \prod_{j=1}^{J_k} m_{kj}(y_{kj}) \qquad \text{et} \qquad m_*(y) = \sum_{k=1}^{K} \omega_k m_k(y),$$

pour $\omega_k \geq 0$ $(1 \leq k \leq K)$ et $\sum_k \omega_k = 1$.

a. Si π_{kj} est une loi a priori sur θ_{kj} et si m_{kj} est la loi marginale correspondante sur y_{kj} $(1 \leq k \leq K,\ 1 \leq j \leq J_k)$, montrer que m_k est la loi marginale de y pour la loi a priori

$$\pi_k(\theta) = \prod_{j=1}^{J_k} \pi_{kj}(\theta_{kj})\,,$$

et que δ_k est la moyenne a posteriori de cette loi.

b. En déduire que

$$\delta^*(y) = y + \nabla \log m_*(y)$$

est l'estimateur de Bayes pour la loi a priori

$$\pi^*(\theta) = \sum_{k=1}^{K} \omega_k \pi_k(\theta).$$

c. Montrer que δ^* peut également s'écrire sous la forme

$$\delta^*(y) = \sum_{k=1}^{K} \varrho_k(y)\delta_k(y),$$

avec $\varrho_k(y) = \omega_k m_k(y)/m_*(y)$, et interpréter ce résultat.

d. Montrer que si, pour $k = 1, \ldots, K$,

$$\mathbb{E}_\theta \left| \frac{\partial^2 m_k(y)}{\partial y_i^2} \Big/ m_k(y) \right| < +\infty,$$
$$\mathbb{E}_\theta ||\nabla \log m_k(y)||^2 < +\infty,$$

alors l'estimateur sans biais du risque de δ^* peut s'écrire

$$\mathscr{D}\delta^*(y) = p - \sum_{k=1}^{K} \varrho_k(y) \left[\mathscr{D}\delta_k(y) - (1/2) \sum_{\ell=1}^{K} \varrho_\ell(y) ||\delta_k(y) - \delta_\ell(y)||^2 \right],$$

avec

$$\mathscr{D}\delta_k(y) = ||\nabla \log m_k(y)||^2 - 2\Delta m_k(y)/m_k(y).$$

(*Indication :* Utiliser le Lemme 10.21 avec $Q = \Sigma = I_p$.)

e. En déduire que, si m_{kj} est *surharmonique*, c'est-à-dire tel que $\Delta m_{kj}(y_{kj}) \leq 0$ pour $1 \leq k \leq K$, $1 \leq j \leq J_k$, δ^* est minimax. [*Note :* Ce résultat peut être décrit par l'assertion qu'une combinaison convexe "propre" d'estimateurs minimax est minimax.]

f. Pour $1 \leq k \leq K$, $1 \leq j \leq J_k$, on note V_{kj} un sous-espace de $\mathbb{R}^{p_{kj}}$, avec $\dim V_{kj} = p_{kj} - q_{kj}$ et $q_{kj} \geq 3$; P_{kj} est le projecteur orthogonal associé de $\mathbb{R}^{p_{jk}}$ sur V_{kj} et $s_{kj} = ||y_{kj} - P_{kj} y_{kj}||^2$. Donner les estimateurs à rétrécisseur multiple δ^* associés à

$$m_{kj}(y_{kj}) = \begin{cases} \left(\dfrac{q_{kj} - 2}{e s_{kj}} \right)^{(q_{kj}-2)/2} & \text{si } s_{kj} \geq q_{kj} - 2, \\ \exp(-s_{kj}/2) & \text{sinon.} \end{cases}$$

(*Indication :* La solution est l'estimateur tronqué de James-Stein.)

Section 10.5.2

10.39 *(Kubokawa *et al.*, 1993a) Soient $x \sim \mathscr{N}_p(\theta, \sigma^2 I_p)$, $y \sim \mathscr{N}_q(\xi, \sigma^2 I_q)$, et $s \sim \sigma^2 \chi^2_n$, avec θ, ξ et σ inconnus. Un estimateur bayésien empirique de θ est l'estimateur de James-Stein

$$\delta^{\mathrm{JS}}(x, s) = \left(1 - \frac{(p-2)s}{(n+2)||x||^2} \right) x.$$

Le but de cet exercice est de montrer que le remplacement de s par un estimateur plus efficace de σ^2 peut conduire à une amélioration de l'estimation de θ.

a. Montrer que, si $\gamma_h(y, s) = sh(||y||^2/s)$ domine $\gamma_0(s) = s/(n+2)$ sous coût quadratique invariant

$$\mathrm{L}(\sigma^2, \gamma) = \left(\frac{\gamma}{\sigma^2} - 1 \right)^2,$$

δ^{JS} est dominé par

$$\hat{\delta}(x, y, s) = \left(1 - \frac{(p-2)\gamma(y, s)}{||x||^2} \right) x$$

sous coût quadratique. (*Indication :* On rappelle que γ_0 est le meilleur estimateur équivariant de σ^2.)

b. Soit
$$\delta_g(x,y,s) = \left(1 - \frac{(p-2)s}{||x||^2} g(||y||^2/s, ||x||^2/s)\right) x.$$
On définit
$$g^*(u,v) = \min\left(g(u,v), \frac{1+u+v}{p+q+n}\right),$$
et supposons que g et g^* sont des fonctions absolument continues de v. Montrer que, si
$$\mathbb{E}\left[\frac{\partial g^*(U,V)}{\partial v} - \frac{\partial g(U,V)}{\partial v}\right] \geq 0,$$
lorsque $U = ||y||^2/s$ et $V = ||x||^2/s$, δ_{g^*} domine δ_g.

c. En déduire que
$$\delta_2(x,y,s) = x - \frac{p-2}{||x||^2} \ \min\left\{\frac{s}{n+2}, \frac{s+||y||^2}{n+q+2}, \frac{s+||x||^2+||y||^2}{n+p+q+2}\right\} \ x$$
domine δ^{JS}.

Section 10.5.3

10.40 *(Casella et Hwang, 1983) Soit $x \sim \mathcal{N}_p(\theta, I_p)$. Sous le coût linéaire
$$\mathrm{L}(\theta, C) = k \ \mathrm{vol}(C) - \mathbb{I}_C(\theta),$$
on rappelle que les estimateurs de Bayes sont des régions HPD de la forme $\{\theta;\ \pi(\theta|x) \geq k\}$ quand $\pi(\{\theta; \pi(\theta|x) = k\}) = 0$. De plus, si
$$k = k_0 = e^{-c^2/2}/(2\pi)^{p/2},$$
Joshi (1967a) montre que la région usuelle
$$C_x^0 = \{\theta;\ ||\theta - x|| \leq c\},$$
est minimax.

a. Montrer que, si $\theta \sim \mathcal{N}_p(0, \tau^2 I_p)$, l'ensemble de Bayes est
$$C_x^\pi = \left\{\theta;\ ||\theta - \delta^\pi(x)||^2 \leq -\frac{2\tau^2}{\tau^2+1} \log\left[k\left(\frac{2\pi\tau^2}{\tau^2+1}\right)^{p/2}\right]\right\},$$
où $\delta^\pi(x) = (\tau^2/\tau^2+1)x$ est l'estimateur de Bayes de θ. Pour $k = k_0$, montrer que cet ensemble peut s'écrire
$$C_x^\pi = \left\{\theta;\ ||\theta - \delta^\pi(x)||^2 \leq \frac{\tau^2}{\tau^2+1}\left[c^2 - p\log\left(\frac{\tau^2}{\tau^2+1}\right)\right]\right\}.$$

b. En déduire qu'un ensemble de Bayes empirique simple est
$$C_x^{\mathrm{EB}} = \left\{\theta;\ \|\theta - \delta^{\mathrm{EB}}(x)\|^2 \leq v^{\mathrm{EB}}(x)\right\},$$
avec $\delta^{\mathrm{EB}}(x) = (1 - [(p-2)/||x||^2])x$ et
$$v^{\mathrm{EB}}(x) = \left(1 - \frac{p-2}{||x||^2}\right)\left(c^2 - p\log\left|1 - \frac{p-2}{||x||^2}\right|\right).$$

c. Expliquer pourquoi il est préférable de considérer

$$\delta^+(x) = \left(1 - \frac{p-2}{||x||^2}\right)^+ x$$

et

$$v^e(x) = \begin{cases} \left(1 - \frac{p-2}{c^2}\right)\left(c^2 - p\log\left[1 - \frac{p-2}{c^2}\right]\right) & \text{si } ||x||^2 < c, \\ \left(1 - \frac{p-2}{||x||^2}\right)\left(c^2 - p\log\left[1 - \frac{p-2}{||x||^2}\right]\right) & \text{sinon.} \end{cases}$$

d. Généraliser au cas où $x \sim \mathcal{N}_p(\theta, \sigma^2 I_p)$ et $s^2 \sim \sigma^2 \chi^2_q$.

Note 10.7.2

10.41 Dans le cadre de l'Exemple 10.38,

a. Montrer que

$$\mathrm{L}(\eta, \hat{\eta}) = \frac{p}{2}\log\left(\frac{\eta}{\hat{\eta}}\right) + \frac{1}{2}\left(\frac{1}{\hat{\eta}} - \frac{1}{\eta}\right) p\,\eta\,,$$

et en déduire (10.26).

b. Montrer que la loi a posteriori associée à π_d est bien définie pour $d > (4-p)/2$.

c. Prouver (10.27) et (10.28).

d. Déduire de l'approximation

$${}_1F_1(a,b,z) = \Gamma(b)\, z^{-a}\left\{1 - \frac{a(1+a-b)}{z} + \mathrm{O}(z^2)\right\}$$

l'équivalence (10.29).

10.42 *(**Suite de l'Exercice 10.41**) En utilisant les conditions STUB (Section 8.5) et l'approximation (10.29), montrer que le choix $d = 1$ est optimal.

10.43 *(**Suite de l'Exercice 10.41**) Déduire d'Alam (1973) la minimaxité de δ_1^{EB}. (*Indication :* Voir l'Exemple 10.26.)

10.44 Dans le cadre de l'Exemple 10.39,

a. Montrer que le coût entropique est donné par (10.30) et en déduire la forme générale des estimateurs de Bayes sous ce coût.

b. Montrer que, lorsque $\pi(\lambda) = \lambda^{-d}$, la distribution a posteriori s'écrit

$$\pi_d(\lambda|x) \propto \lambda^{n-d}(\lambda+1)^{-\sum_i x_i - n}$$

et en déduire que l'estimateur de Bayes de λ est donné par (10.31).

c. Montrer que la distribution conditionnelle $\pi(\theta|x,\lambda) \propto \theta^x e^{-(\lambda+1)\theta}$ et en déduire que l'estimateur de Bayes de θ_i conditionnellement à λ est

$$\mathbb{E}[\theta_|x_i, \lambda] = \frac{x_i + 1}{\lambda + 1}\,.$$

10.45 (Suite de l'Exercice 10.44) Montrer que l'estimateur classique de θ, $\gamma_0(x) = x$, a un risque infini sous coût entropique.

10.46 (Suite de l'Exercice 10.44) Montrer que l'estimateur de Bayes associé à l'a priori intégré

$$\pi(\theta) = \int \pi(\theta, \lambda) d\lambda$$

est donné par

$$\delta^{\pi}(x) = \left(1 - \frac{n - d + 1}{\sum_{i=1}^{n} x_i + n}\right)(x + \mathbf{1})\,.$$

Déduire des différences sur les risques que l'estimateur de Bayes domine son équivalent empirique sous coût entropique.

10.7 Notes

10.7.1 Modèles graphiques[79]

Les modèles graphiques sont des modèles statistiques où interviennent des structures de graphes (au sens de la théorie mathématique des graphes). Ils ont été développés essentiellement pour représenter les relations d'indépendance conditionnelle, d'abord dans le domaine des *systèmes experts* (Whittaker, 1990, Spiegelhalter et Lauritzen, 1990, Spiegelhalter et Cowell, 1992, Spiegelhalter *et al.*, 1993). L'application de l'approche bayésienne à ces modèles, qui a permis d'intégrer l'incertitude de modélisation, a été grandement facilitée par les progrès des techniques MCMC, comme le montrent Madigan et York (1995) dans un article introductif à partir duquel la présente note a été écrite.

La construction d'un modèle graphique est établie à partir d'une collection d'hypothèses d'indépendance représentées par un graphe. Nous rappelons ici quelques points essentiels de théorie des graphes et renvoyons à Lauritzen (1996) pour plus de détails. Un *graphe* est défini par un ensemble de *sommets* ou *nœuds*, $\alpha \in \mathscr{V}$, qui représente les variables aléatoires ou facteurs d'étude, et par un ensemble d'*arêtes*, $(\alpha, \beta) \in \mathcal{V}^2$, qui peuvent être associées à une direction (le graphe est alors dit *orienté*) ou non (le graphe est *non orienté*) et représentent les liens de dépendance entre les variables. Dans un graphe non orienté, les variables α et β sont reliées par une arête si, conditionnellement à toutes les autres variables, elles *ne* sont *pas* indépendantes. Dans un graphe orienté, α est un *parent* de β si (α, β) est une arête (et β est alors un *fils* de α)[80]. Une hypothèse usuelle sur les graphes est de les supposer *acycliques*, c'est-à-dire sans chemin orienté reliant un nœud α à lui-même. On obtient alors la notion de *graphes acycliques orientés* définie par Kiiveri et Speed (1982) et souvent représentée par l'acronyme DAG.

Dans l'optique de construction de modèles probabilistes sur les graphes, une notion importante est celle de *clique*. Une *clique* C est un sous-ensemble maximal de nœuds tous connectés deux à deux (maximal au sens où il n'existe pas de sous-ensemble contenant C et vérifiant cette condition). Un ordre des cliques d'un graphe non orienté $(C_1, \ldots, C_n)$ est dit *parfait* si les nœuds de chaque

[79] Cette note est fortement inspirée de la Note 7.6.6 de Robert et Casella, 1990.

[80] Les graphes orientés peuvent être transformés en graphes non orientés en ajoutant des liens entre les nœuds qui ont un fils commun et en ignorant les directions.

clique C_i qui apparaissaient dans une clique antérieure sont tous membres de la même clique précédente (ces nœuds sont appelés les *séparateurs*, $\alpha \in S_i$). Dans ce cas, la densité jointe de la variable aléatoire V à valeurs dans $\mathscr{V}$ est

$$p(V) = \prod_{v \in V} p(v|\mathscr{P}(v)) \,,$$

où $\mathscr{P}(v)$ désigne les parents de v. Ceci peut également s'écrire

$$p(V) = \frac{\prod_{i=1}^{n} p(C_i)}{\prod_{i=1}^{n} p(S_i)}, \tag{10.24}$$

et le modèle est alors dit *décomposable*; voir Spiegelhalter et Lauritzen (1990), Dawid et Lauritzen (1993) ou Lauritzen (1996). Comme le soulignent Spiegelhalter *et al.* (1993), la représentation (10.24) induit un *principe de traitement local*, qui permet de construire une loi a priori ou de simuler selon une loi conditionnelle à partir d'une seule clique. (Autrement dit, la loi est de *Markov par rapport au graphe non orienté*, ce que montrent Dawid et Lauritzen, 1993.) L'intérêt de cette propriété apparaît alors clairement dans le cadre d'une implémentation de l'échantillonnage de Gibbs.

Lorsque les densités ou probabilités sont paramétrées, les paramètres sont notés θ_A pour la loi marginale de $V \in A$, $A \subset \mathscr{V}$. (Dans le cas des modèles discrets, $\theta = \theta_V$ peut coïncider avec p lui-même; voir l'Exemple 10.36.) La loi a priori $\pi(\theta)$ doit alors être compatible avec la structure du graphe : Dawid et Lauritzen (1993) montrent qu'il existe une solution de la forme

$$\pi(\theta) = \frac{\prod_{i=1}^{n} \pi_i(\theta_{C_i})}{\prod_{i=1}^{n} \tilde{\pi}_i(\theta_{S_i})} \,, \tag{10.25}$$

reproduisant ainsi la décomposition en cliques (10.24).

Exemple 10.36. On considère un graphe décomposable tel que les variables aléatoires correspondant à tous les nœuds de $\mathscr{V}$ soient discrètes. Soient $w \in W$ une valeur possible pour le vecteur de ces variables aléatoires et $\theta(w)$ la probabilité associée. Pour la décomposition parfaite en cliques $(C_1, \ldots, C_n)$, on note $\theta(w_i)$ la probabilité marginale que le vecteur inclus $(v, v \in C_i)$ prenne la valeur w_i $(\in W_i)$ et, de façon similaire, $\theta(w_i^s)$ est la probabilité que le vecteur inclus $(v, v \in S_i)$ prenne la valeur w_i^s en notant $(S_1, \ldots, S_n)$ la suite de séparateurs correspondante. Dans ce cas,

$$\theta(w) = \frac{\prod_{i=1}^{n} \theta(w_i)}{\prod_{i=1}^{n} \theta(w_i^s)} \,.$$

Comme le montrent Madigan et York (1995), une loi a priori de Dirichlet peut être construite sur $\theta_W = (\theta(w), w \in W)$. Il induit de vraies lois a priori de Dirichlet sur les $\theta_{W_i} = (\theta(w_i), w_i \in W_i)$, sous la contrainte que les poids de Dirichlet soient identiques à l'intersection de deux cliques. Dawid et Lauritzen (1993) prouvent que cette loi a priori est unique, si les deux lois a priori marginales sur les cliques sont données. ‖

Exemple 10.37. Giudici et Green (1999) donnent un autre exemple de spécification a priori dans le cas d'un *modèle graphique gaussien*, $\mathbf{X} \sim \mathcal{N}_p(0, \Sigma)$, la matrice de précision $K = \{k_{ij}\} = \Sigma^{-1}$ devant être compatible avec les relations d'indépendance conditionnelles du graphe. Par exemple, si $\mathbf{X}_v$ et $\mathbf{X}_w$ sont indépendants sachant le reste du graphe, alors $k_{vw} = 0$. La vraisemblance peut alors être factorisée en

$$f(\mathbf{x}|\Sigma) = \frac{\prod_{i=1}^{n} f(\mathbf{x}_{C_i}|\Sigma^{C_i})}{\prod_{i=1}^{n} f(\mathbf{x}_{S_i}|\Sigma^{S_i})},$$

avec les mêmes notations pour les cliques et les séparateurs que ci-dessus et avec $f(\mathbf{x}_C|\Sigma^C)$ densité normale $\mathcal{N}_{p_C}(0, \Sigma^C)$, d'après (10.24). L'a priori sur Σ peut être défini comme les lois a priori inverses conjuguées de Wishart sur les Σ^{C_i}, sous certaines conditions de compatibilité. ‖

Madigan et York (1995) utilisent ce cadre pour proposer une approche MCMC à la sélection de modèles et à la moyennisation de modèles. Dellaportas et Forster (1996) et Giudici et Green (1999) implémentent des algorithmes à sauts réversibles pour déterminer la structure de graphe la plus probable associée à un ensemble de données, les seconds le faisant sous une hypothèse gaussienne.

10.7.2 Approche bayésienne empirique

Comme nous l'avons évoqué dans la Section 10.4, la difficulté de l'approche bayésienne empirique est qu'elle effectue l'estimation en deux étapes, la première consistant à estimer l'hyperparamètre à partir de lois marginales et la seconde à estimer le paramètre à partir du pseudo-a priori où l'hyperparamètre est remplacé par son estimation. Bien que l'inefficacité de cette procédure–comparée à une méthode vraiment bayésienne–ne puisse être complètement levée, il semble logique d'utiliser la technique d'estimation la plus efficace possible à la première étape, à savoir une approche bayésienne non informative. Puisque le premier niveau d'estimation n'est pas induit par un problème de décision, il est très probable qu'aucune fonction de coût ne soit disponible. Les coûts intrinsèques présentés en Section 2.5.4 font alors figure d'option par défaut naturelle dans ce cas. De manière surprenante, cette solution, qui permet pourtant de s'affranchir de l'arbitraire lié à l'estimation des hyperparamètres du bayésien empirique, n'est pas utilisée dans la littérature. Les deux exemples ci-dessous viennent de Fourdrinier et Robert (1995).

Exemple 10.38. (Suite de l'Exemple 10.32) La distribution marginale de x, $m(x|\eta)$, est $\mathcal{N}_p(0, \eta\, I_p)$, avec $\eta = 1 + \tau^2$, et le coût entropique correspondant pour l'estimation de η est

$$\begin{aligned} \mathrm{L}(\eta, \hat{\eta}) &= \int \log\left(\frac{m(x|\eta)}{m(x|\hat{\eta})}\right) m(x|\eta)dx \\ &= \frac{p}{2}\left(\frac{\eta}{\hat{\eta}} - \log\left(\frac{\eta}{\hat{\eta}}\right) - 1\right). \end{aligned} \tag{10.26}$$

Puisque η est un paramètre d'échelle pour la distribution marginale, une famille de lois a priori non informatives naturelle est $\pi_d(\eta) = \eta^{-d}$ sur $[1, \infty)$ et l'estimateur correspondant de η est

$$\hat{\eta}_d = \frac{\int_0^1 \nu^{(p/2)+d-3} e^{-||x||^2\nu/2} d\nu}{\int_0^1 \nu^{(p/2)+d-2} e^{-||x||^2\nu/2} d\nu}. \tag{10.27}$$

L'estimateur bayésien empirique est alors

$$\begin{aligned} \delta_d^{\mathrm{EB}}(x) &= (1 - \hat{\eta}^{-1})x \\ &= \frac{\int_0^1 \nu^{(p/2)+d-3}(1-\nu) e^{-||x||^2\nu/2} d\nu}{\int_0^1 \nu^{(p/2)+d-2} e^{-||x||^2\nu/2} d\nu}\, x \\ &= \frac{2}{p+2d-2}\, \frac{{}_1F_1(2, d+p/2, ||x||^2/2)}{{}_1F_1(2, d+p/2-1, ||x||^2/2)}\, x\,, \end{aligned} \tag{10.28}$$

où ${}_1F_1$ est la fonction confluente hypergéométrique (Abramowitz et Stegun, 1964, Chapitre 13). Puisque $\delta_d^{\mathrm{EB}}(x)$ est asymptotiquement équivalent à

$$\left(1 - \frac{p + 2(d-2)}{||x||^2}\right) x\,, \tag{10.29}$$

le choix $d = 1$, c'est-à-dire $\pi(\eta) = 1/\eta$, est le choix optimal de d (Exercice 10.42). ‖

Exemple 10.39. (Suite de l'Exemple 10.29) Le coût entropique associé à $m(x|\lambda)$ est

$$\mathrm{L}(\lambda, \hat{\lambda}) = \log\left(\frac{\lambda}{\hat{\lambda}}\right) + \left(1 + \frac{1}{\lambda}\right) \log\left(\frac{\hat{\lambda}+1}{\lambda+1}\right) \tag{10.30}$$

et, pour $\pi(\lambda) = \lambda^{-d}$, l'estimateur de Bayes correspondant de λ est

$$\hat{\lambda} = \frac{n-d}{\sum_{i=1}^n x_i + n - 1}. \tag{10.31}$$

En utilisant également un coût entropique pour l'estimation de λ, $\mathrm{L}(\theta, \hat{\theta}) = \hat{\theta} - \theta - \log(\hat{\theta}/\theta)$, l'estimateur bayésien empirique de $\theta = (\theta_1, \ldots, \theta_n)$ est

$$\theta^{\mathrm{EB}}(x) = \left(1 - \frac{n-d}{\sum_{i=1}^n x_i + n - 1}\right)(x + \mathbf{1})\,, \tag{10.32}$$

avec $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^n$. Fourdrinier et Robert (1995) montrent en outre qu'il existe un choix optimal d^* de d pour le coût entropique, avec $d^* \leq 2$ et que, sur un intervalle de valeurs donné de d, θ^{EB} domine $\hat{\theta}_0 = x + \mathbf{1}$. ‖

11

Une défense du choix bayésien

> "A series of steps, each taken for good cause or pure necessity, each seeming so reasonable at the time, and each leading to things he had never imagined. He always seemed to find himself caught in that sort of dance."
>
> **Robert Jordan**, *Lord of Chaos.*

Ce livre a introduit les aspects les plus importants de l'approche bayésienne, principalement sous l'angle de la Théorie de la Décision. Il ne s'agit évidemment pas d'une couverture exhaustive : d'une part, un traitement approfondi de la plupart des notions abordées ici est possible et d'une certaine manière souhaitable. D'autre part, il existe un nombre considérable d'applications de l'analyse bayésienne, et qui va croissant en particulier du fait des nouvelles possibilités offertes par les méthodes de calcul présentées au Chapitre 6. Nous pouvons ainsi mentionner les *Biostatistiques* (voir, par exemple, Berry et Stangl, 1996) ; l'*Économétrie* (Zellner, 1971, 1984, Box et Tiao, 1973, Poirier, 1995, Bauwens *et al.*, 1999, ou Geweke, 1999) ; l'*Environnemétrie* (Parent *et al.*, 1998) ; les *systèmes experts* (Gilks *et al.*, 1993, Cowell *et al.*, 1999) ; la *Finance* (Jacquier *et al.*, 1994, Pitt et Shephard, 1999) ; le *traitement d'image et la reconnaissance d'objets* (Geman et Geman, 1984, Besag, 1986 ou Fitzgerald *et al.*, 1999) ; les *réseaux de neurones* (Ripley, 1992, Neal, 1999), le *traitement du signal* (Andrieu et Doucet, 1999, Andrieu *et al.*, 2000) ; les *réseaux bayésiens* (Chickering et Heckerman, 2000, Kontkanen *et al.*, 2000). (On pourra aussi

consulter la revue par Berger, 2000, ainsi que Gatsonis *et al.*, 1993, 1995, 1997, 1999, Gilks *et al.*, 1996 et Carlin et Louis, 2000a pour des références additionnelles.) Des ouvrages consacrés aux applications de l'analyse bayésienne sont aussi disponibles, comme par exemple Pole *et al.* (1994), Congdon (2001, 2003), Gill (2002) ou Holmes *et al.* (2002).

Par conséquent, il nous semble utile, à ce stade du livre, de replacer l'analyse bayésienne dans cette approche théorique et décisionnelle plutôt que de commencer à l'illustrer dans quelques applications choisies. Premièrement, ce positionnement permet d'appréhender la cohérence de l'analyse bayésienne, en soi et par rapport à d'autres théories statistiques. Deuxièmement, il n'est certainement pas inutile de bien posséder les bases théoriques d'une méthodologie pour pouvoir l'appliquer efficacement dans *toutes* les situations réelles. C'est pourquoi ce dernier chapitre contient une justification globale de l'approche bayésienne qui résume et parfois étend les arguments développés jusqu'à présent[81]. Le ton de ce chapitre est donc vaguement philosophique, plutôt que méthodologique (ou mathématique), et les lecteurs pourront juger si le sentiment qu'ils retirent de la lecture de cet ouvrage coïncide avec la perspective défendue ci-après.

(1) **Le choix d'une représentation probabiliste**

Le fait de proposer une loi sur les paramètres inconnus d'un modèle statistique est en quelque sorte une *probabilisation de l'incertain*. Nous voulons traduire par ce néologisme une réduction axiomatique de la notion d'inconnu à la notion d'aléatoire. Cette réduction étant acceptable–et elle l'est en général pour la quasi-totalité des statisticiens–pour les modèles d'échantillonnage, elle devrait l'être tout autant pour les paramètres qui dirigent ces modèles. En particulier, la distinction entre échantillon et paramètres n'est jamais absolue. Il suffit par exemple de considérer le cas des modèles à effets aléatoires (Chapitre 10) ou celui des vecteurs d'allocation dans un modèle de mélange (Chapitre 6).

Plus fondamentalement, un modèle probabiliste n'est souvent qu'une *interprétation* d'un certain phénomène–et non pas une *explication*. Si on prend, par exemple, le cas des modèles *économétriques*, où les différences entre les réalisations des variables *endogènes* et leur prévision linéaire par rapport aux variables *exogènes* sont expliquées par une perturbation aléatoire, il est évident que la nature aléatoire de cette différence est de peu d'importance, ne serait-ce que parce que l'expérience ne peut être

[81] La présentation de ce chapitre est très différente de celle des chapitres précédents, en ce qu'il ne contient ni théorème ni exemple, mais simplement une suite de points (de (1) à (10)) qui sont des argumentations en faveur de l'approche bayésienne. Ces points sont suivis de courtes réfutations des critiques les plus usuelles avancées contre l'analyse bayésienne en Statistique.

reproduite[82]. Par conséquent, la représentation de phénomènes inconnus par un modèle probabiliste, au niveau des observations aussi bien qu'au niveau des paramètres, n'a pas besoin de correspondre effectivement–ou physiquement–à une génération issue d'une loi de probabilité, ni même de nous obliger à accepter un schéma superdéterministe pour les phénomènes analysés. Fondamentalement, l'absence de répétabilité de la plupart des expériences ou collectes de données vient effacer en grande partie la frontière entre aléatoire et non aléatoire.

En fait, les représentations probabilistes de phénomènes partiellement expliqués devraient surtout être comprises comme un *outil* simplificateur mais efficace permettant l'analyse de ces phénomènes (voir le point (4) ci-dessous). Cette perspective est similaire à la manière dont la Physique peut aussi être vue comme une interprétation du monde, donc comme un *outil*, suffisamment efficace pour permettre une meilleur compréhension de l'univers (et, incidemment, la perpétuation du progrès technique), sans qu'on ait à défendre l'existence d'une "vérité" de toute manière inatteignable[83].

(2) **Conditionner par rapport aux données**

La base de l'inférence statistique est fondamentalement un processus d'*inversion*, puisqu'elle cherche à déduire les causes des effets, en prenant en compte la nature probabiliste du modèle et l'influence de facteurs complètement aléatoires (c'est-à-dire non intégrés dans l'analyse). Dans ses versions discrètes comme dans ses versions continues, le théorème de Bayes formalise cette inversion, comme le fait aussi la notion de vraisemblance $\ell(\theta|x)$, qui remplace la densité $f(x|\theta)$. L'échec de la *Statistique fiduciaire* à fournir un système inférenciel satisfaisant (voir la Note 1.8.1) peut en fait être associé à un refus (par cette théorie) de poursuivre cette inversion jusqu'au bout de ses conséquences logiques et, corrélativement, à une perpétuation de la confusion entre observations et variables aléatoires.

D'un point de vue probabiliste, si une analyse quantitative sur les paramètres θ est opérée *conditionnellement* à x, elle demande nécessairement une distribution sur les paramètres θ, $\pi(\theta)$, pour pouvoir inverser les lois. Si on intègre cette contrainte, l'approche bayésienne est la seule axiomatique *cohérente* qui respecte la perspective d'inversion des probabilités. La difficulté pratique de la détermination de la loi a priori π n'apparaît pas au même niveau conceptuel (voir le point (ii) ci-dessous).

(3) **Construire une véritable vraisemblance**

Continuant l'argumentation des points (1) et (2) ci-dessus, nous pouvons également faire remarquer qu'une modélisation a priori sur les paramètres

[82] En d'autres termes, un nombre arbitraire peut toujours être perçu comme une *unique* réalisation d'une infinité de distributions !

[83] Voir aussi Popper (1983) pour sa justification alternative de la modélisation scientifique via le *réalisme métaphysique* qu'il oppose en fait à cette approche *instrumentale*.

du modèle autorise une approche inférentielle *complète* sur ces paramètres, donc la détermination d'une véritable vraisemblance de θ conditionnellement aux observations x. Par comparaison, les approches classiques en Statistique n'aboutissent pas à cette complétude. En particulier, tant que θ est pris comme *inconnu* mais *fixe*, on ne peut pas donner un sens probabiliste précis à la fonction de vraisemblance.

Cette impossibilité qu'a l'analyse classique à conduire à des conclusions quantitatives justifiées est par exemple illustrée par le cas des régions de confiance et des tests, puisque cette analyse propose une problématique inappropriée (et donc une réponse inappropriée). Comme le décrit le Chapitre 5, les procédures classiques, que ce soient un intervalle de confiance à 95% ou une p-value, tirent leur nature probabiliste d'une analyse fréquentiste du problème. De leur point de vue, ce n'est plus le paramètre θ qui appartient à un intervalle donné avec probabilité 95% conditionnellement à x, mais plutôt l'intervalle déduit de x qui contient la valeur (fixe mais inconnue) de θ avec probabilité 0.95. De nouveau, le manque de répétabilité de la plupart des expériences invalide fortement ce point de vue fréquentiste (voir aussi le point (9) ci-dessous).

(4) **Voir les lois a priori comme outils ou résumés**

Le choix d'une loi a priori π ne nécessite pas un quelconque degré de *croyance* en cette distribution. Il est en fait plutôt rare de disposer d'une loi a priori complètement spécifiée, le cas de la boule de billard de Thomas Bayes étant, paradoxalement, un contre-exemple exceptionnel où la construction de l'expérience détermine la loi a priori. D'un point de vue plus général, π doit être considéré soit comme un *outil* qui fournit une procédure inférentielle unificatrice qui possède des propriétés fréquentistes acceptables (voir les points (6) et (8)), soit comme une manière *de résumer* l'information a priori disponible ainsi que l'incertitude qui l'entoure. Que l'analyse bayésienne puisse s'étendre à des contextes *non informatifs*–avec quelques difficultés parfois, comme dans le cas des tests–est en fait la preuve de cette polyvalence. De plus, que de nombreux estimateurs standard puissent se représenter via une modélisation non informative montre bien que l'utilisation d'une loi a priori n'implique pas toujours un biais dans le processus statistique mais, au contraire, qu'elle autorise en sus le traitement quantitatif mentionné au point (3). En fait, ces coïncidences sont des arguments supplémentaires en faveur de la validité de l'approche bayésienne, puisqu'elle fournit un champ inférentiel incluant aussi les estimateurs classiques.

(5) **Intégrer les bases subjectives de la connaissance**

D'un point de vue plus philosophique, il est globalement accepté que la connaissance procède de la confrontation entre ses a priori et des expériences, voulues ou subies. Par exemple, selon Kant, *bien que la connaissance débute avec l'expérimentation, il ne s'ensuit pas que la connaissance soit entièrement déduite de l'expérimentation.* En effet, sans

a priori, ce qui signifie ici, sans une structure préétablie du monde, l'observation n'a pas de sens, car elle ne procède plus d'une confirmation ou d'une confrontation à un modèle de référence. Par conséquent, la construction de la connaissance par l'expérimentation n'est possible que par l'existence d'un système de représentation a priori, évidemment primitif à l'origine, qui se trouve progressivement enrichi et actualisé au travers de ces expériences successives. Dans cette perspective, *l'apprentissage* est perçu comme le réexamen critique de systèmes de référence préexistants à la lumière d'expériences successives.

Ce point de vue est aussi partagé par Poincaré (1902) :

> *On dit souvent qu'il faut expérimenter sans idée préconçue. Cela n'est pas possible ; non seulement ce serait rendre toute expérience stérile, mais on le voudrait qu'on ne le pourrait pas. Chacun porte en soi sa conception du monde dont il ne peut se défaire si aisément.*

L'approche bayésienne est, bien entendu, en concordance avec cette perspective, puisque les distributions a priori sont le plus souvent fondées sur les résultats d'expériences antérieures. En fait, même l'aspect *subjectif* du choix de la loi a priori peut être assimilé par cette théorie de la connaissance, puisqu'elle implique que chaque acquisition de connaissance est essentiellement *subjective*, résultant d'une interaction entre les perceptions individuelles et la réalité extérieure[84].

Dans sa théorie (radicale) de l'Épistémologie, Feyerabend (1975) soutient que l'individualisme (qu'on peut aussi traduire par *subjectivité*) est un facteur important, mais totalement passé sous silence, des découvertes scientifiques. Bien qu'il s'oppose fortement à cette vision subjective de la connaissance, Popper (1983) reconnaît également le rôle des intuitions a priori (qu'il appelle *systèmes*), même si elles ne sont pas toujours fondées sur l'expérience, dans l'histoire des Sciences–l'exemple le plus frappant étant, de son point de vue, l'*atomisme*, c'est-à-dire la représentation de la matière comme étant formée d'atomes, théorie qui mit plus de vingt siècles avant d'être vérifiée expérimentalement.

(6) **Choisir un système inférentiel unique et cohérent**

Le but ultime de la Statistique est, sans conteste, de conduire à une *inférence* sur un paramètre θ à partir d'observations x reliées à θ via une loi de probabilité $f(x|\theta)$. De plus, il semble raisonnable de rechercher l'*efficacité* (voire l'*optimalité*) dans cette démarche inférentielle, cette notion d'optimalité pouvant être définie *explicitement* par le statisticien (ou le décideur). Forcer l'inférence à se couler dans un moule décisionnel par le

[84] En exagérant un peu, on pourrait soutenir que la Statistique bayésienne répond à ce souhait de Kant, exprimé dans l'introduction à sa *Critique de la raison pure* : *"La Philosophie a besoin d'une science qui détermine la possibilité, les principes et l'étendue de toutes nos connaissances a priori."*

choix d'une *fonction de coût* conduit à une clarification (nécessaire) de la manière dont les outils inférentiels sont évalués, donc force le statisticien ou le client à révéler ses préférences. En sus, quand le cadre décisionnel se trouve complété par le choix de la loi a priori, les buts et desiderata inférentiels mentionnés ci-dessus sont automatiquement remplis, puisque l'approche bayésienne offre très généralement une *unique procédure*, qui dépend bien sûr des propriétés du coût et de la connaissance a priori. Bien entendu, l'*unicité* d'une procédure décisionnelle n'est pas un argument *en soi*, puisque de nombreuses méthodologies, si insignifiantes soient-elles, peuvent également posséder cette propriété.

Par conséquent, la plus importante caractéristique d'une démarche bayésienne est que les estimateurs de Bayes sont construits par un processus *logique* : démarrant de propriétés imposées aux procédures inférentielles, traduites par la fonction de coût et la loi a priori, l'approche bayésienne déduit la meilleure solution sous ces contraintes. À l'inverse, les procédures classiques sont construites sans principe général, au sens où elles partent d'un estimateur "arbitraire" (estimateur du maximum de vraisemblance, moindres carrés, etc.) et ensuite seulement elles examinent ses propriétés fréquentistes, d'ailleurs pas toujours dans un contexte décisionnel et sans prétendre à une optimalité globale, comme le montre l'effet Stein. Dans certaines situations, les approches classiques partent aussi d'un critére pour le choix d'un estimateur (meilleur estimateur sans biais, meilleur estimateur équivariant, test uniformément plus puissant, etc.), mais elles ne peuvent pas produire une méthodologie universellement constructive, c'est à dire un algorithme, même formel, pour l'obtention des estimateurs optimaux (voir aussi le point (10)), et il est parfois nécessaire de restreindre plus avant la classe des estimateurs considérés comme, par exemple, dans le cas des tests uniformément plus puissants sans biais.

Cette opposition fondamentale dans les bases logiques des deux théories renforce notre argument de *cohérence* de l'approche bayésienne, puisque c'est la *seule*–en considérant les meilleurs estimateurs équivariants dans un cadre d'estimation bayésienne sous la mesure de Haar appropriée–à fournir une *méthodologie universelle et implémentable issue des contraintes inférentielles*.

(7) **Mettre en œuvre le principe de vraisemblance**

Le *principe de vraisemblance*, comme l'a montré le Chapitre 1, est fondé sur les principes (logiques) de *conditionnement* et d'*exhaustivité*. Par conséquent, il devrait toujours régir le choix des procédures d'estimation, créant ainsi une propriété désirable de plus par rapport à celles déjà mentionnées au point (6). La théorie bayésienne offre une technique d'implémentation de ce principe, puisqu'elle permet la construction de décisions compatibles avec ces diverses contraintes.

De plus, bien qu'elle incorpore la théorie (partielle) du maximum de vraisemblance comme cas particulier (pour $\pi(\theta) = 1$), l'approche bayésienne

peut aussi éviter certains paradoxes de vraisemblance comme ceux présentés dans la Section 4.1, grâce à l'utilisation des lois non informatives de Jeffreys, même s'il faut garder à l'esprit l'incompatibilité de ces lois avec le principe de vraisemblance. Un avantage supplémentaire et non négligeable, par rapport à la théorie du maximum de vraisemblance, est que l'approche bayésienne incorpore aussi naturellement les contraintes imposées par une fonction de coût et donc agrèg le principe de vraisemblance et la Théorie de la Décision.

(8) **Chercher des procédures fréquentistes optimales**

Du point de vue de l'approche fréquentiste, l'argument majeur en faveur de l'approche bayésienne est qu'elle s'accorde très fortement avec les trois notions d'optimalité classique que sont la minimaxité, l'admissibilité et l'équivariance. Nous avons en effet pu constater dans les Chapitres 2, 8 et 9 que la plupart des estimateurs optimaux suivant l'un de ces critères sont des estimateurs de Bayes ou des limites d'estimateurs de Bayes (la notion de *limite* dépendant du contexte). Par conséquent, non seulement il est possible de produire des estimateurs de Bayes qui satisfont à un, à deux ou à trois de ces critères d'optimalité mais, plus fondamentalement, les estimateurs de Bayes sont essentiellement les seuls à atteindre ce but. Ainsi, un statisticien bayésien peut être opposé à toute intervention *subjective* dans son traitement inférentiel sans pour autant devoir s'interdire l'utilisation systématique d'estimateurs de Bayes ou de Bayes généralisés, puisque la plupart d'entre eux satisfont ces critères[85]. On peut aussi insister sur la dérivation aisée des estimateurs de Bayes, qui fournit une méthode quasi universelle de construction d'estimateurs optimaux. Dans cette perspective utilitaire, les lois a priori sont bien sûr à considérer comme des instruments d'estimation et non plus comme des résumés exhaustifs de l'information a priori, mais leur forme et leur utilisation a posteriori demeurent bien évidemment les mêmes. L'optimalité des procédures bayésiennes est aussi satisfaite sous l'angle des principaux critères asymptotiques, puisque, sous les conditions assurant l'efficacité de l'estimateur du maximum de vraisemblance, la plupart des estimateurs de Bayes sont asymptotiquement efficaces et deviennent équivalents à l'estimateur du maximum de vraisemblance quand la taille de l'échantillon croît (voir Lehmann, 1983, et Ibragimov et Has'minskii, 1981), même si cette optimalité n'a pas été abordée dans cet ouvrage, car elle ne correspond pas particulièrement à notre vision de la Statistique.

(9) **Résoudre le véritable problème**

Il est aussi nécessaire de pouvoir fournir une alternative à l'approche fréquentiste d'un point de vue *pratique*. En effet, les méthodes fréquentistes

[85] De ce point de vue, on peut faire remarquer que les prétentions fréquentistes à l'*objectivité* (souvent opposée à la subjectivité inhérente à l'approche bayésienne) sont quelque peu amoindries quand on prend en compte la nécessité qu'a cette approche de sélectionner ses estimateurs avant de les comparer !

sont justifiées par un argument de long terme. Par exemple, un intervalle de confiance au niveau 0.95 utilisé pour des problèmes indépendants aura un taux de succès global proche de 95%, ce qui fournit une évaluation du statisticien. Au contraire, pour un décideur (le "client"), ces propriétés de long terme n'ont pas vraiment d'intérêt puisqu'il ou elle est peu intéressé(e) par les performances de long terme de la procédure proposée par le statisticien. Ce qui compte est la garantie d'une performance acceptable *pour le problème qui l'occupe* ! Par exemple, le fait qu'un médicament soit efficace dans 99% des cas n'est pas un élément important pour un patient : il désire connaître *ses* chances de guérison. Une telle demande sur la validation des procédures statistiques implique évidemment une structure qui raisonne conditionnellement à x, ce qui nous ramène nécessairement à l'approche bayésienne (voir le point (2)).

Cet argument ne semble pas s'appliquer à des cadres statistiques impliquant des expériences répétées, où la décision est prise par le même individu à chaque fois, comme en *contrôle de qualité*. Il n'empêche que ces situations justifient tout autant une résolution bayésienne, puisqu'elles sont propices à une exploitation des informations fournies par les résultats antérieurs.

(10) **Calculer des estimateurs via un programme d'optimisation**

Un dernier point argumentant en faveur du choix bayésien est que les procédures bayésiennes sont plus *faciles à calculer* que les procédures d'autres théories. Une telle assertion peut sembler pour le moins paradoxale quand on se réfère aux dévelopements des Chapitres 6 et 10 et, par exemple, aux difficultés rencontrées dans le traitement des mélanges de distributions ; plus généralement, nous avons bien vu dans les chapitres précédents que les estimateurs de Bayes apparaissent très rarement sous une forme explicite, sauf dans le cas très particulier des lois conjuguées. Cependant, on peut facilement argumenter que l'approche bayésienne fournit un *programme universel* de calcul de ses procédures, quels que soient le coût, la loi des observations et la loi a priori, puisque la solution consiste toujours en la minimisation du coût a posteriori, même si la résolution pratique de cette minimisation demande l'utilisation de techniques numériques ou de Monte Carlo.

Au contraire, l'approche fréquentiste ne dispose pas d'un algorithme universel de construction des estimateurs minimax ou admissibles, à l'exception peut-être d'émuler l'approche bayésienne par l'utilisation, respectivement, de lois a priori les moins favorables (mais sans indiquer comment obtenir ces lois les moins favorables !) ou de lois propres[86]. De même,

[86] Bien que les deux approches minimisent un coût, la différence majeure entre elles est que, pour l'approche fréquentiste, la minimisation s'opère sur un espace fonctionnel–l'espace des estimateurs–tandis que, pour l'approche bayésienne, elle est effectuée sur l'espace de décision–l'espace des estimations. Les complexités respectives de ces deux espaces sont, généralement, considérablement différentes.

la seule technique générique de construction des *meilleurs estimateurs équivariants* repose sur la connexion avec les mesures de Haar et leur représentation bayésienne, comme nous l'avons démontré au Chapitre 9.

On peut argumenter que les estimateurs du maximum de vraisemblance reposent également sur un programme général d'optimisation. Mais il faut prendre en compte la nature intrinsèquement limitée de l'approche vraisembliste, qui ne propose pas une couverture complète du champ inférentiel. De plus, les estimateurs de Bayes autorisent des représentations intégrales sous les coûts usuels, tandis que les estimateurs du maximum de vraisemblance n'existent pas toujours. C'est par exemple le cas pour les mélanges normaux, où la fonction de vraisemblance n'est pas bornée, ou pour les distributions où il existe plusieurs maxima globaux de la fonction de vraisemblance.

Par ailleurs, d'un pur point de vue pratique, il est certain que le calcul *effectif* des estimateurs de Bayes est souvent plus délicat, car il implique à la fois la résolution d'un programme de minimisation *et* des intégrations multiples. Bien que ce soit certainement un problème véritable pour le praticien, il faut cependant relativiser en prenant en compte l'existence de techniques (et de logiciels) génériques, comme par exemple les méthodes MCMC et winBUGS. En fait, la "révolution bayésienne" des années 1990 que représente le développement simultané et symbiotique de nouvelles techniques de calcul et de nouveaux champs d'application de l'analyse bayésienne en est la preuve.

Pour d'autres perspectives sur les avantages d'une approche bayésienne, on pourra se reporter aux ouvrages indiqués dans les chapitres précédents, en particulier Jeffreys (1961), Lindley (1971), Berger (1985b, Section 4.1 et Section 4.12), Berger et Wolpert (1988), Bernardo et Smith (1994), Carlin et Louis (2000a), O'Hagan et Forster (2002) et Gelman *et al.* (2003).

Les critiques de l'approche bayésienne sont nombreuses et nous ne voulons pas en dresser une liste exhaustive, d'autant qu'elles échouent à exhiber des incohérences fondamentales dans cette approche (les estimateurs de Bayes non convergents mentionnés dans la Note 1.8.4 relèvent plus d'un phénomène exotique que d'une difficulté intrinsèque). Par conséquent, nous considérons seulement trois questions usuelles sur les lois a priori, puisqu'elles constituent généralement l'aspect le plus critiqué de l'approche bayésienne.

(i) *Le passage de l'information a priori, qui peut être vague ou mal spécifiée, à la loi a priori n'est pas expliqué par les axiomes bayésiens.*

Une réponse partielle, bien que superficielle, est que la même critique s'applique aux distributions d'échantillonnage, qui sont presque toujours supposées connues. Dans de nombreux cas, et pour la plupart des approches, la modélisation a toujours une influence décisive sur l'analyse résultante, mais elle ne peut pas être formalisée au même degré que la méthodologie

qui en résulte. La diversité des sources d'information, les différents degrés de précision de cette information et l'évaluation des conséquences de la sélection de la loi a priori font que la modélisation demeure plus un art qu'une science. De plus, comme nous l'avons vu dans la Note 3.8.1, des axiomes de cohérence sur l'ordonnancement des probabilités (ou des vraisemblances) a priori justifient en partie l'existence d'une loi a priori, même si c'est généralement sur une σ-algèbre moins fine que désiré.

D'un point de vue pragmatique, la construction d'une loi a priori dépend de la capacité des individus à pouvoir représenter leurs connaissances et leurs incertitudes au travers d'une distribution de probabilité. Que les décideurs ne soient pas en mesure de le faire à présent ne signifie pas qu'ils ou elles ne peuvent pas acquérir cette capacité, à condition qu'ils ou elles puissent être formé(e)s dans ce but. L'éducation a permis à la quasi-totalité des individus des pays développés de traiter et de manipuler des quantités numériques. Elle peut de même les former à manipuler l'incertain. (Voir aussi Smith, 1988.)

Un autre argument qui mérite d'être mentionné est que l'analyse bayésienne fournit également des outils permettant de faire face aux imprécisions sur la loi a priori, via les approches hiérarchique et robuste. La caractéristique importante de la composante arbitraire dans le choix d'une loi a priori est l'influence de l'information a priori sur l'inférence a posteriori. Si différentes modélisations conduisent à des inférences similaires, l'arbitraire n'a que peu d'importance. Si, au contraire, des divergences apparaissent, elles signalent que la construction de la loi a priori doit être mieux fondée et que ses aspects les plus fragiles doivent être évalués via une *analyse de sensibilité*, sans pour autant rejeter l'information a priori disponible. Cette décomposition des facteurs influant sur l'inférence nous semble en fait constituer un avantage de l'analyse bayésienne (voir aussi ci-dessous).

(ii) *La subjectivité n'est qu'un prétexte à déviances et manipulations de toutes sortes, comme par exemple le choix de la réponse désirée* a priori.

De nouveau, il est possible d'adresser exactement la même critique à la plupart des méthodologies alternatives, par exemple à propos du choix de la fonction de coût ou de la classe d'estimateurs étudiés. Une illustration due à Brown (1980) est que, pour toute dimension p_0, il existe une fonction de coût telle que l'effet Stein ne se produit que lorsque la dimension du problème est supérieure à p_0 (voir la Note 2.8.2).

Cette mise au point étant faite, la critique est aussi justifiée envers l'analyse bayésienne, au sens où l'introduction d'un facteur additionnel dans le processus inférentiel peut toujours être détournée de son but originel. Une illustration immédiate est l'emploi de masses de Dirac comme lois a priori ! Mais il existe évidemment des stratégies beaucoup plus discrètes pour produire une inférence "à la commande"... C'est hélas une conséquence inévitable des capacités d'inclusion (de l'information a priori) et d'adapta-

tion qu'offre l'approche bayésienne. Bien entendu, dans les fondements de cette approche existe un présupposé implicite d'honnêteté du statisticien ou de l'expérimentateur qui est que le choix de la loi a priori doit pouvoir se justifier (ou se *falsifier* en langage poppérien), au sens où le ou la statisticien(ne) ou l'expérimentateur(trice) est responsable du passage de l'information dont il ou elle dispose vers la loi a priori–même si cette justification accepte les arguments de recherche de simplicité ou d'intuition personnelle, jusqu'à un certain niveau.

Insister sur la possibilité d'une vérification des sources de la modélisation n'est pas sans rappeler l'impératif de *répétabilité* des expériences dans les disciplines expérimentales, mais cette contrainte est curieusement absente dans d'autres méthodologies statistiques, ce qui signale l'ambiguïté inhérente au choix d'une procédure d'estimation, comme par exemple l'opposition entre estimateur du maximum de vraisemblance et estimateur des moindres carrés. On peut ici défendre la thèse opposée que l'approche bayésienne est à un certain point *plus* objective que les autres méthodes inférentielles parce que, d'une part, elle identifie et sépare clairement les différentes sources d'apport subjectif dans le processus inférentiel (distribution d'échantillonnage, loi a priori, fonction de coût), ce qui permet par la suite de faire d'éventuelles modifications sur ces facteurs. D'autre part, elle développe des outils objectifs d'analyse d'influence (lois non informatives, analyse de sensibilité, etc.). De ce point de vue, Poincaré (1902) fournit un argument supplémentaire dans la suite de la citation du point (5) :

> *Chacun porte en soi sa conception du monde dont il ne peut se défaire si aisément. Il faut bien, par exemple, que nous nous servions du langage, et notre langage n'est pétri que d'idées préconçues et ne peut l'être d'autre chose. Seulement ce sont des idées préconçues inconscientes, mille fois plus dangereuses que les autres. Dirons-nous que si nous en faisons intervenir d'autres, dont nous aurions pleine conscience, nous ne ferons qu'aggraver le mal ! Je ne le crois pas ; j'estime plutôt qu'elles se serviront mutuellement de contrepoids, j'allais dire d'antidote. [...] C'est assez pour nous affranchir ; on n'est plus esclave quand on peut choisir son maître.*

même si le tout dernier argument de cette citation est plus que discutable ! Le *principe des règles d'arrêt* illustre cette objectivité, au sens où la décision bayésienne est indépendante du critère d'arrêt, donc n'est pas influencée par les motivations subjectives qui ont conduit à cet échantillon. Encore une fois, si on se place dans une perspective fréquentiste, les choix des distributions d'échantillonnage et des fonctions de coût sont aussi des facteurs déterminants qu'on passe souvent sous silence (Good, 1973).

(iii) *Dans un contexte intégralement non informatif, l'utilisation de lois prétendues non informatives n'a aucune justification et ne sert que comme argument à une extension factice du champ bayésien.*

Bien entendu, nous ne voyons pas de contradiction intrinsèque à vouloir étendre le champ bayésien mais, plus fondamentalement, il nous semble que les contextes totalement non informatifs ne sont pas légions et qu'il existe toujours des indications a priori à exploiter, à moins que les conditions de l'expérience *n'exigent* un traitement non informatif ou de référence. On peut noter que les points (2), (3), (4) et (6) ci-dessus fournissent des réponses partielles à cette critique. En effet, dans un contexte non informatif, la loi a priori ne peut pas correspondre à une traduction de l'information a priori mais elle peut cependant être comprise comme outil d'inférence efficace[87]. Vu sous cet angle, les méthodes bayésiennes non informatives ne sont ni plus ni moins arbitraires que les méthodes par maximum de vraisemblance, puisque toutes sont issues de la loi des observations, qui représente la seule information a priori disponible. Si une fonction de coût est aussi fournie par le décideur ou le contexte, elle donne une information supplémentaire dont l'approche bayésienne peut faire bon usage, au contraire de la méthode du maximum de vraisemblance. Enfin, comme ces approches non informatives fournissent la plupart des estimateurs usuels, elles ne peuvent être rejetées uniquement parce qu'elles sont bayésiennes ! Au contraire, d'un point de vue strictement bayésien, on argumenterait que les bonnes performances de ces estimateurs *découlent* de leur caractère bayésien (Jaynes, 1980).

Dans les points précédents, nous avons également insisté sur la nécessité de conditionner en l'observation x. Ce conditionnement implique *stricto sensu* l'existence d'une modélisation probabiliste sur θ, donc une loi a priori, puisque l'approche par maximum de vraisemblance ne peut pas fournir une inférence statistique complète et ne fonctionne que très rarement comme distribution "objective" sur θ.

La technique de détermination de lois non informatives due à Jeffreys n'est donc qu'une *technique* qui prend en compte l'information présente dans le modèle (ce qui signifie dans ce cas l'information apportée par les x sur θ), tout en conservant le riche éventail des outils bayésiens, en restant compatible avec les contraintes intuitives comme l'invariance *et* en incluant la plupart des procédures usuelles. La nécessité d'une telle approche apparaît clairement en théorie des tests, où la perspective de Neyman-Pearson est déficiente à de nombreux points de vue[88].

[87] Cette critique sur les lois non informatives procède d'un argument qui rejette l'utilisation de l'information a priori, sauf lorsque celle-ci n'est pas disponible !

[88] La difficulté de traiter des hypothèses ponctuelles par des lois impropres, abordée dans les Chapitres 5 et 7, est réelle et ne doit pas être sous-estimée. La diversité de réponses possibles présentée par exemple au Chapitre 7 est cependant rassurante.

Bien que le traitement des *paramètres de nuisance* conduise à des difficultés *techniques* (comme par exemple les paradoxes de marginalisation de la Section 3.5), la généralisation par les lois de référence proposée par Bernardo (1979) fournit une solution partielle à cette difficulté. Un autre problème mentionné dans ce livre est celui de l'estimation des *modèles de mélange*, vue au Chapitre 6, qui est fondamentalement lié au manque d'*identifiabilité* de ces modèles. Il admet néanmoins une résolution non informative via un changement de paramètres (Robert et Titterington, 1998).

Un point positif de ces critiques est qu'elles soulignent l'importance de construire rigoureusement la loi a priori dans l'analyse bayésienne. Elles poussent également à des études plus avancées sur les techniques non informatives, comme par exemple l'exploitation de l'information contenue dans la fonction de coût ou la cohérence des suites de lois utilisées dans la comparaison de modèles imbriqués. Elles signalent en sus le besoin de voir se développer des techniques "automatiques" (ou semi-automatiques) de détermination des lois a priori afin de permettre une utilisation plus universelle des méthodes bayésiennes en Statistique appliquée. Des logiciels bayésiens sont dès à présent disponibles (voir la Note 6.6.2, et Berger, 2000). En symbiose avec les méthodes numériques d'approximation présentées au Chapitre 6, ces techniques devraient favoriser la diffusion de la méthodologie bayésienne dans de nombreuses communautés. La croissance exponentielle des applications bayésiennes dans les dix dernières années est un signal fort que cette diffusion est en cours (voir Berger, 2000).

Pour achever cette conclusion, notons enfin que le caractère antagoniste des approches bayésiennes et non bayésiennes est parfois démesurément amplifié. Pour un observateur extérieur non statisticien, et en particulier pour les étudiant(e)s, la querelle entre classiques et bayésiens n'est pas compréhensible et donne l'image d'une discipline peu fiable puisque les experts n'arrivent pas à y définir un standard unique ! Par ailleurs, l'utilisation toujours croissante de techniques de traitement de données par des non statisticiens a tendance à effacer les motivations philosophiques et les frontières théoriques entre méthodes pour se consacrer à leur applicabilité. (On peut regretter cette prise de pouvoir mais elle a déjà eu lieu !) D'un point de vue théorique, les récents développements de la Théorie de la Décision (paramétrique et non paramétrique) ont renforcé les fondations bayésiennes des notions classiques d'optimalité (voir (6)), tandis que l'état de l'art en robustesse bayésienne cherche à réduire les problèmes de mauvais choix de la loi a priori en prenant en compte des critères fréquentistes (comme la minimaxité ou la *minimaxité bayésienne* étudiée dans Kempthorne, 1988). Par ailleurs, des chercheurs éminents des deux communautés sont engagés dans la production d'un contexte décisionnel qui donnerait des procédures acceptables par les deux écoles, comme nous le décrivons dans la Note 5.7.4. En pratique, les approximations fréquentistes sont aussi bien souvent nécessaires lorsque la

construction de la loi a priori est délicate, par exemple lorsque l'information de Fisher n'existe pas sous forme explicite ou quand la dimension de l'espace des paramètres est trop grande.

Le choix de l'approche bayésienne peut donc être fondé sur la réconciliation de la plupart des procédures classiques avec une analyse bayésienne ou bayésienne généralisée, sur l'attrait indéniable de sa complétude et de sa cohérence globale, et aussi sur sa capacité à élargir le champ des inférences possibles, sans avoir besoin de rejeter *toutes* les procédures classiques. L'approche bayésienne nous semble tout simplement plus en harmonie avec ce que doit être l'inférence statistique, tout en étant plus attrayante intellectuellement.

A

Distributions de probabilité

Nous donnons dans cet appendice quelques rappels sur les distributions les plus couramment utilisées dans ce livre, au travers de leur densité et des deux premiers moments. Une revue quasi exhaustive des distributions usuelles est fournie par les livres de Johnson et Kotz (1972), Johnson *et al.* (1994), Johnson *et al.* (1995) et Johnson et Hoeting (2003). À chaque fois, les densités sont données par rapport à la mesure de Lebesgue ou à la mesure de comptage, suivant le contexte (variable réelle ou entière).

A.1. **Loi normale,** $\mathcal{N}_p(\theta, \Sigma)$

($\theta \in \mathbb{R}^p$ et Σ est une matrice symétrique $(p \times p)$ définie positive)

$$f(\mathbf{x}|\theta, \Sigma) = (\det \Sigma)^{-1/2}(2\pi)^{-p/2} e^{-(\mathbf{x}-\theta)^t \Sigma^{-1}(\mathbf{x}-\theta)/2}$$

$\mathbb{E}_{\theta,\Sigma}[\mathbf{X}] = \theta$ et $\mathbb{E}_{\theta,\Sigma}[(\mathbf{X}-\theta)(\mathbf{X}-\theta)^t] = \Sigma$. Quand Σ n'est pas définie positive, la loi $\mathcal{N}_p(\theta, \Sigma)$ n'a pas de densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^p$. Pour $p = 1$, la loi *log-normale* est définie comme la loi de $\exp X$ quand $X \sim \mathcal{N}(\theta, \sigma^2)$.

A.2. **Loi gamma,** $\mathcal{G}a(\alpha, \beta)$

($\alpha, \beta > 0$)

$$f(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \mathbb{I}_{[0,+\infty)}(x)$$

$\mathbb{E}_{\alpha,\beta}[X] = \alpha/\beta$ et $\text{var}_{\alpha,\beta}(X) = \alpha/\beta^2$. Des cas particuliers de la loi gamma sont les lois d'*Erlang*, $\mathcal{G}a(\alpha, 1)$, *exponentielle* $\mathcal{G}a(1, \beta)$ (notée $\mathcal{E}xp(\beta)$) et la loi du *khi deux*, $\mathcal{G}a(\nu/2, 1/2)$ (notée χ^2_ν). (Remarquons que la convention inverse est parfois adoptée pour le second paramètre, donc que la loi $\mathcal{G}a(\alpha, \beta)$ peut parfois être donnée comme $\mathcal{G}a(\alpha, 1/\beta)$. Voir, par exemple, Berger, 1985b, qui la définit ainsi pour pouvoir utiliser une loi gamma comme loi conjuguée.)

A.3. **Loi bêta,** $\mathscr{B}e(\alpha,\beta)$

$(\alpha,\beta>0)$

$$f(x|\alpha,\beta)=\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}\mathbb{I}_{[0,1]}(x)$$

où

$$B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

$\mathbb{E}_{\alpha,\beta}[X]=\alpha/(\alpha+\beta)$ et $\mathrm{var}_{\alpha,\beta}(X)=\alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)]$. La loi bêta s'obtient comme la loi de $Y_1/(Y_1+Y_2)$ quand $Y_1\sim\mathscr{G}a(\alpha,1)$ et $Y_2\sim\mathscr{G}a(\beta,1)$.

A.4. **Loi de Student,** $\mathscr{T}_p(\nu,\theta,\Sigma)$

($\nu>0$, $\theta\in\mathbb{R}^p$, et Σ est une matrice symétrique a $(p\times p)$ définie positive.)

$$f(\mathbf{x}|\nu,\theta,\Sigma)=\frac{\Gamma((\nu+p)/2)/\Gamma(\nu/2)}{(\det\Sigma)^{1/2}(\nu\pi)^{p/2}}\left[1+\frac{(\mathbf{x}-\theta)^t\Sigma^{-1}(\mathbf{x}-\theta)}{\nu}\right]^{-(\nu+p)/2}$$

$\mathbb{E}_{\nu,\theta,\Sigma}[\mathbf{X}]=\theta$ $(\nu>1)$ et $\mathbb{E}_{\theta,\Sigma}[(\mathbf{X}-\theta)(\mathbf{X}-\theta)^t]=\nu\Sigma/(\nu-2)$ $(\nu>2)$. Quand $p=1$, un cas particulier de la loi de Student est la loi de *Cauchy*, $\mathscr{C}(\theta,\sigma^2)$, qui correspond à $\nu=1$. La loi de Student $\mathscr{T}_p(\nu,0,I)$ s'obtient comme loi de $\mathbf{X}/Z$ lorsque $\mathbf{X}\sim\mathscr{N}_p(0,I)$ et $\nu Z^2\sim\chi^2_\nu$. (On l'appelle aussi, pour des raisons historiques, la loi du t de Student.)

A.5. **Loi de Fisher,** $\mathscr{F}(\nu,\varrho)$

$(\nu,\varrho>0)$

$$f(x|\nu,\varrho)=\frac{\Gamma((\nu+\varrho)/2)\nu^{\varrho/2}\varrho^{\nu/2}}{\Gamma(\nu/2)\Gamma(\varrho/2)}\frac{x^{(\nu-2)/2}}{(\nu+\varrho x)^{(\nu+\varrho)/2}}\mathbb{I}_{[0,+\infty)}(x)$$

$\mathbb{E}_{\nu,\varrho}[X]=\varrho/(\varrho-2)$ $(\varrho>2)$ et $\mathrm{var}_{\nu,\varrho}(X)=2\varrho^2(\nu+\varrho-2)/[\nu(\varrho-4)(\varrho-2)^2]$ $(\varrho>4)$.

La loi $\mathscr{F}(p,q)$ est aussi la loi de $(\mathbf{X}-\theta)^t\Sigma^{-1}(\mathbf{X}-\theta)/p$ lorsque $\mathbf{X}\sim\mathscr{T}_p(q,\theta,\Sigma)$. De plus, si $X\sim\mathscr{F}(\nu,\varrho)$, $\nu X/(\varrho+\nu X)\sim\mathscr{B}e(\nu/2,\varrho/2)$. (On l'appelle aussi, toujours pour des raisons historiques, la loi du F de Fisher.)

A.6. **Loi gamma inverse,** $\mathscr{IG}(\alpha,\beta)$

$(\alpha,\beta>0)$

$$f(x|\alpha,\beta)=\frac{\beta^\alpha}{\Gamma(\alpha)}\frac{e^{-\beta/x}}{x^{\alpha+1}}\mathbb{I}_{[0,+\infty[}(x)$$

$\mathbb{E}_{\alpha,\beta}[X]=\beta/(\alpha-1)$ $(\alpha>1)$ et $\mathrm{var}_{\alpha,\beta}(X)=\beta^2/((\alpha-1)^2(\alpha-2))$ $(\alpha>2)$. Cette loi est celle de X^{-1} lorsque $X\sim\mathscr{G}a(\alpha,\beta)$. Elle apparaît naturellement dans l'analyse bayésienne des lois gamma et normale.

A.7. **Loi du khi deux décentré,** $\chi^2_\nu(\lambda)$

$(\lambda \geq 0)$

$$f(x|\lambda) = \frac{1}{2}(x/\lambda)^{(p-2)/4} I_{(p-2)/2}(\sqrt{\lambda x}) e^{-(\lambda+x)/2}$$

$\mathbb{E}_\lambda[X] = p + \lambda$ et $\text{var}_\lambda(X) = 3p + 4\lambda$. Cette loi est celle de $X_1^2 + \cdots + X_p^2$ lorsque $X_i \sim \mathcal{N}(\theta_i, 1)$ et $\theta_1^2 + \ldots + \theta_p^2 = \lambda$.

A.8. **Loi de Dirichlet,** $\mathscr{D}_k(\alpha_1, \ldots, \alpha_k)$

$(\alpha_1, \ldots, \alpha_k > 0$ et $\alpha_0 = \alpha_1 + \cdots + \alpha_k)$

$$f(x|\alpha_1, \ldots, \alpha_k) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_k)} x_1^{\alpha_1 - 1} \ldots x_k^{\alpha_k - 1} \mathbb{I}_{\{\sum x_i = 1\}}$$

$\mathbb{E}_\alpha[X_i] = \alpha_i/\alpha_0$, $\text{var}(X_i) = (\alpha_0 - \alpha_i)\alpha_i/[\alpha_0^2(\alpha_0 + 1)]$ et $\text{cov}(X_i, X_j) = -\alpha_i\alpha_j/[\alpha_0^2(\alpha_0 + 1)]$ $(i \neq j)$. Comme cas particulier, notons que $(X, 1 - X) \sim \mathscr{D}_2(\alpha_1, \alpha_2)$ est équivalent à $X \sim \mathscr{B}e(\alpha_1, \alpha_2)$.

A.9. **Loi de Pareto,** $\mathscr{P}a(\alpha, x_0)$

$(\alpha > 0$ et $x_0 > 0)$

$$f(x|\alpha, x_0) = \alpha \frac{x_0^\alpha}{x^{\alpha+1}} \mathbb{I}_{[x_0, +\infty[}(x)$$

$\mathbb{E}_{\alpha,x_0}[X] = \alpha x_0/(\alpha - 1)$ $(\alpha > 1)$ et $\text{var}_{\alpha,x_0}(X) = \alpha x_0^2/[(\alpha - 1)^2(\alpha - 2)]$ $(\alpha > 2)$.

A.10. **Loi binomiale,** $\mathscr{B}(n, p)$

$(0 \leq p \leq 1)$

$$f(x|p) = \binom{n}{x} p^x (1 - p)^{n-x} \mathbb{I}_{\{0,\ldots,n\}}(x)$$

$\mathbb{E}_p(X) = np$ et $\text{var}(X) = np(1 - p)$.

A.11. **Loi multinomiale,** $\mathscr{M}_k(n; p_1, \ldots, p_k)$

$(p_i \geq 0$ $(1 \leq i \leq k)$ et $\sum_i p_i = 1)$

$$f(x_1, \ldots, x_k|p_1, \ldots, p_k) = \binom{n}{x_1 \ \ldots \ x_k} \prod_{i=1}^k p_i^{x_i} \mathbb{I}_{\sum x_i = n}$$

$\mathbb{E}_p(X_i) = np_i$, $\text{var}(X_i) = np_i(1 - p_i)$, et $\text{cov}(X_i, X_j) = -np_ip_j$ $(i \neq j)$. Notons que, si, $X \sim \mathscr{M}_k(n; p_1, \ldots, p_k)$, alors $X_i \sim \mathscr{B}(n, p_i)$, et aussi que la loi $X \sim \mathscr{B}(n, p)$ correspond à $(X, n - X) \sim \mathscr{M}_2(n; p, 1 - p)$. Cette loi est celle du tirage indépendant avec remise.

A.12. **Loi de Poisson,** $\mathscr{P}(\lambda)$

$(\lambda > 0)$

$$f(x|\lambda) = e^{-\lambda} \frac{\lambda^x}{x!} \mathbb{I}_{\mathbb{N}}(x)$$

$\mathbb{E}_\lambda[X] = \lambda$ et $\text{var}_\lambda(X) = \lambda$.

A.13. **Loi binomiale négative,** $\mathscr{N}eg(n,p)$

$(0 \leq p \leq 1)$

$$f(x|p) = \binom{n+x+1}{x} p^n (1-p)^x \mathbb{I}_{\mathbb{N}}(x)$$

$\mathbb{E}_p[X] = n(1-p)/p$ et $\text{var}_p(X) = n(1-p)/p^2$.

A.14. **Loi hypergéométrique,** $\mathscr{H}yp(N;n;p)$

$(0 \leq p \leq 1,\ n < N$ et $pN \in \mathbb{N})$

$$f(x|p) = \frac{\binom{pn}{x}\binom{(1-p)N}{n-x}}{\binom{N}{n}} \mathbb{I}_{\{n-(1-p)N,\ldots,pN\}}(x) \mathbb{I}_{\{0,1,\ldots,n\}}(x)$$

$\mathbb{E}_{N,n,p}[X] = np$ et $\text{var}_{N,n,p}(X) = (N-n)np(1-p)/(N-1)$. Cette loi est celle du tirage sans remise.

B

Notations

mathématiques

$A \prec B$	$(B - A)$ est une matrice définie positive
$\|A\|, \det(A)$	déterminant de la matrice A
a^+	$\max(a, 0)$
C_n^p, $\binom{n}{p}$	coefficient binomial
D_α	fonction logistique
$\Delta f(z)$	laplacien de $f(z)$
${}_1F_1(a; b; z)$	fonction confluente hypergéométrique
F^-	inverse généralisée de F
$f(t) \propto g(t)$	les fonctions f et g sont proportionnelles
$\Gamma(a)$	fonction gamma ($a > 0$)
$\mathbf{h} = (h_1, \ldots, h_n) = \{h_i\}$	un caractère gras représente un vecteur
$H = \{h_{ij}\} = \|\|h_{ij}\|\|$	une majuscule représente une matrice
I, $\mathbf{1}$, $J = \mathbf{1}\mathbf{1}'$	matrice identité, vecteur identité, matrice unitaire
$\mathbb{I}_A(t)$	fonction indicatrice (1 si $t \in A$, 0 sinon)
$I_\nu(z)$	fonction de Bessel modifiée ($z > 0$)
$\lambda_{\max}(A)$	plus grande valeur propre de la matrice A
$\binom{n}{n_1 \ldots n_k}$	coefficient multinomial
$\nabla f(z)$	gradient de $f(z)$, soit le vecteur de terme générique $(\partial/\partial z_i) f(z)$ ($f(z) \in \mathbb{R}$ et $z \in \mathbb{R}^p$)
$\nabla^t f(z)$	divergence de $f(z)$, $\sum(\partial/\partial z_i) f(z)$ ($f(z) \in \mathbb{R}^p$ et $z \in \mathbb{R}$)
$\|\| \cdot \|\|_{TV}$	norme de la variation totale
$\Psi(x)$	fonction digamma ($x > 0$)
$\Sigma \otimes \Psi$	produit tensoriel des matrices Σ et Ψ
$\text{supp}(f)$	support de f
$\text{tr}(A)$	trace de la matrice A
$\|\mathbf{x}\| = (\Sigma x_i^2)^{1/2}$	norme euclidienne
$\lfloor x \rfloor$	partie entière de x, le plus grand entier inférieur à x

$\lceil x \rceil$	plus petit entier plus grand que x
$x \vee y$	maximum de x et y
$x \wedge y$	minimum de x et y
$< x, y >$	produit scalaire de x et y dans $\mathbb{R}^p$

probabilistes

X, Y	variable aléatoire (majuscule)
$(\mathcal{X}, \mathcal{P}, \mathcal{B})$	triplet probabiliste : espace des variables, loi, et σ-algèbre
β_n	coefficient de β-mélangeance
$\delta_{\theta_0}(\theta)$	masse de Dirac en θ_0
$\mathscr{E}(\theta)$	fonction énergie d'une distribution de Gibbs
$\mathfrak{E}(\pi)$	entropie de la loi π
$\mathbb{E}_\theta[g(X)]$	espérance de $g(x)$ pour la loi $f(x\|\theta)$ sur X
$\mathbb{E}^V[h(V)]$	espérance de $h(v)$ pour la loi de V
$\mathbb{E}^\pi[h(\theta)\|x]$	espérance de $h(\theta)$ pour la loi de θ conditionnellement à x, $\pi(\theta\|x)$
$F(x\|\theta)$	fonction de répartition de X, conditionnellement au paramètre θ
$f(x\|\theta)$	densité de x indicée par le paramètre θ par rapport à la mesure de Lebesgue ou à la mesure de comptage
iid	indépendant et identiquement distribué (pour un échantillon)
$\lambda(dx)$	mesure de Lebesgue (encore notée $d\lambda(x)$)
P_θ	distribution de probabilité indexée par le paramètre θ
$\varphi(t)$	densité de la loi normale $\mathscr{N}(0,1)$
$\Phi(t)$	fonction de répartition de la loi normale $\mathscr{N}(0,1)$
$f(x\|\theta)$	
P_θ	loi de probabilité, indexée par le paramètre θ
$p \star q$	produit de convolution des lois p et q, soit loi de la somme de $X \sim p$ et de $Y \sim q$
$p^{n\star}$	n-ième produit de convolution, soit, loi de la somme de n variables iid de loi p
$\varphi(t)$	densité de la loi normale $\mathscr{N}(0,1)$
$\Phi(t)$	fonction de répartition de la loi normale $\mathscr{N}(0,1)$
$O(n), o(n)$	grand "O", petit "o". Quand $n \to \infty$, $\frac{O(n)}{n} \to$ constante,
ou $O_p(n), o_p(n)$	$\frac{o(n)}{n} \to 0$, et l'indice p signifie *en probabilité*
$X_t, X^{(t)}$	élément générique d'une chaîne de Markov
$x \sim f(x\|\theta)$	x est distribuée suivant la loi de densité

de distributions

$\mathscr{B}(n,p)$	loi binomiale
$\mathscr{B}e(\alpha,\beta)$	loi bêta
$\mathscr{C}(\theta,\sigma^2)$	loi de Cauchy
$\mathscr{D}_k(\alpha_1,\ldots,\alpha_k)$	loi de Dirichlet

$\mathscr{E}xp(\lambda)$	loi exponentielle
$\mathscr{F}(p,q)$	loi de Fisher
$\mathscr{G}a(\alpha,\beta)$	loi gamma
$\mathscr{IG}(\alpha,\beta)$	loi inverse gamma
$\mathscr{IN}(\alpha,\mu)$	loi normale inverse
χ^2_p	loi du khi deux centrée
$\chi^2_p(\lambda)$	loi du khi deux non centrée
$\mathscr{M}_k(n;p_1,..,p_k)$	loi multinomiale
$\mathscr{N}(\theta,\sigma^2)$	loi normale unidimensionnelle
$\mathscr{N}_p(\theta,\Sigma)$	loi normale multidimensionnelle
$\mathscr{N}eg(n,p)$	loi binomiale négative
$\mathscr{P}(\lambda)$	loi de Poisson
$\mathscr{P}(x_0,\alpha)$	loi de Pareto
$\mathscr{T}_p(\nu,\theta,\Sigma)$	loi de Student multidimensionnelle
$\mathscr{U}_{[a,b]}, \mathscr{U}([a,b])$	loi uniforme (continue)
$\mathscr{W}e(\alpha,c)$	loi de Weibull
$\mathscr{W}_k(p,\Sigma)$	loi de Wishart

décisionnelles

$\mathscr{D}$	espace des décisions
$\mathscr{G}$	groupe agissant sur $\mathscr{X}$
$\bar{g}$	élément de $\bar{\mathscr{G}}$ associé à $g \in \mathscr{G}$
$\bar{\mathscr{G}}$	groupe induit par $\mathscr{G}$ agissant sur Θ
$\tilde{g}$	élément de $\tilde{\mathscr{G}}$ associé à $g \in \mathscr{G}$
$\tilde{\mathscr{G}}$	groupe induit par $\mathscr{G}$ agissant sur $\mathscr{D}$
$\mathrm{L}(\theta,\delta)$	fonction de coût de δ en θ
$\mathscr{M}_0, \mathscr{M}_k$	modèles considérés
$R(\theta,\delta)$	risque fréquentiste de δ en θ
$r(\pi,\delta)$	risque de Bayes de δ pour la loi a priori π
$\varrho(\pi,\delta\|x)$	risque a posteriori de δ pour la loi a priori π
Θ	espace des paramètres
$\mathscr{X}$	espace des observations

statistiques

$\mathrm{AR}(p)$	processus autorégressif d'ordre p
$\mathrm{ARMA}(p,q)$	processus autorégressif à moyenne mobile d'ordre (p,q)
$B^\pi(x)$	facteur de Bayes
$B^A_{12}(x)$, B^G_{12}, B^M_{12}	pseudo-facteur de Bayes
$\underline{B}$	borne inférieure sur un facteur de Bayes
C_α	région de confiance (ou crédible)
$\delta^{\mathrm{JS}}(x)$	estimateur de James-Stein

$\delta^\pi(x)$	estimateur de Bayes
$\delta^+(x)$	estimateur de James-Stein tronqué
$\delta^\star(x)$	estimateur randomisé
H_0	hypothèse nulle
H_1, H_a	hypothèse alternative
$I(\theta)$	information de Fisher
$\mathrm{L}(\theta, \delta)$	fonction de coût de δ en θ
$\ell(\theta\|x)$	vraisemblance, en tant que fonction de θ, identique à $f(x\|\theta)$
$\ell^P(\theta\|x)$	vraisemblance profilée
$m(x)$	loi marginale
$\mathrm{MA}(q)$	processus à moyenne mobile d'ordre q
$\overset{\mathrm{P}}{\succ}$	domination au sens de Pitman
$\pi(\theta)$	loi a priori générique sur θ
$\pi^{\mathrm{J}}(\theta)$	loi a priori de Jeffreys sur θ
$\pi(\theta\|x)$	loi a posteriori générique sur θ
s^2	somme des carrés des écarts à la moyenne empirique
θ, λ	paramètres (lettres grecques minuscules)
Θ	espace des paramètres (lettres grecques majuscules)
$\bar{x}$	moyenne empirique
x^*, y^*	données latentes ou manquantes

Références

Abraham, C. (2001). Asymptotic limit of the Bayes actions set derived from a class of loss functions. *J. Multiv. Analysis*, 79(2), 251–274.

Abraham, C. et Daurés, J. (2000). Global robustness with respect to the loss function and the prior. *Theory and Decision*, 48(4), 359–381.

Abramovich, F., Spatinas, T., et Silverman, B. (1998). Wavelet thresholding via a Bayesian approach. *J. Royal Statist. Soc. Series B*, 60, 725–749.

Abramowitz, M. et Stegun, I. (1964). *Handbook of Mathematical Functions*. Dover, New York.

Adams, M. (1987). *William Ockham*. University of Notre Dame Press, Notre Dame, Indiana.

Aitkin, M. (1991). Posterior Bayes factors (with discussion). *J. Royal Statist. Soc. Series B*, 53, 111–142.

Akaike, H. (1978). A new look at the Bayes procedure. *Biometrika*, 65, 53–59.

Akaike, H. (1983). Information measure and model selection. *Bull. Int. Statist. Inst.*, 50, 277–290.

Alam, K. (1973). A family of admissible minimax estimators of the mean of a multivariate normal distribution. *Ann. Statist.*, 1, 517–525.

Albert, J. (1988). Computational methods using a Bayesian hierarchical generalized linear model. *J. American Statist. Assoc.*, 83, 1037–1044.

Anderson, T. (1984). *An Introduction to Multivariate Statistical Analysis*. John Wiley, New York, seconde édition.

Andrieu, C. et Doucet, A. (1999). Joint Bayesian detection and estimation of noisy sinusoids via reversible jump MCMC. *IEEE Trans. Signal Proc.*, 47(10), 2667–2676.

Andrieu, C., Doucet, A., et Fitzgerald, W. (2000). On Monte Carlo methods for Bayesian data analysis. In Mees, A. et R.L., S., éditeurs, *Nonlinear Dynamics and Statistics*. Birkhauser, Boston.

Angers, J. (1987). *Development of robust Bayes estimators for a multivariate normal mean*. PhD thesis, Purdue University, West Lafayette, Indiana.

Angers, J. (1992). Use of the Student's t-prior for the estimation of normal means : A computational approach. In Bernardo, J., Berger, J., Dawid, A., et Smith, A., éditeurs, *In Bayesian Statistics*, volume 4, pages 567–575. Oxford University Press, Oxford.

Angers, J. et MacGibbon, K. (1990). Hierarchical Bayes estimation in linear models with robustness against partial prior misspecification. Technical Report 69, Dépt. de Mathématiques et d'Informatique, Université de Sherbrooke.

Arrow, K. (1956). *Social Choice and Individual Values*. John Wiley, New York.

Bar-Lev, S., Enis, P., et Letac, G. (1994). Models which admit a given exponential family as an a priori conjugate model. *Ann. Statist.*, 22(3), 1555–1586.

Baranchick, A. (1970). A family of minimax estimators of the mean of a multivariate normal distribution. *Ann. Mathemat. Statist.*, 41, 642–645.

Barbieri, M., Liseo, B., et Petrella, L. (1999). Bayes factor at work in a challenging class of problems. In Racugno, W., éditeur, *Model Choice Collana Atti di Congressi*, pages 109–132. Pitagora Editrice, Bologna.

Barnard, G. (1949). Statistical inference (with discussion). *J. Royal Statist. Soc. Series B*, 11, 115–159.

Barnett, G., Kohn, R., et Sheather, S. (1996). Bayesian estimation of an autoregressive model using Markov chain Monte Carlo. *J. Econometrics*, 74, 237–254.

Barron, A. (1988). The exponential convergence of posterior probabilities with implication for Bayes estimators of density functions. Technical report, Dept. of Statistics, University of Illinois.

Barron, A. (1998). Information-theoretic characterization of Bayes performances and the choice of priors in parametric and nonparametric problems (with discussion). In Bernardo, J., Berger, J., Dawid, A., et Smith, A., éditeurs, *Bayesian Statistics*, volume 6, pages 27–52. Oxford University Press, Oxford.

Barron, A., Schervish, M., et Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.*, 27(2), 536–561.

Bartlett, M. (1937). Properties of sufficiency and statistical tests. *Proc. Roy. Soc. London*, 130, 268–282.

Basu, D. (1988). *Statistical Information and Likelihood*. Springer-Verlag, New York.

Baum, L. et Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Mathemat. Statist.*, 37, 1554–1563.

Bauwens, L. (1984). *Bayesian Full Information of Simultaneous Equations Models Using Integration by Monte Carlo*, volume 232 dans *Lecture Notes in Economics and Mathematical Systems*. Springer-Verlag, New York.

Bauwens, L. (1991). The "pathology" of the natural conjugate prior density in the regression model. *Ann. Econom. Statist.*, 23, 49–64.

Bauwens, L., Lubrano, M., et Richard, J. (1999). Bayesian inference in dynamic econometric models. In Granger, C. et Mizon, G., éditeurs, *Advanced Texts in Econometrics*. Oxford University Press, Oxford.

Bayarri, M. et DeGroot, M. (1988). Gaining weight : a Bayesian approach. In Bernardo, J., DeGroot, M., D., L., et A.F.M., S., éditeurs, *Bayesian Statistics*, volume 3, pages 25–44. Oxford University Press, Oxford.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc.*, 53, 370–418.

Bechofer, R. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variance. *Ann. Mathemat. Statist.*, 25, 16–39.

Bensmail, H., Celeux, G., Raftery, A., et Robert, C. (1997). Inference in model-based cluster analysis. *Statist. Comp.*, 7(1), 1–10.

Beran, R. (1996). Stein estimation in high dimension : A retrospective. In *Research Developments in Probability and Statistics : Madan L. Puri Festschrift*, pages 91–110. Universiteit Utrecht.

Bergé, P., Pommeau, Y., et Vidal, C. (1984). *Order Within Chaos*. John Wiley, New York.

Berger, J. (1975a). Admissibility results for generalized Bayes estimators of a location vector. *Ann. Statist.*, 4, 334–356.

Berger, J. (1975b). Minimax estimation of location vectors for a wide class of densities. *Ann. Statist.*, 3, 1318–1328.

Berger, J. (1980a). Improving on inadmissible estimators in continuous exponential families with applications to simultaneous estimation of gamma scale parameters. *Ann. Statist.*, 8, 545–571.

Berger, J. (1980b). A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *Ann. Statist.*, 8, 716–761.

Berger, J. (1982a). Estimation in continuous exponential families : Bayesian estimation subject to risk restrictions and inadmissibility results. In Gupta, S. et Berger, J., éditeurs, *Statistical Decision Theory and Related Topics*, volume 3, pages 109–142. Academic Press, New York.

Berger, J. (1982b). Selecting a minimax estimator of a multivariate normal mean. *Ann. Statist.*, 10, 81–92.

Berger, J. (1984). The robust Bayesian viewpoint (with discussion). In Kadane, J., éditeur, *Robustness of Bayesian Analysis*. North-Holland, Amsterdam.

Berger, J. (1985a). Discussion of 'quantifying prior opinion' by Diaconis and Ylvisaker. In Bernardo, J., DeGroot, M., Lindley, D., et Smith, A., éditeurs, *Bayesian Statistics*, volume 3, Amsterdam. North-Holland.

Berger, J. (1985b). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York, second édition.

Berger, J. (1990). Robust Bayesian analysis : sensitivity to the prior. *J. Statist. Plann. Inference*, 25, 303–328.

Berger, J. (2000). Bayesian analysis : A look at today and thoughts of tomorrow. *J. American Statist. Assoc.*, 95, 1269–1277.

Berger, J. et Berliner, L. (1986). Robust Bayes and empirical Bayes analysis with ε-contamined priors. *Ann. Statist.*, 14, 461–486.

Berger, J. et Bernardo, J. (1989). Estimating a product of means : Bayesian analysis with reference priors. *J. American Statist. Assoc.*, 84, 200–207.

Berger, J. et Bernardo, J. (1992a). On the development of the reference prior method. In Bernardo, J., Berger, J., Dawid, A., Lindley, D., et Smith, A., éditeurs, *Bayesian Statistics 4*, pages 35–60, London. Oxford University Press.

Berger, J. et Bernardo, J. (1992b). Ordered group reference priors with application to the multinomial problem. *Biometrika*, 79, 25–37.

Berger, J. et Bock, M. (1976). Eliminating singularities of Stein-type estimators of location vectors. *J. Royal Statist. Soc. Series B*, 39, 166–170.

Berger, J., Boukai, B., et Wang, Y. (1997). Unified frequentist and Bayesian testing of a precise hypothesis (with discussion). *Statistical Science*, 12, 133–160.

Berger, J., Brown, L., et Wolpert, R. (1994). A unified conditional frequentist and Bayesian test for fixed and sequential hypothesis testing. *Ann. Statist.*, 22, 1787–1807.

Berger, J. et Deely, J. (1988). A Bayesian approach to ranking and selection of related means with alternatives to ANOVA methodology. *J. American Statist. Assoc.*, 83, 364–373.

Berger, J. et Delampady, M. (1987). Testing precise hypotheses (with discussion). *Statist. Science*, 2, 317–352.

Berger, J. et Mortera, J. (1991). Interpreting the stars in precise hypothesis testing. *International Statistical Review*, 59, 337–353.

Berger, J. et Pericchi, L. (1996a). The intrinsic Bayes factor for linear model. In Bernardo, J., Berger, J., Dawid, A., Lindley, D., et Smith, A., éditeurs, *Bayesian Statistics 5*, pages 23–42, Oxford. Oxford University Press.

Berger, J. et Pericchi, L. (1996b). The intrinsic Bayes factor for model selection and prediction. *J. American Statist. Assoc.*, 91, 109–122.

Berger, J. et Pericchi, L. (1998). Accurate and stable Bayesian model selection : the median intrinsic Bayes factor. *Sankhya B*, 60, 1–18.

Berger, J. et Pericchi, L. (2001). Objective Bayesian methods for model selection : introduction and comparison. In Lahiri, P., éditeur, *Model Selection*, volume 38 dans *Lecture Notes – Monograph Series*, pages 135–207, Beachwood Ohio. Institute of Mathematical Statistics.

Berger, J., Philippe, A., et Robert, C. (1998). Estimation of quadratic functions : reference priors for non-centrality parameters. *Statistica Sinica*, 8(2), 359–375.

Berger, J. et Robert, C. (1990). Subjective hierarchical Bayes estimation of a multivariate normal mean : on the frequentist interface. *Ann. Statist.*, 18, 617–651.

Berger, J. et Sellke, T. (1987). Testing a point-null hypothesis : the irreconcilability of significance levels and evidence (with discussion). *J. American Statist. Assoc.*, 82, 112–122.

Berger, J. et Srinivasan, C. (1978). Generalized Bayes estimators in multivariate problems. *Ann. Statist.*, 6, 783–801.

Berger, J. et Wolpert, R. (1988). *The Likelihood Principle*, volume 9 dans *IMS Lecture Notes–Monograph Series*. IMS, Hayward California, second édition.

Berger, J. et Yang, R. (1994). Noninformative priors and Bayesian testing for the AR(1) model. *Econometric Theory*, 10, 461–482.

Bergman, N., Doucet, A., et Gordon, N. (2001). Optimal estimation and Cramér-Rao bounds for partial non-Gaussian state-space models. *Ann. Inst. Statist. Math.*, 52(1), 97–112.

Bernardo, J. (1979). Reference posterior distributions for Bayesian inference (with discussion). *J. Royal Statist. Soc. Series B*, 41, 113–147.

Bernardo, J. (1980). A Bayesian analysis of classical hypothesis testing. In Bernardo, J., DeGroot, M. H., Lindley, D. V., et Smith, A., éditeurs, *Bayesian Statistics*. Oxford University Press.

Bernardo, J. et Smith, A. (1994). *Bayesian Theory*. John Wiley, New York.

Berry, D. et Stangl, D. (1996). *Bayesian Biostatistics*. Marcel Dekker, New York.

Bertrand, J. (1889). *Calcul des Probabilités*. Gauthier-Villars, Paris.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Royal Statist. Soc. Series B*, 36, 192–326.

Besag, J. (1986). On the statistical analysis of dirty pictures. *J. Royal Statist. Soc. Series B*, 48, 259–302.

Besag, J. (2000). Markov chain Monte Carlo for statistical inference. Technical Report 9, University of Washington, Center for Statistics and the Social Sciences.

Besag, J. et Green, P. (1993). Spatial statistics and Bayesian computation (with discussion). *J. Royal Statist. Soc. Series B*, 55, 25–38.

Best, N., Cowles, M., et Vines, K. (1995). CODA : Convergence diagnosis and output analysis software for Gibbs sampling output, version 0.30. Technical report, MRC Biostatistics Unit, University of Cambridge.

Bhattacharya, R. et Rao, R. (1986). *Normal approximations and asymptotic expansions*. John Wiley, New York, seconde édition.

Bickel, P. (1981). Minimax estimation of the mean of a normal distribution when the parameter space is restricted. *Ann. Statist.*, 9, 1301–1309.

Bickel, P. et Ghosh, J. (1990). A decomposition for the likelihood ratio statistic and the Bartlett correction : a Bayesian argument. *Ann. Statist.*, 18, 1070–1090.

Billingsley, P. (1985). *Probability and Measure*. John Wiley, New York, seconde édition.

Billingsley, P. (1995). *Probability and Measure*. John Wiley, New York, third édition.

Billio, M., Monfort, A., et Robert, C. (1998). The simulated likelihood ratio method. Technical Report 9821, CREST, INSEE, Paris.

Billio, M., Monfort, A., et Robert, C. (1999). Bayesian estimation of switching ARMA models. *J. Econometrics*, 93, 229–255.

Bilodeau, M. (1988). On the simultaneous estimation of scale parameters. *Canad. J. Statist.*, 14, 169–174.

Binder, D. (1978). Bayesian cluster analysis (with discussion). *Biometrika*, 65, 31–38.

Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). *J. American Statist. Assoc.*, 57, 269–326.

Bjørnstad, J. (1990). Predictive likelihood : a review. *Statist. Science*, 5, 242–265.

Blackwell, D. et Girshick, M. (1954). *Theory of Games and Statistical Decisions*. John Wiley, New York.

Blattberg, R. et George, E. (1991). Shrinkage estimation of price and promotion elasticities : seemingly unrelated equations. *J. American Statist. Assoc.*, 86, 304–315.

Blyth, C. (1951). On minimax statistical decisions procedures and their admissibility. *Ann. Mathemat. Statist.*, 22, 22–42.

Blyth, C. (1972a). Discussion of Robert, Hwang and Strawderman. *J. American Statist. Assoc.*, 88, 72–74.

Blyth, C. (1972b). Some probability paradoxes in choice from among random alternatives (with discussion). *J. American Statist. Assoc.*, 67, 366–387.

Blyth, C. et Pathak, P. (1985). Does an estimator distribution suffice ? In Cam, L. L. et Olshen, A., éditeurs, *Proc. Berkeley Conf. in Honor of J. Neyman and J. Kiefer*, volume 1. Wadsworth, Belmont, California.

Blyth, C.R.and Hutchinson, D. (1961). Tables of Neyman-shortest confidence interval for the binomial parameter. *Biometrika*, 47, 381–391.

Bock, M. (1985). Minimax estimators that shift towards a hypersphere for location of spherically symmetric distributions. *J. Multiv. Analysis*, 9, 579–588.

Bock, M. (1988). Shrinkage estimators : pseudo-Bayes rules for normal vectors. In Gupta, S. et Berger, J., éditeurs, *Statistical Decision Theory and Related Topics*, volume 4, pages 281–297. Springer-Verlag, New York.

Bock, M. et Robert, C. (1985). Bayes estimators with respect to uniform distributions on spheres (i) : the empirical Bayes approach. Unpublished notes.

Bohning, D. (1999). *Computer-Assisted Analysis of Mixtures and Applications*. Chapman and Hall, New York.

Bondar, J. (1987). How much improvement can a shrinkage estimator give. In McNeill, I. et Umphreys, G., éditeurs, *Foundations of Statistical Inference*. Reidel, Dordrecht.

Bondar, J. et Milnes, P. (1981). Amenability : a survey for statistical applications of Hunt-Stein and related conditions on groups. *Z. Wahrsch. verw. Gebiete*, 57, 103–128.

Boole, G. (1854). *A Investigation of the Laws of Thought*. Walton and Maberly, London.

Bose, S. (1992). Some properties of posterior Pitman closeness. *Comm. Statist.*, 20, 3697–3412.

Box, G. et Jenkins, G. (1976). *Time Series Analysis : Forecasting and Control*. Holden-Bay, San Francisco.

Box, G. et Muller, M. (1958). A note on the generation of random normal variates. *Ann. Mathemat. Statist.*, 29, 610–611.

Box, G. et Tiao, G. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, Massachusetts.

Brandwein, A. et Strawderman, W. (1980). Minimax estimators of location parameters for spherically symmetric distributions with concave loss. *Ann. Statist.*, 8, 279–284.

Brandwein, A. et Strawderman, W. (1990). Stein estimation : the spherically symmetric case. *Statist. Science*, 5, 356–569.

Brandwein, A., Strawderman, W., et Ralescu, S. (1992). Stein estimation for non-normal spherically symmetric location families in three dimensions. *J. Multiv. Analysis*, 42, 35–50.

Brewster, J. et Zidek, J. (1974). Improving on equivariant estimators. *Ann. Statist.*, 2, 21–38.

Brockwell, P. et Davis, P. (1998). *Introduction to Time Series and Forecasting*. Springer Texts in Statistics. Springer-Verlag, New York.

Broniatowski, M., Celeux, G., et Diebolt, J. (1983). Reconnaissance de mélanges de densités par un algorithme d'apprentissage probabiliste. In Diday, E., éditeur, *Data Analysis and Informatics*. North-Holland, Amsterdam.

Brown, L. (1966). On the admissibility of invariant estimators of one or more location parameters. *Ann. Mathemat. Statist.*, 37, 1087–1136.

Brown, L. (1967). The conditional level of Student's t-test. *Ann. Mathemat. Statist.*, 38, 1068–1071.

Brown, L. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary-value problems. *Ann. Mathemat. Statist.*, 42, 855–903.

Brown, L. (1975). Estimation with incompletely specified loss functions. *J. American Statist. Assoc.*, 70, 417–426.

Brown, L. (1976). Notes on statistical decision theory. Technical report, Ithaca, New York.

Brown, L. (1978). A contribution to Kiefer's theory of conditional confidence procedures. *Ann. Statist.*, 6, 59–71.

Brown, L. (1980). Examples of Berger's phenomenon in the estimation of independent normal means. *Ann. Statist.*, 9, 1289–1300.

Brown, L. (1981). A complete class theorem for statistical problems with finite sample spaces. *Ann. Statist.*, 9, 1289–1300.

Brown, L. (1986a). An ancilarity paradox which appears in multiple linear regression (with discussion). *Ann. Statist.*, 18, 471–538.

Brown, L. (1986b). *Foundations of Exponential Families*, volume 6 dans *IMS lecture notes Monograph Series*. Hayward California.

Brown, L. (1988). The differential inequality of a statistical estimation problem. In Gupta, S. et Berger, J., éditeurs, *Statistical Decision Theory and Related Topics*, volume 4. Springer-Verlag, New York.

Brown, L. (1993). Minimaxity, more or less. In Gupta, S. et Berger, J., éditeurs, *Statistical Decision Theory and Related Topics*, volume 5, pages 1–18. Springer-Verlag, New York.

Brown, L. (2000). An essay on statistical decision theory. *J. American Statist. Assoc.*, 95, 1277–1282.

Brown, L. et Farrell, R. (1985). Complete class theorems for estimation of multivariate Poisson means and related problems. *Ann. Statist.*, 8, 377–398.

Brown, L. et Hwang, J. (1982). A unified admissibility proof. In Gupta, S. et Berger, J., éditeurs, *Statistical Decision Theory and Related Topics*, volume 3, pages 205–230. Academic Press, New York.

Brown, L. et Hwang, J. (1989). Universal domination and stochastic domination : U-admissibility and u-inadmissibility of the least-squares estimator. *Ann. Mathemat. Statist.*, 17, 252–267.

Buehler, R. (1959). Some validity criteria for statistical inference. *Ann. Statist.*, 30, 845–863.

Cappé, O., Guillin, A., Marin, J., et Robert, C. (2004). Population Monte Carlo. *J. Comput. Graph. Statist.*, 13(4), 907–929.

Cappé, O., Moulines, E., et Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer-Verlag, New York.

Cappé, O. et Robert, C. (2000). MCMC : Ten years and still running ! *J. American Statist. Assoc.*, 95(4), 1282–1286.

Cappé, O., Robert, C., et Rydén, T. (2003). Reversible jump, birth-and-death, and more general continuous time MCMC samplers. *J. Royal Statist. Soc. Series B*, 65(3), 679–700.

Carlin, B. et Chib, S. (1995). Bayesian model choice through Markov chain Monte Carlo. *J. Roy. Statist. Soc. (Ser. B)*, 57(3), 473–484.

Carlin, B. et Gelfand, A. (1991). A sample reuse method for accurate parametric empirical Bayes confidence intervals. *J. Royal Statist. Soc. Series B*, 53, 189–200.

Carlin, B. et Louis, T. (2000a). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall, New York.

Carlin, B. et Louis, T. (2000b). Empirical Bayes : Past, present and future. *J. American Statist. Assoc.*, 95, 1286–1290.

Carlin, B. et Louis, T. (2001). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall, New York, seconde édition.

Caron, N. (1994). *Approches alternatives d'une théorie non-informative des tests bayésiens*. Thèse de doctorat, Université de Rouen, Dépt. de Mathématique.

Carota, C., Parmigiani, G., et Polson, N. (1996). Diagnostic measures for model criticism. *J. American Statist. Assoc.*, 91, 753–762.

Carter, G. et Rolph, J. (1974). Empirical Bayes methods applied to estimating fire alarm probabilities. *J. American Statist. Assoc.*, 69, 882–885.

Casella, G. (1980). Minimax ridge regression estimation. *Ann. Statist.*, 8, 1036–1056.

Casella, G. (1985a). Condition number and minimax ridge regression estimation. *J. American Statist. Assoc.*, 80, 753–758.

Casella, G. (1985b). An introduction to empirical Bayes data analysis. *The American Statistician*, 39, 83–87.

Casella, G. (1987). Conditionally acceptable recentered set estimators. *Ann. Statist.*, 15, 1364–1371.

Casella, G. (1990). Estimators with nondecreasing risks : application of a chi-squared identity. *Statist. Prob. Lett.*, 10, 107–109.

Casella, G. (1992). Conditional inference for confidence sets. In Ghosh, M. et Pathak, P., éditeurs, *Current Issues in Statistical Inference : Essays in Honor of D. Basu*, volume 17 dans *IMS lectures notes Monograph Series*, pages 1–12. Hayward, California.

Casella, G. (1996). Statistical theory and Monte Carlo algorithms (with discussion). *TEST*, 5, 249–344.

Casella, G. et Berger, R. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *J. American Statist. Assoc.*, 82, 106–111.

Casella, G. et Berger, R. (2001). *Statistical Inference*. Wadsworth, Belmont, CA, seconde édition.

Casella, G. et George, E. (1992). An introduction to Gibbs sampling. *Ann. Mathemat. Statist.*, 46, 167–174.

Casella, G. et Hwang, J. (1983). Empirical Bayes confidence sets for the mean of a multivariate normal distribution. *J. American Statist. Assoc.*, 78, 688–698.

Casella, G. et Hwang, J. (1987). Employing vague prior information in the construction of confidence sets. *J. Multiv. Analysis*, 21, 79–104.

Casella, G., Hwang, J., et Robert, C. (1993a). *Loss function for set estimation*, pages 237–252. Springer-Verlag, New York.

Casella, G., Hwang, J., et Robert, C. (1993b). A paradox in decision-theoretic set estimation. *Statist. Sinica*, 3, 141–155.

Casella, G., Robert, C., et Wells, M. (2000). Mixture models, latent variables and partitioned importance sampling. Technical Report 2000-03, CREST, INSEE, Paris.

Casella, G. et Strawderman, W. (1981). Estimating a bounded normal mean. *Ann. Statist.*, 4, 283–300.

Casella, G. et Wells, M. (1993). Discussion of Robert, Hwang and Strawderman. *J. American Statist. Assoc.*, 88, 70–71.

Castledine, B. (1981). A Bayesian analysis of multiple-recapture sampling for a closed population. *Biometrika*, 67, 197–210.

Castro, I., Conigliani, C., et O'Hagan, A. (1999). Bayesian assessment of goodness of fit against nonparametric alternatives (with discussion). In Racugno, W., éditeur, *Model Selection*, Collana Atti di Congressi. Pitagora Editrice, Bologna.

Celeux, G. et Diebolt, J. (1990). Une version de type recuit simulé de l'algorithme EM. *Comptes Rendus Acad. Sciences Paris*, 310, 119–124.

Celeux, G., Forbes, F., Robert, C., et Titterington, D. (2005). Deviance criteria in missing data models. *Bayesian Analysis*. (To appear.).

Celeux, G., Hurn, M., et Robert, C. (2000). Computational and inferential difficulties with mixture posterior distribution. *J. American Statist. Assoc.*, 95(3), 957–979.

Cellier, D., Fourdrinier, D., et Robert, C. (1989). Robust shrinkage estimators of the location parameter for elliptically symmetric distributions. *J. Multiv. Analysis*, 29, 39–52.

Chamberlain, G. (2000). *Econometrics*. Springer-Verlag, New York.

Chen, M. et Shao, Q. (1997). On Monte Carlo methods for estimating ratios of normalizing constants. *Ann. Statist.*, 25, 1563–1594.

Chen, M., Shao, Q., et Ibrahim, J. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer-Verlag, New York.

Chernoff, H. et Yahav, J. (1977). A subset selection employing a new criterion. In Gupta, S. et Moore, D., éditeurs, *Statistical Decision Theory and Related Topics*. Academic Press, New York, New York.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *J. American Statist. Assoc.*, 90, 1313–1321.

Chib, S. et Greenberg, E. (1994). Bayes inference in regression models with ARMAi (p, q) errors. *J. Econometrics*, 64, 183–206.

Chickering, D. et Heckerman, D. (2000). A comparison of scientific and engineering criteria forBayesian model selection. *Statist. Comp.*, 10, 55–62.

Chow, M. (1987). A complete class theorem for estimating a non-centrality parameter. *Ann. Statist.*, 15, 869–876.

Chow, M. et Hwang, J. (1990). The comparison of estimators for the non-centrality of a chi-square distribution. Technical report, Cornell University, Dept. of Mathematics, New York.

Chow, Y. et Teicher, H. (1988). *Probability Theory*. Springer-Verlag, New York.

Chrystal, G. (1891). On some fundamental principles in the theory of probability. *Trans. Actuarial Soc. Edinburgh*, 2, 421–439.

Clarke, B. et Wasserman, L. (1993). Noninformative priors and nuisance parameters. *J. American Statist. Assoc.*, 88, 1427–1432.

Clevenson, M. et Zidek, J. (1975). Simultaneous estimation of the mean of independant Poisson laws. *J. American Statist. Assoc.*, 70, 698–705.

Clyde, M. (1999). Bayesian model averaging and model search strategies. In Bernardo, J., Dawid, A., Berger, J., et Smith, A., éditeurs, *Bayesian Statistics*, volume 6, pages 157–185. Oxford University Press, Oxford.

Cohen, A. (1972). Improved confidence intervals for the variance of a normal distribution. *J. American Statist. Assoc.*, 67, 382–387.

Cohen, A. et Sackrowitz, H. (1984). Decision theoretic results for vector risks with applications. *Statist. Decisions, Supplement Issue*, 1, 159–176.

Cohen, A. et Strawderman, W. (1973). Admissible confidence intervals and point estimators for translation or scale parameters. *Ann. Statist.*, 1, 545–550.

Congdon, P. (2001). *Bayesian Statistical Modelling*. John Wiley, New York.

Congdon, P. (2003). *Applied Bayesian Modelling*. John Wiley, New York.

Cowell, R., Dawid, A., Lauritzen, S., et Spiegelhalter, D. (1999). *Probabilistic Networks and Expert Systems*. Springer-Verlag, New York.

Cox, D. R. (1958). Some problems connected with statistical inference. *Ann. Statist.*, 29, 357–425.

Cox, D. R. (1990). Role of models in statistical analysis. *Statist. Science*, 5, 169–174.

Cox, D. R. et Hinkley, D. (1987). *Theoretical Statistics*. Chapman and Hall, New York.

Cox, D. R. et Reid, N. (1987). Orthogonal parameters and approximate conditional inference (with discussion). *J. Royal Statist. Soc. Series B*, 49, 1–18.

Crawford, S., DeGroot, M., Kadane, J., et Small, M. (1992). Modelling lake-chemistry distributions : Approximate Bayesian methods for estimating a finite-mixture model. *Technometrics*, 34, 441–453.

Cressie, N. (1993). *Spatial Statistics*. John Wiley, New York.

Dacunha-Castelle, D. et Gassiat, E. (1999). Testing the order of a model using locally conic parametrization : population mixtures and stationary ARMA processes. *Ann. Statist.*, 27, 1178–1209.

Dalal, S. et Hall (1983). Approximating priors by mixtures of natural conjugate priors. *J. Royal Statist. Soc. Series B*, 45, 278–286.

Dale, A. (1991). *A History of Inverse Probability*. Springer-Verlag, New York.

Damien, P., Wakefield, J., et Walker, S. (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *J. Royal Statist. Soc. Series B*, 61(2), 331–344.

Darroch, J. (1958). The multiple-recapture census. I : Estimation of a closed population. *Biometrika*, 45, 343–359.

Das Gupta, A. (1958). Admissibility in the Gamma distribution : two examples. *Sankhya, Ser. A*, 46, 395–407.

Das Gupta, A. et Sinha, B. (1986). Estimation in the multiparameter exponential family : admissibility and inadmissibility results. *Statist. Decisions*, 4, 101–130.

Das Gupta, A. et Studden, W. (1988). Frequentist behavior of smallest volume robust Bayes confidence sets. Technical Report Technical Report 94-20, Purdue University, West Lafayette, Indiana.

Datta, G. et Ghosh, M. (1995a). On priors providing frequentist validity for Bayesian inference. *Biometrika*, 82, 37–45.

Datta, G. et Ghosh, M. (1995b). Some remarks on noninformative priors. *J. American Statist. Assoc.*, 90, 1357–1363.

Dawid, A. (1984). Probability forecasts. Technical report, University College, London.

Dawid, A. (1992). Frequential analysis, stochastic complexity and Bayesian inference. In Berger, J., Bernardo, J., Dawid, A., et Smith, A., éditeurs, *Bayesian Statistics 4*, volume 4. Oxford University Press, Oxford.

Dawid, A. (2002). Discussion of "Bayesian measures of model complexity and fit" by Spiegelhalter *et al. J. Royal Statist. Soc. Series B*, 64, 583–640.

Dawid, A. et Lauritzen, S. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.*, 21, 1272–1317.

Dawid, A., Stone, N., et Zidek, J. (1973). Marginalization paradoxes in Bayesian and structural inference (with discussion). *J. Royal Statist. Soc. Series B*, 35, 189–233.

de Finetti, B. (1972). *Probability, Induction and Statistics*. John Wiley, New York.

de Finetti, B. (1974). *Theory of Probability*, volume 1. John Wiley, New York.

Deely, J. et Gupta, S. (1968). On the property of subset selection per order. *Sankhya, Ser. A*, 30, 37–50.

Deely, J. et Lindley, D. (1981). Bayes empirical Bayes. *J. American Statist. Assoc.*, 76, 833–841.

DeGroot, M. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York.

DeGroot, M. (1973). Doing what comes naturally : Interpreting a tail area as a posterior probability or as a likelihood ratio. *J. American Statist. Assoc.*, 68, 966–969.

DeGroot, M. et Fienberg, S. (1983). The comparison and evaluation of forecasters. *The Statistician*, 32, 12–22.

Delampady, M. (1989). Lower bounds on Bayes factors for invariant testing situations. *J. Multiv. Analysis*, 28, 227–246.

Dellaportas, P. et Forster, J. (1996). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. Technical report, Univ. of Southampton.

Dempster, A. (1968). A generalization of Bayesian inference (with discussion). *J. Royal Statist. Soc. Series B*, 30, 205–248.

Dempster, A., Laird, N., et Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Royal Statist. Soc. Series B*, 39, 1–38.

der Meulen B., V. (1992). *Assessing weights of evidence for discussing classical statistical hypotheses*. PhD thesis, University of Groningen.

DeRobertis, L. et Hartigan, J. (1981). Bayesian inference using intervals of measures. *Ann. Statist.*, 9, 235–244.

Dette, H. et Studden, W. (1997). *The Theory of Canonical Moments with Applications in Statistics, Probability and Analysis*. John Wiley, New York.

Devroye, L. (1985). *Non-Uniform Random Variate Generation*. Springer-Verlag, New York.

Devroye, L. et Györfi, L. (1985). *Nonparametric Density Estimation : the L_1 View*. John Wiley, New York.

Dey, D., Müller, P., et Sinha, D. (1998). *Practical Nonparametrics and Semiparametrics in Bayesian Statistical Inference*, volume 133 dans *Lecture Notes in Statistics*. Springer-Verlag, New York.

Diaconis, P. et Freedman, D. (1986). On the consistency of Bayes estimates. *Ann. Statist.*, 14, 1–26.

Diaconis, P. et Kemperman, J. (1996). Some new tools for Dirichlet priors (with discussion). In Bernardo, J., Berger, J., Dawid, A., Lindley, D., et Smith, A., éditeurs, *Bayesian Statistics 5*, pages 97–106. Oxford University Press, Oxford.

Diaconis, P. et Mosteller, F. (1989). Methods for studying coincidences. *J. American Statist. Assoc.*, 84, 853–861.

Diaconis, P. et Ylvisaker, D. (1979). Conjugate priors for exponential families. *Ann. Statist.*, 7, 269–281.

Diaconis, P. et Ylvisaker, D. (1985). Quantifying prior opinion. In Bernardo, J., DeGroot, M., Lindley, D., et Smith, A., éditeurs, *Bayesian Statistics 2*, pages 163–175. North-Holland, Amsterdam.

Diaconis, P. et Zabell, S. (1991). Closed form summation for classical distribution variations on a theme of De Moivre. *Statist. Science*, 6, 284–302.

DiCiccio, T. J. et Stern, S. (1993). On Bartlett adjustments for approximate Bayesian inference. *Biometrika*, 80, 731–740.

DiCiccio, T. J. et Stern, S. (1994). Frequentist and Bayesian Bartlett correction of test statistics based on adjusted profile likelihoods. *J. Royal Statist. Soc. Series B*, 56, 397–408.

Dickey, J. (1968). Three multidimensional integral identities with Bayesian applications. *Ann. Statist.*, 39, 1615–1627.

Diebolt, J. et Robert, C. (1994). Estimation of finite mixture distributions by Bayesian sampling. *J. Royal Statist. Soc. Series B*, 56, 363–375.

Douc, R., Guillin, A., Marin, J.-M., et Robert, C. (2005). Convergence of adaptive sampling schemes. Technical Report 2005-6, University Paris Dauphine.

Doucet, A., de Freitas, N., et Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, New York.

Doucet, A., Godsill, S., et Robert, C. (2002). Marginal maximum a posteriori estimation using Markov chain Monte Carlo. *Statistics and Computing*, 12, 77–84.

Draper, D. (1995). Assessment and propagation of model uncertainty. *J. Royal Statist. Soc. Series B*, 57, 45–98.

Drèze, J.H.and Morales, J. (1976a). Bayesian full information analysis of the simultaneous equations. *J. American Statist. Assoc.*, 71, 919–923.

Drèze, J.H.and Morales, J. (1976b). Bayesian regression analysis using poly-t densities. *J. American Statist. Assoc.*, 71, 919–923.

Dudewicz, E. et Koo, J. (1982). *The Complete Categorized Guide to Statistical Selection and Ranking Procedures*. American Science Press, Columbus, Ohio.

Dumouchel, W. et Harris, J. (1983). Bayes methods for combining the results of cancer studies in human and other species (with discussion). *J. American Statist. Assoc.*, 78, 293–315.

Dupuis, J. (1995a). *Analyse stochastique bayésienne de modèles de capture-recapture*. Thèse de doctorat, Université Paris VI.

Dupuis, J. (1995b). Bayesian estimation of movement probabilities in open populations using hidden Markov chains. *Biometrika*, 82(4), 761–772.

Dupuis, J. et Robert, C. (2001). Bayesian variable selection in qualitative models by Kullback-Leibler projections. *J. Statist. Plann. Inference*, 111, 77–94.

Dynkin, E. (1951). Necessary and sufficient statistics for a family of probability distributions. *Selected Transl. Math. Statist. Prob.*, 1, 23–41.

Eaton, M. (1982). *Multivariate Statistics*. John Wiley, New York.

Eaton, M. (1986). A characterization of spherical distributions. *J. Multivariate Anal.*, 20, 272–276.

Eaton, M. (1989). *Group Invariance Applications in Statistics*, volume 1 dans *Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics, Hayward, California.

Eaton, M. (1992). A statistical dyptich : Admissible inferences - recurrence of symmetric Markov chains. *Ann. Statist.*, 20, 1147–1179.

Eaton, M. (1999). Markov chain conditions for admissibility in estimation problems with quadratic loss. Technical Report Report PN1 R9904, Centrum voor Wiskunde en Informatica, Amsterdam.

Eberly, L. E. et Casella, G. (1999). Comparison of Bayesian credible intervals in hierarchical models. Technical report, Division of Biostatistics University of Minnesota.

Efron, B. (1975). Biased versus unbiased estimation. *Adv. in Math.*, 16, 259–277.

Efron, B. (1982). The Jacknife, the Bootstrap and other resampling plans. In *Regional Conference in Applied Mathematics*, volume 38. SIAM, Philadelphia.

Efron, B. (1992). Regression percentile using asymmetric squared error loss. *Statist. Sinica*, 1, 93–125.

Efron, B. et Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *J. American Statist. Assoc.*, 70, 311–319.

Efron, B. et Thisted, R. (1976). Estimating the number of species : How many words did Shakespeare know ? *Biometrika*, 63, 435–447.

Eichenauer, J. et Lehn, J. (1989). Gamma-minimax estimators for a bounded normal mean under squared error-loss. *Statist. Decisions*, 7, 37–62.

Engle, R. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50, 987–1008.

Escobar, M. (1989). *Estimating the means of several normal populations by estimating the distribution of the means*. PhD thesis, Yale University.

Escobar, M. et West, M. (1995). Bayesian prediction and density estimation. *J. American Statist. Assoc.*, 90, 577–588.

Evans, M., Fraser, D., et Monette, G. (1986). On principles and arguments to likelihood (with discussion). *Canadian J. Statist.*, 14, 181–199.

Fabius, J. (1964). Asymptotic behavior of Bayes estimates. *Ann. Mathemat. Statist.*, 34, 846–856.

Fan, K. et Anderson, T. (1990). *Statistical Inference in Elliptically Contoured and Related Distributions*. Allerton Press, New York.

Farrell, R. (1968a). On a necessary and sufficient condition for admissibility of estimators when strictly convex loss is used. *Ann. Mathemat. Statist.*, 38, 23–28.

Farrell, R. (1968b). Towards a theory of generalized Bayes tests. *Ann. Mathemat. Statist.*, 38, 1–22.

Feller, W. (1970). *An Introduction to Probability Theory and its Applications*, volume 1. John Wiley, New York.

Feller, W. (1971). *An Introduction to Probability Theory and its Applications*, volume 2. John Wiley, New York.

Ferguson, T. (1967). *Mathematical Statistics : a Decision-Theoretic Approach*. Academic Press, New York.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1, 209–230.

Ferguson, T. (1974). Prior distributions in spaces of probability measures. *Ann. Statist.*, 2, 615–629.

Fernandez, C. et Steel, M. (1999). On the dangers of modelling through continuous distributions : a Bayesian perspective (with discussion). In Bernardo, J., Berger, J., Dawid, A., et Smith, A., éditeurs, *Bayesian Statistics 6*, pages 213–238. Oxford University Press, Oxford.

Feyerabend, P. (1975). *Against Method*. New Left Books, London.

Field, C. et Ronchetti, E. (1990). *Small Sample Asymptotics*. IMS Lecture Notes - Monograph Series, Hayward, CA.

Fieller, E. (1954). Some problems in interval estimation. *J. Royal Statist. Soc. Series B*, 16, 175–185.

Fienberg, S. (2005). When did Bayesian statistics become Bayesian ? *Bayesian Analysis*. (To appear.).

Fishburn, P. (1988). *Non-Linear Preferences and Utility Theory*. Harvester Wheatsheaf, Brighton, Sussex.

Fisher, R. (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics*, 41, 155–160.

Fisher, R. (1922). On the mathematical foundations of theoretical Statistics. *Philos. Trans. Roy. Soc. London*, 222, 309–368.

Fisher, R. (1930). Inverse probability. *Proc. Cambridge Philos. Soc*, 26, 528–535.

Fisher, R. (1956). *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh.

Fisher, R. (1959). Mathematical probability in the natural sciences. *Technometrics*, 1, 21–29.

Fishman, G. (1996). *Monte Carlo*. Springer-Verlag, New York.

Fitzgerald, W., Godsill, S., Kokaram, A., et Stark, J. (1999). Bayesian methods in signal and image processing. In Bernardo, J., Berger, J., Dawid, A., et Smith, A., éditeurs, *Bayesian Statistics 6*, pages 239–254, Oxford. Oxford University Press.

Foster, D. et George, E. (1998). A simple ancillarity paradox. *Scand. J. Statist.*, 23, 233–242.

Fouley, J., San Cristobal, M., Gianola, D., et Im, S. (1992). Marginal likelehood and Bayesian approaches to the analysis of heterogeneous residual variances in mixed linear Gaussian models. *Comput. Statist. Data Anal.*, 13, 291–305.

Fourdrinier, D. et Robert, C. (1995). A note on empirical Bayes via entropy. *Statist. Prob. Letters*, 23(1), 35–44.

Fourdrinier, D., Strawderman, W., et Wells, M. (1998). On the construction of Bayes minimax estimators. *Ann. Statist.*, 26(2), 660–671.

Fourdrinier, D. et Wells, M. (1993). Risk comparison of variable selection rules. Technical report, Université de Rouen.

Fraisse, A., Raoult, J., Robert, C., et Roy, M. (1990). Une condition nécessaire d'admissibilité et ses conséquences sur les estimateurs à rétrécisseur de la moyenne d'une loi normale. *Canadian J. Statist.*, 18, 213–220.

Fraisse, A., Robert, C., et Roy, M. (1998). Semi-tail upper bounds for admissible estimators in exponential families with nuisance parameters. *Statistics & Decisions*, 16(2), 147–162.

Francq, C. et Zakoïan, J. (2001). Stationarity of multivariate Markov-switching ARMA models. *J. Econometrics*, 102(2), 339–364.

Fraser, D., Monette, G., et Ng, K. (1984). Marginalization,likelihood and structural models. In Krishnaiah, P., éditeur, *Multivariate Analysis*, volume 6. North-Holland, Amsterdam.

Gatsonis, C., Hodges, J., Kass, R., et McCulloch, R. (1997). *Case Studies in Bayesian Statistics II*, volume 121 dans *Lecture Notes in Statistics*. Springer-Verlag, New York.

Gatsonis, C., Hodges, J., Kass, R., et Singpurwalla, N. (1993). *Case Studies in Bayesian Statistics*, volume 83 dans *Lecture Notes in Statistics*. Springer-Verlag, New York.

Gatsonis, C., Hodges, J., Kass, R., et Singpurwalla, N. (1995). *Case Studies in Bayesian Statistics II*, volume 105 dans *Lecture Notes in Statistics*. Springer-Verlag, New York.

Gatsonis, C., Kass, R., Carlin, B., Carriquiry, A., Gelman, A., et Verdinelli, I. (1999). *Case Studies in Bayesian Statistics IV*, volume 140. Springer-Verlag, New York.

Gatsonis, C., MacGibbon, K., et Strawderman, W. (1987). On the estimation of a truncated normal mean. *Statist. Prob. Letters*, 6, 21–30.

Gauss, C. (1810). *Méthode des Moindres Carrés. Mémoire sur la Combination des Observations*. Mallet-Bachelier, Paris. Transl. J. Bertrand.

Geisser, S. et Cornfield, J. (1963). Posterior distributions for multivariate normal parameters. *J. Royal Statist. Soc. Series B*, 25, 368–376.

Gelfand, A. (1996). Model determination using sampling-based methods. In W.R. Gilks, S. R. et Spiegelhalter, D., éditeurs, *Markov Chain Monte Carlo in Practice*, pages 145–162. Chapman and Hall, New York.

Gelfand, A. (2000). Gibbs sampling. *J. American Statist. Assoc.*, 95, 1300–1304.

Gelfand, A. et Dey, D. (1994). Bayesian model choice : asymptotics and exact calculations. *J. Roy. Statist. Soc. (Ser. B)*, 56, 501–514.

Gelfand, A. et Smith, A. (1990). Sampling based approaches to calculating marginal densities. *J. American Statist. Assoc.*, 85, 398–409.

Gelfand, A., Smith, A., et Lee, T. (1992). Bayesian analysis of constrained parameters and truncated data problems using Gibbs sampling. *J. American Statist. Assoc.*, 87, 523–532.

Gelman, A. (1996). Inference and monitoring convergence. In Gilks, W., Richardson, S., et Spiegelhalter, D., éditeurs, *Markov chain Monte Carlo in Practice*, pages 131–143. Chapman and Hall, New York.

Gelman, A., Carlin, J., Stern, H., et Rubin, D. (2003). *Bayesian Data Analysis*. Chapman and Hall, New York, second édition.

Gelman, A. et Meng, X. (1998). Simulating normalizing constants : From importance sampling to bridge sampling to path sampling. *Statist. Science*, 13, 163–185.

Gelman, A. et Rubin, D. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statist. Science*, 7, 457–511.

Geman, S. et Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6, 721–741.

Genest, C. et Zidek, J. (1986). Combining probability distributions : A critique and an annotated bibliography. *Statist. Science*, 1, 114–135.

Gentle, J. E. (1998). *Random Number Generation and Monte Carlo Methods*. Springer-Verlag, New York.

George, E. (1986a). Combining minimax shrinkage estimators. *J. American Statist. Assoc.*, 81, 437–445.

George, E. (1986b). Minimax multiple shrinkage estimators. *Ann. Statist.*, 14, 188–205.

George, E. et Casella, G. (1994). Empirical Bayes confidence estimation. *Statist. Sinica*, 4(2), 617–638.

George, E. et Foster, D. (1999). Empirical Bayes variable selection. In Racugno, W., éditeur, *Model Choice*. Pitagora Editrice, Bologna.

George, E. et McCulloch, R. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, 7, 339–374.

George, E. et Robert, C. (1992). Calculating Bayes estimates for capture-recapture models. *Biometrika*, 79(4), 677–683.

Geweke, J. (1991). Efficient simulation from the multivariate normal and student t-distributions subject to linear constraints. *Computer Sciences and Statistics : Proc. 23d Symp. Interface*.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments (with discussion). In Bernardo, J., Berger, J., Dawid, A., et Smith, A., éditeurs, *Bayesian Statistics 4*, pages 169–193. Oxford University Press, Oxford.

Geweke, J. (1999). Using simulation methods for Bayesian econometric models : inference, development and communication (with discussion). *Econometric Reviews*.

Geyer, C. (1992). Practical Monte Carlo Markov chain (with discussion). *Statist. Science*, 7, 473–511.

Geyer, C. (1995). Conditioning in Markov chain Monte Carlo. *J. Comput. Graph. Statis.*, 4, 148–154.

Geyer, C. (1996). Estimation and optimization of functions. In Gilks, W., Richardson, S., et Spiegelhalter, D., éditeurs, *Markov chain Monte Carlo in Practice*, pages 241–258. Chapman and Hall, New York.

Geyer, C. et Thompson, E. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. Royal Statist. Soc. Series B*, 54, 657–699.

Ghosh, M., Carlin, B. P., et Srivastiva, M. S. (1995). Probability matching priors for linear calibration. *TEST*, 4, 333–357.

Ghosh, M., Hwang, J., et Tsui, K. (1983). Construction of improved estimators in multiparameter estimation for discrete exponential families (with discussion). *Ann. Mathemat. Statist.*, 11, 351–376.

Ghosh, M., Keating, J., et Sen, P. (1993). Discussion of Robert, Hwang and Strawderman. *J. American Statist. Assoc.*, 88, 63–66.

Ghosh, M. et Mukerjee, R. (1992a). Bayesian and frequentist Bartlett corrections for likelihood ratio tests. *J. Royal Statist. Soc. Series B*, 56, 396–408.

Ghosh, M. et Mukerjee, R. (1992b). Noninformative priors (with discussion). In Bernardo, J., Berger, J., Dawid, A., et Smith, A., éditeurs, *Bayesian Statistics*. Oxford University Press, Oxford.

Ghosh, M. et Mukerjee, R. (1993). Frequentist validity of highest posterior density regions in the multiparameter case. *Ann. Inst. Statist. Math.*, 45, 293–302.

Ghosh, M. et Saleh, A. (1989). Empirical Bayes subset estimation in regression models. *Statist. Decisions*, 7, 15–35.

Ghosh, M. et Sen, P. (1989). Median unbiasedness and Pitman closeness. *J. American Statist. Assoc.*, 84, 1089–1091.

Ghosh, M. et Yang, M. (1996). Noninformative priors for the two sample normal problem. *Test*, 5, 145–157.

Gibbons, J., Olkin, I., et Sobel, M. (1977). *Selecting and Ordering Populations*. John Wiley, New York.

Gilks, W., Best, N., et Tan, K. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *Applied Statist. (Ser. C)*, 44, 455–472.

Gilks, W., Clayton, D., Spiegelhalter, D., Best, N., McNeil, A., Sharples, L., et Kirby, A. (1993). Modelling complexity : applications of Gibbs sampling in medicine. *J. Royal Statist. Soc. Series B*, 55, 39–52.

Gilks, W., Richardson, S., et Spiegelhalter, D., éditeurs (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, New York, London.

Gilks, W. et Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.*, 41, 337–348.

Gill, J. (2002). *Bayesian Methods : A Social and Behavioral Sciences Approach*. CRC Press.

Gill, W. et Levit, B. (1995). Applications of the Van Trees inequality : a Bayesian Cramér-Rao bound. *Bernouilli*, 1, 59–79.

Giudici, P. et Green, P. (1999). Decomposable graphical Gaussian model determination. *Biometrika*, 86(4), 785–801.

Givens, G., Smith, D., et Tweedie, R. (1997). Publication bias in meta-analysis : a Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. *StatSci*, 12, 221–250.

Gleick, J. (1987). *Chaos*. Penguin, New York.

Gleser, L. et Healy, J. (1976). Estimating the mean of a normal distribution with known coefficient of variation. *J. American Statist. Assoc.*, 71, 977–981.

Gleser, L. et Hwang, J. (1987). The non-existence of $100(1-\alpha)\%$ confidence sets of finite expected diameters in errors-in-variable and related models. *Ann. Statist.*, 15, 1351–1362.

Goel, P. et Rubin, H. (1977). On selecting a subset containing the best population-a Bayesian approach. *Ann. Statist.*, 5, 969–983.

Goldstein, M. et Smith, A. (1974). Ridge-type estimators for regression analysis. *J. Royal Statist. Soc. Series B*, 36, 284–219.

Good, I. (1952). Rational decisions. *J. Royal Statist. Soc. Series B*, 14, 107–114.

Good, I. (1973). The probabilistic explication of evidence, causality, explanation and utility. In Godambe, V. et Sprott, D., éditeurs, *Foundations of Statistical Inference*. Rinehart and Winston, Toronto.

Good, I. (1975). Estimation methods for two-way contingency tables. *J. Royal Statist. Soc. Series B*, 37, 23–37.

Good, I. (1980). Some history of the hierarchical Bayesian methodology. In Bernardo, J., DeGroot, M., Lindley, D., et Smith, A., éditeurs, *Bayesian Statistics 2*. North-Holland, Amsterdam.

Good, I. (1983). *Good Thinking : The Foundations of Probability and Its Applications*. University of Minnesota Press, Minneapolis.

Gouriéroux, C. (1997). *ARCH Model*. Springer-Verlag, New York.

Gouriéroux, C. et Monfort, A. (1996). *Statistics and Econometric Models*. Cambridge University Press, Cambridge.

Goutis, C. (1990). Ranges of posterior measures for some classes of priors with specified moments. Technical Report 70, University College London, London.

Goutis, C. (1994). Ranges of posterior measures for some classes of priors with specified moments. *International Statistical Review*, 62(2), 245–256.

Goutis, C. et Casella, G. (1991). Improved invariant confidence intervals for a normal variance. *Ann. Statist.*, 19, 2015–2031.

Goutis, C. et Casella, G. (1992). Increasing the confidence in student's t-interval. *Ann. Statist.*, 20(3), 1501–1513.

Goutis, C., Casella, G., et Wells, M. (1996). Assessing evidence in multiple hypotheses. *J. American Statist. Assoc.*, 91, 1268–1277.

Goutis, C. et Robert, C. (1998). Model choice in generalized linear models : a Bayesian approach via Kullback–Leibler projections. *Biometrika*, 85, 29–37.

Green, P. (1995). Reversible jump MCMC computation and Bayesian model determination. *Biometrika*, 82(4), 711–732.

Grenander, U. et Miller, M. (1994). Representations of knowledge in complex systems (with discussion). *J. Royal Statist. Soc. Series B*, 56, 549–603.

Gruet, M., Philippe, A., et Robert, C. (1999). MCMC control spreadsheets for exponential mixture estimation. *J. Comput. Graph. Statist.*, 8, 298–317.

Guihenneuc-Jouyaux, C., Richardson, S., et Lasserre, V. (1998). Convergence assessment in latent variable models : application to longitudinal modelling of a marker of HIV progression. In Robert, C., éditeur, *Discretization and MCMC Convergence Assessment*, volume 135 dans *Lecture Notes in Statistics*, chapter 7, pages 147–160. Springer-Verlag, New York.

Gupta, S. (1965). On multiple decision (selection and ranking) rules. *Tech*, 7, 222–245.

Gupta, S. et Panchapakesan, S. (1979). *Multiple Decision Procedures*. John Wiley, New York.

Gutmann, S. (1982). Stein's paradox is impossible in problems with finite sample space. *Ann. Statist.*, 10, 1017–1020.

Hadjicostas, P. et Berry, S. (1999). Improper and proper posteriors with improper priors in a Poisson-gamma hierarchical model. *Test*, 8, 147–166.

Haff, L. et Johnstone, R. (1986). The superharmonic condition for simultaneous estimation of means in exponential families. *Canadian J. Statist.*, 14, 43–54.

Hàjek, B. et Sidàk, Z. (1968). *Theory of Rank Test*. Academic Press, New York.

Hald, A. (1998). *An History of Mathematical Statistics*. John Wiley, New York.

Haldane, J. (1931). A note on inverse probability. *Proc. Cambridge Philos. Soc.*, 28, 55–61.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, New York.

Hamilton, J. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2), 357–384.

Hammersley, J. (1974). Discussion of Besag's paper. *J. Royal Statist. Soc. Series B*, 36, 230–231.

Hansen, M. et Yu, B. (2000). Model selection and minimum description length principle. *J. American Statist. Assoc.* (To appear.).

Hartigan, J. A. (1983). *Bayes Theory*. Springer-Verlag, New York.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their application. *Biometrika*, 57, 97–109.

Heath, D. et Sudderth, W. (1989). Coherent inference from improper priors and from finitely additive priors. *Ann. Statist.*, 17, 907–919.

Heidelberger, P. et Welch, P. (1983). A spectral method for confidence interval generation and run length control in simulations. *Comm. Assoc. Comput. Machinery*, 24, 233–245.

Heitjan, D. et Rubin, D. (1991). Ignorability and coarse data. *Ann. Statist.*, 19, 2244–2253.

Helland, I. (1999). Statistical inference under a fixed symmetry group. Technical report, Dept. of Mathematics and Statistics, University of Oslo.

Hesterberg, T. (1998). Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37, 185–194.

Hills, S. et Smith, A. (1992). Parametrization issues in Bayesian inference. In Bernardo, J., Berger, J., Dawid, A., et Smith, A., éditeurs, *Bayesian Statistics 4*, pages 641–649. Oxford University Press, Oxford.

Hinkley, D. (1997). Discussion of "unified frequentist and Bayesian testing of a precise hypothesis". *Statist. Science*, 12, 155–156.

Hjort, N. (1996). Bayesian approaches to non- and semiparametric density estimation (with discussion). In Bernardo, J., Berger, J., Dawid, A., Lindley, D., et Smith, A., éditeurs, *Bayesian Statistics 5*, pages 223–253. Oxford University Press, Oxford.

Hoaglin, D., Mosteller, F., et Tukey, J. (1996). *Exploring Data Tables, Trends, and Shapes*. John Wiley, New York.

Hobert, J. (2000a). Stability relationships among the Gibbs sampler and its subchains. Technical report, University of Florida.

Hobert, J. et Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear models. *J. American Statist. Assoc.*, 91, 1461–1473.

Hobert, J. et Casella, G. (1998). Functional compatibility, Markov chains, and Gibbs sampling with improper posteriors. *J. Comput. Graph. Statist.*, 7, 42–60.

Hobert, J. et Robert, C. (1999). Eaton's Markov chain, its conjugate partner and p-admissibility. *Ann. Statist.*, 27, 361–373.

Hobert, J. et Robert, C. (2004). Moralizing perfect sampling. *Ann. Applied Prob.*, 14(3), 1295–1305.

Hobert, J. P. (2000b). Hierarchical models : a current computational perspective. *J. American Statist. Assoc.*, 95, 1312–1316.

Hoerl, A. et Kennard, R. (1970). Ridge regression : biased estimators for non-orthogonal problems. *Technometrics*, 12, 55–67.

Holmes, C., Denison, D., Mallick, B., et Smith, A. (2002). *Bayesian Methods for Nonlinear Classification and Regression*. John Wiley, New York.

Hora, R. et Buehler, R. (1966). Fiducial theory and invariant estimation. *Ann. Statist.*, 37, 361–379.

Huber, P. (1964a). Robust estimation of a location parameter. *Ann. Statist.*, 35, 73–101.

Huber, P. (1964b). Robust statistics : a review. *Ann. Statist.*, 67, 1041–1067.

Huerta, G. et West, M. (1999). Priors and component structures in autoregressive time series models. *J. Royal Statist. Soc. Series B*, 61(4), 881–899.

Hui, S. et Berger, J. (1983). Empirical Bayes estimation of rates in longitudinal studies. *J. American Statist. Assoc.*, 78, 753–760.

Hwang, J. (1982a). Improving upon standard estimators in discrete exponential families with applications to Poisson and negative binomial cases. *Ann. Statist.*, 10, 857–867.

Hwang, J. (1982b). Semi-tail upper bounds on the class of admissible estimators in discrete exponential families, with applications to Poisson and negative binomial distributions. *Ann. Statist.*, 10, 1137–1147.

Hwang, J. (1985). Universal domination and stochastic domination : decision theory simultaneously under a broad class of loss functions. *Ann. Statist.*, 13, 295–314.

Hwang, J. et Brown, L. (1991). Estimated confidence under the validity constraint. *Ann. Statist.*, 19, 1964–1977.

Hwang, J. et Casella, G. (1982). Minimax confidence sets for the mean of a multivariate normal distribution. *Ann. Statist.*, 10, 868–881.

Hwang, J. et Casella, G. (1984). Improved set estimators for a multivariate normal mean. *Statist. Decisions*, 1, 3–16. Supplement Issue.

Hwang, J., Casella, G., Wells, M., et Farrel, R. (1992). Estimation of accuracy in testing. *Ann. Statist.*, 20, 490–509.

Hwang, J. et Chen, J. (1986). Improved confidence sets for the coefficients of a linear model with spherically symmetric errors. *Ann. Statist.*, 14, 444–460.

Hwang, J. et Pemantle, R. (1994). Evaluation of estimators of statistical significance under a class of proper loss functions. *Statist. Decisions*, 15, 103–128.

Hwang, J. et Ullah, A. (1994). Confidence sets recentered at James-Stein estimators-a surprise concerning the unknown variance case. *Econometrics*, 60(1-2), 145–156.

Ibragimov, I. et Has'minskii, R. (1981). *Statistical Estimation. Asymptotic Theory*. Springer-Verlag.

Jacquier, E., Polson, N., et Rossi, P. (1994). Bayesian analysis of stochastic volatility models (with discussion). *J. Business Economic Stat.*, 12, 371–417.

James, W. et Stein, C. (1961). Estimation with quadratic loss. In *Proc. Fourth Berkeley Symp. Math. Statist. Probab.*, volume 1, pages 361–380. University of California Press.

Jaynes, E. (1980). Marginalization and prior probabilities. In Zellner, A., éditeur, *Bayesian Analysis in Econometrics and Statistics*. North-Holland, Amsterdam.

Jaynes, E. (1983). *Papers on Probability, Statistics and Statistical Physics*. R.D. Rosencrantz, Reidel, Dordrecht.

Jefferys, W. et Berger, J. (1992). Ockham's razor and Bayesian analysis. *American Scientist*, 80, 64–72.

Jeffreys, H. (1939). *Theory of Probability*. Oxford University Press, Oxford.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London(Ser. A)*, 186, 453–461.

Jeffreys, H. (1961). *Theory of Probability (3rd edition)*. Oxford University Press, Oxford.

Johnson, B. (1971). On the admissible estimators for certain fixed sample binomial problems. *Ann. Statist.*, 41, 1579–1587.

Johnson, D. et Hoeting, J. (2003). Autoregressive models for capture-recapture data : A Bayesian approach. *Biometrics*, 59(2), 341–350.

Johnson, D. et Lindley, D. (1995). Bayesian inference given data "significant at α" : tests of point-null hypotheses. *Theory and Decision*, 38(1), 51–60.

Johnson, N. et Kotz, S. (1969-1972). *Distributions in Statistics (4 vols.)*. John Wiley, New York.

Johnson, N., Kotz, S., et Balakrishnan, N. (1994). *Continuous Univariate Distributions*, volume 1. John Wiley, New York, seconde édition.

Johnson, N., Kotz, S., et Balakrishnan, N. (1995). *Continuous Univariate Distributions*, volume 2. John Wiley, New York, seconde édition.

Johnstone, I. (1984). Admissibility, difference equations, and recurrence in estimating a Poisson mean. *Ann. Statist.*, 12, 1173–1198.

Johnstone, I. (1998). On the inadmissibility of Stein's unbiased estimate of loss. In Gupta, S. et Berger, J., éditeurs, *Statistical Decision Theory and Related Topics*. Springer-Verlag, New York.

Johnstone, I. et MacGibbon, B. (1992). Minimax estimation of a constrained Poisson vector. *Ann. Statist.*, 20, 807–831.

Jones, M. (1987). Randomly choosing parameters from the stationarity and invertibility region of autoregressive-moving average models. *Applied Statistics (Series C)*, 38, 134–138.

Joshi, V. (1967a). Admissibility of the usual confidence set for the mean of a multivariate normal population. *Ann. Statist.*, 38, 1868–1875.

Joshi, V. (1967b). The censoring concept and the likelihood principle. *J. Statist. Plann. Inference*, 26, 109–111.

Judge, G. et Bock, M. (1978). *Implications of Pre-Test and Stein Rule Estimators in Econometrics*. North-Holland, Amsterdam.

Kadane, J. et Chuang, D. (1978). Stable decision problems. *Ann. Statist.*, 6, 1095–1111.

Kariya, T., Giri, N., et Perron, F. (1988). Invariant estimation of mean vector μ of $\mathcal{N}(\mu, \Sigma)$ with $\mu'\Sigma^{-1}\mu = 1$ or $\Sigma^{-1/2}\mu = C$ or $\Sigma = \delta^2\mu\mu' I$. *J. Multiv. Analysis*, 27, 270–283.

Karlin, S. (1958). Admissibility for estimation with quadratic loss. *Ann. Statist.*, 29, 406–436.

Karlin, S. et Rubin, H. (1956). The theory of decision procedures for distributions with monotone likelihood ratio. *Ann. Statist.*, 27, 272–299.

Kass, R. (1989). The geometry of asymptotic inference. *Statist. Science*, 4, 188–234.

Kass, R. et Raftery, A. (1995). Bayes factor and model uncertainty. *J. American Statist. Assoc.*, 90, 773–795.

Kass, R. et Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *J. American Statist. Assoc.*, 87, 717–726.

Kass, R. et Wasserman, L. (1996). Formal rules of selecting prior distributions : a review and annotated bibliography. *J. American Statist. Assoc.*, 91, 343–1370.

Keating, J. et Mason, R. (1988). James-Stein estimation from an alternative perspective. *Amer. Statist.*, 42, 160–164.

Keeney, R. et Raiffa, H. (1976). *Decisions with Multiple Objectives*, volume 42. J. Wiley, New York.

Kelker, D. (1970). Distribution theory of spherical distributions and a location-scale parameter generalization. *Sankhya (Ser. A)*, 32, 419–430.

Kempthorne, P. (1988). Controlling risks under different loss functions : the compromise decision problem. *Ann. Statist.*, 16, 1594–1608.

Kendall, M. et Stuart, A. (1979). *Inference and Relationships*. The Advanced Theory of Statistics. Macmillan, New York, 4th edition édition.

Keynes, J. (1921). *A Treatise on Probability*. Macmillan, London.

Kiefer, J. (1957). Invariance, minimax sequential estimation and continuous time-processes. *Ann. Mathemat. Statist.*, 28, 573–601.

Kiefer, J. (1977). Conditional confidence statements and confidence estimators (theory and methods). *J. American Statist. Assoc.*, 72, 789–827.

Kiiveri, H. et Speed, T. (1982). Structural analysis of multivariate data : A review. In Leinhardt, S., éditeur, *Sociological Methodology*, pages 209–289. Jossey Bass, San Francisco.

Kirby, A. J. et Spiegelhalter, D. J. (1994). Statistical modelling for the precursors of cervical cancer. In Lange, N., éditeur, *Case Studies in Biometry*. John Wiley, New York.

Kleibergen, F. et Van Dijk, H. (1993). Non-stationarity in GARCH models : a Bayesian analysis. *J. of Appl. Econometrics*, 8, 41–61.

Kontkanen, P., Myllymäki, P., Silander, T., Tirri, H., et Grünwald, P. (2000). On predictive distributions and Bayesian networks. *Statist. Comp.*, 10, 39–54.

Koopman, B. (1936). On distributions admitting a sufficient statistic. *Trans. Amer. Math. Soc.*, 39, 399–409.

Kubokawa, T. (1991). An approach to improving James-Stein estimator. *J. Multiv. Analysis*, 36, 121–126.

Kubokawa, T., Morita, S., Makita, S., et Nagakura, K. (1993a). Estimation of the variance and its applications. *J. Statist. Plann. Inference*, 35, 319–333.

Kubokawa, T. et Robert, C. (1994). New perspectives on linear calibration. *J. Multiv. Analysis*, 51, 178–200.

Kubokawa, T., Robert, C., et Saleh, A. (1991). Robust estimation of common regression coefficients under spherical symmetry. *Ann. Inst. Statist. Math.*, 43, 677–688.

Kubokawa, T., Robert, C., et Saleh, A. (1992). Empirical Bayes estimation of the covariance matrix of a normal distribution with unknown mean under an entropy loss. *Sankhya*, 54, 402–410. Ser. A.

Kubokawa, T., Robert, C., et Saleh, A. (1993b). Estimation of noncentrality parameters. *Canadian J. Statist.*, 21, 54–58.

Lad, F. (1996). *Operational Subjective Statistical Methods : a Mathematical, Philosophical and Historical Introduction*. John Wiley, New York.

Laird, N. et Louis, T. (1987). Confidence intervals based on bootstrap samples. *J. American Statist. Assoc.*, 82, 739–750.

Laplace, P. (1773). Mémoire sur la probabilité des causes par les événements. *Mémoires de l'Académie Royale des Sciences présentés par divers savants*, 6, 621–656. Reprinted in Laplace (1878).

Laplace, P. (1786). Sur les naissances, les mariages et les morts à Paris depuis 1771 jusqu'à 1784 et dans toute l'étendue de la France, pendant les années 1781 et 1782. *Mémoires de l'Académie Royale des Sciences présentés par divers savants*, 11, 35–46. Reprinted in Laplace (1878).

Laplace, P. (1795). *Essai Philosophique sur les Probabilités*. Epistémé. Christian Bourgeois, Paris. Reprinted in 1986.

Lauritzen, S. (1996). *Graphical Models*. Oxford University Press, Oxford.

Lavielle, M. et Moulines, E. (1997). On a stochastic approximation version of the em algorithm. *Statist. Comput.*, 7, 229–236.

Lavine, M. (1992). Some aspects of Pólya tree distributions for statistical modeling. *Ann. Statist.*, 22, 1222–1235.

Lawley, D. (1956). A general method for approximating to the distribution of the likelihood ratio criteria. *Biometrika*, 43, 295–303.

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, New York.

Le Cam, L. (1990). Maximum likelihood : an introduction. *Statist. Science*, 58, 153–172.

Legendre, A. (1805). *Nouvelles Méthodes pour la Détermination des Orbites des Comètes*. Courcier, Paris.

Lehmann, E. (1983). *Theory of Point Estimation*. John Wiley, New York.

Lehmann, E. (1986). *Testing Statistical Hypotheses*. John Wiley, New York.

Lehmann, E. (1990). Model specification. *Statist. Science*, 5, 160–168.

Lehmann, E. et Casella, G. (1998). *Theory of Point Estimation*. Springer-Verlag, New York, revised édition.

Lenk, P. (1999). Bayesian inference for semiparametric regression using a Fourier representation. *J. Royal Statist. Soc. Series B*, 61, 863–879.

Leonard, T. (1982). Comments on Lejeune and Faulkenberry. *J. American Statist. Assoc.*, 77, 657–658.

Letac, G. et Mora, M. (1990). Natural real exponential families with cubic variance functions. *Ann. Statist.*, 18, 1–37.

Lindley, D. (1957). A statistical paradox. *Biometrika*, 44, 187–192.

Lindley, D. (1961). The use of prior probability distributions in statistical inference and decision. In *Proc. Fourth Berkeley Symp. Math. Statist. Probab*, volume 1, pages 453–468. University of California Press.

Lindley, D. (1962). Discussion of professor Stein's paper 'confidence sets for the mean of a multivariate normal distribution'. *J. Royal Statist. Soc. Series B*, 24, 265–296.

Lindley, D. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint*, volume Parts 1 and 2. Cambridge University Press, Cambridge.

Lindley, D. (1971). *Bayesian Statistics : A Review*. SIAM, Philadelphia.

Lindley, D. (1980). Approximate Bayesian methods. In Bernardo, J., DeGroot, M., Lindley, D., et Smith, A., éditeurs, *Bayesian Statistics 2*. North-Holland, Amsterdam.

Lindley, D. (1985). *Making Decisions*. John Wiley, New York.

Lindley, D. (1990). The present position in Bayesian statistics (with discussion). *Statist. Science*, 5(1), 44–89.

Lindley, D. et Smith, A. (1972). Bayes stimates for the linear model. *J. Royal Statist. Soc. Series B*, 34, 1–41.

Liseo, B. (1993). Elimination of nuisance parameters with reference priors. *Biometrika*, 80(2), 295–304.

Liu, J., Wong, W., et Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and sampling schemes. *Biometrika*, 81, 27–40.

Liu, J., Wong, W., et Kong, A. (1995). Correlation structure and convergence rate of the Gibbs sampler with various scans. *J. Royal Statist. Soc. Series B*, 57, 157–169.

Liu, J. et Wu, Y. (1999). Parameter expansion scheme for data augmentation. *J. American Statist. Assoc.*, 94, 1264–1274.

Louis, T. (1997). Discussion of "unified frequentist and Bayesian testing of a precise hypothesis". *Statist. Science*, 12, 152–155.

Lu, K. et Berger, J. (1989a). Estimated confidence procedures for multivariate normal means. *J. Statist. Plann. Inference*, 23, 1–19.

Lu, K. et Berger, J. (1989b). Estimation of normal means : frequentist estimators of loss. *Ann. Statist.*, 17, 890–907.

Maatta, J. et Casella, G. (1990). Developments in decision theoretic variance estimation (with discussion). *Statist. Science*, 5, 90–120.

Machina, G. (1982). Expected utility analysis without the independence axiom. *Econometrica*, 50, 277–323.

Machina, G. (1987). Choice under uncertainty : problems solved and unsolved. *Econom. Perspectives*, 1, 121–154.

MacLachlan, G. et Basford, K. (1987). *Mixture Models*. Marcel Dekker, New York.

MacLachlan, G. et Krishnan, T. (1997). *The EM Algorithm and Extensions*. John Wiley, New York.

Madigan, D. et Raftery, A. (1991). Model selection and accounting for model uncertainty in graphical models using Occam's window. Technical report 213, University of Washington.

Madigan, D. et Raftery, A. (1995). Bayesian graphical models for discrete data. *Int. Statist. Rev.*, 63, 215–232.

Madigan, D. et York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review*, 63, 215–232.

Marin, J., Mengersen, K., et Robert, C. (2004). Bayesian modelling and inference on mixtures of distributions. In Rao, C. et Dey, D., éditeurs, *Handbook of Statistics*, volume 25 (To appear.). Springer-Verlag, New York.

Maritz, J. et Lwin, T. (1989). *Empirical Bayes Methods*. Chapman and Hall, New York, seconde édition.

Marsaglia, G. et Zaman, A. (1993). The KISS generator. Technical report, Dept. of Statistics, Univ. of Florida.

McCullagh, P. et Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall, New York.

McCulloch, C. et Rossi, P. (1992). Bayes factors for nonlinear hypotheses and likelihood distributions. *Biometrika*, 79, 663–676.

McKeganey, N., Barnard, M., Leyland, A., Coote, I., et Follet, E. (1992). Female streetworking prostitutes and HIV infection in Glasgow. *British Medical Journal*, 305, 801–804.

Meng, X. et Van Dyk, D. (1997). The EM algorithm–an old folk-song sung to a new tune (with discussion). *J. Royal Statist. Soc. Series B*, 59, 511–568.

Meng, X. et Van Dyk, D. (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika*, 86, 301–320.

Meng, X. et Wong, W. (1996). Simulating ratios of normalizing constants via a simple identity : a theoretical exploration. *Statist. Sinica*, 6, 831–860.

Mengersen, K. et Robert, C. (1996). Testing for mixtures : A Bayesian entropic approach (with discussion). In Berger, J., Bernardo, J., Dawid, A., Lindley, D., et Smith, A., éditeurs, *Bayesian Statistics 5*, pages 255–276. Oxford University Press, Oxford.

Mengersen, K., Robert, C., et Guihenneuc-Jouyaux, C. (1999). MCMC convergence diagnostics : a "reviewww". In Berger, J., Bernardo, J., Dawid, A., Lindley, D., et Smith, A., éditeurs, *Bayesian Statistics 6*, pages 415–440. Oxford University Press, Oxford.

Mengersen, K. et Tweedie, R. (1995). Meta-analysis approaches to dose-response relationships with application in studies of lung cancer and passive smoking. *Statist. Medicine*, 14, 545–69.

Mengersen, K. et Tweedie, R. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.*, 24, 101–121.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., et Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21, 1087–1092.

Metropolis, N. et Ulam, S. (1949). The Monte Carlo method. *J. American Statist. Assoc.*, 44, 335–341.

Meyn, S. et Tweedie, R. (1993). *Markov Chains and Stochastic Stability*. Springer-Verlag, New York.

Meyn, S. et Tweedie, R. (1994). Computable bounds for convergence rates of Markov chains. *Ann. Appl. Prob.*, 4, 981–1011.

Mira, A., Møller, J., et Roberts, G. (2001). Perfect slice samplers. *J. Royal Statist. Soc. Series B*, 63, 583–606.

Monahan, J. (1984). A note on enforcing stationarity in autoregressive-moving average models. *Biometrika*, 71, 403–404.

Moors, J. (1981). Inadmissibilité of linearly invariant estimators in truncated parameter spaces. *J. American Statist. Assoc.*, 76, 910–915.

Morisson, D. (1979). Purchase intentions and purchase behavior. *J. Marketing*, 43, 65–74.

Morris, C. (1982). Natural exponential families with quadratic variance functions. *Ann. Statist.*, 10, 65–80.

Morris, C. (1983a). Natural exponential families with quadratic variance functions : statistical theory. *Ann. Statist.*, 11, 515–529.

Morris, C. (1983b). Parametric empirical Bayes inference : theory and applications. *J. American Statist. Assoc.*, 78, 47–65.

Mosteller, F. et Chalmers, T. (1992). Some progress and problems in meta-analysis of clinical trials. *Statist. Science*, 7, 227–236.

Mosteller, F. et Wallace, D. (1984). *Applied Bayesian and Classical Inference*. Springer-Verlag, New York.

Mukerjee, R. et Dey, D. (1993). Frequentist validity of posterior quantiles in the presence of a nuisance parameter : higher order asymptotics. *Biometrika*, 80, 499–505.

Müller, P. et Vidakovic, B. (1999). *Bayesian Inference in Wavelet-Based Models*, volume 141 dans *Lecture Notes in Statistics*. Springer-Verlag, New York.

Murphy, A.H.and Winkler, R. (1984). Probability forecasting in meteorology. *J. American Statist. Assoc.*, 79, 489–500.

Musio, M. et Racugno, W. (1999). Discussion of Fernandez and Steel's paper. In Bernardo, J., Berger, J., Dawid, A., et Smith, A., éditeurs, *Bayesian Statistics 6*, pages 231–233. Oxford University Press.

Nachbin, L. (1965). *The Haar Integral*. Van Nostrand, New York.

Naylor, J. et Smith, A. (1982). Application of a method for the efficient computation of posterior distributions. *Applied Statistics*, 31, 214–225.

Neal, R. (1999). *Bayesian Learning for Neural Networks*, volume 118 dans *Lecture Notes*. Springer-Verlag, New York.

Nelson, D. (1990). Stationarity and persistence in the GARCH(1,1) model. *Econometric Theory*, 6, 318–334.

Newton, M. et Raftery, A. (1994). Approximate Bayesian inference by the weighted likelihood boostrap (with discussion). *J. Royal Statist. Soc. Series B*, 56, 1–48.

Neyman, J. (1934). On the two different aspects of the representative method : The method of stratified sampling and the method of purposive selection. *J. Royal Statist. Soc. Series B*, 97, 558–625.

Neyman, J. (1937). "Smooth" test for goodness of fit. *Skand. Aktvariebidokr*, 20, 150–199.

Neyman, J. et Pearson, E. (1933a). On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. Royal Soc. Ser. A*, 231, 289–337.

Neyman, J. et Pearson, E. (1933b). The testing of statistical hypotheses in relation to probabilities a priori. *Proc. Cambridge Philos. Soc.*, 24, 492–510.

Neyman, J. et Scott, E. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16, 1–32.

Novick, M. et Hall, W. (1965). A Bayesian indifference procedure. *J. American Statist. Assoc.*, 60, 1104–1117.

O'Hagan, A. (1994). *Bayesian Inference*. Numero 2B dans Kendall's Advanced Theory of Statistics. Chapman and Hall, New York.

O'Hagan, A. (1995). Fractional Bayes factors for model comparisons. *J. Royal Statist. Soc. Series B*, 57, 99–138.

O'Hagan, A. (1997). Properties of intrinsic and fractional Bayes factors. *Test*, 6, 101–118.

O'Hagan, A. et Forster, J. (2002). *Bayesian Inference*. Numero 2B dans Kendall's Advanced Theory of Statistics. Chapman and Hall, New York, seconde édition.

O'Hagan, A.and Berger, J. (1988). Ranges of posterior probabilities for quasi-unimodal priors with specified quantiles. *J. American Statist. Assoc.*, 83, 503–508.

Olkin, I., Petkau, A., et Zidek, J. (1981). A comparison of n estimators for the binomial distribution. *J. American Statist. Assoc.*, 76, 637–642.

Olver, F. (1974). *Asymptotics and Special Functions*. Academic Press, New York.

Osborne, C. (1991). Statistical calibration : a review. *International Statistical Review*, 59, 309–336.

Owen, A. et Zhou, Y. (2000). Safe and effective importance sampling. *J. American Statist. Assoc.*, 95, 135–143.

Parent, E., Bobée, B., Hubert, P., et Miquel, J. (1998). Statistical and Bayesian methods in hydrological sciences. In *Selected Proceedings from the UNESCO conference in honnor of Pr. Bernier*. Unesco. IHP-V Technical Documents in Hydrology N°20.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA.

Pearson, K. (1894). Contribution to the mathematical theory of evolution. *Proc. Trans. Roy. Soc. A*, 185, 71–110.

Peddada, S. et Khattree, R. (1986). On Pitman nearness and variance of estimators. *Comm. Stat.*, 15, 3005–3018.

Perk, W. (1947). Some observations on inverse probabilities including a new indifference rule. *J. Inst. Actuaries*, 73, 285–312.

Perron, F. et Giri, N. (1990). On the best equivariant estimator of mean of a multivariate normal population. *Multivariate Anal.*, 32(4), 1–16.

Petrone, S. et Wasserman, L. (2002). Consistency of Bernstein polynomial posteriors. *J. Royal Statist. Soc. Series B*, 64, 79100.

Pettit, L. (1992). Bayes factors for outlier models using the device of imaginary observations. *J. American Statist. Assoc.*, 87, 541–545.

Pfangzagl, J. (1968). A characterization of the one parameter exponential family by existence of uniformly most powerful tests. *Sankhya, Ser. A*, 30, 147–156.

Phillips, D. et Smith, A. (1996). Bayesian model comparison via jump diffusions. In Gilks, W., Richardson, S., et Spiegelhalter, D., éditeurs, *Markov chain Monte Carlo in Practice*, pages 215–240. Chapman and Hall, New York.

Phillips, P. (1991). Bayesian routes and unit roots : de rebus prioribus semper est disputandum. *J. Appl. Econometrics*, 6, 435–474.

Pierce, D. (1973). On some difficulties in a frequency theory of inference. *Ann. Statist.*, 1, 241–250.

Pitman, E. (1936). Sufficient statistics and intrinsic accuracy. *Proc. Cambridge Philos. Soc.*, 32, 567–579.

Pitman, E. (1937). The closest estimates of statistical parameters. *Proc. Cambridge Philos. Soc.*, 33, 212–222.

Pitman, E. (1939). The estimation of location and scale parameters of a continuous population of any given form. *Biometrika*, 30, 391–421.

Pitt, M. et Shephard, N. (1999). Filtering via simulation : Auxiliary particle filters. *J. American Statist. Assoc.*, 94(446), 590–599.

Plessis, B. (1989). Context dependent enhancements for digitized radiographs. Master's thesis, Dept. of Electrical Engineering, University of Ottawa.

Poincaré, H. (1902). *La Science et l'Hypothèse.* Flammarion, Paris. Réimpression par Champs, 1989.

Poirier, D. (1995). *Intermediate Statistics and Econometrics : a Comparative Approach.* Cambridge, Mass.

Pole, A., West, M., et Harrison, J. (1994). *Applied Bayesian Forecasting and Time Series Analysis.* Chapman and Hall, New York, New York.

Pollock, K. (1991). Modelling capture, recapture and removal statistics for estimation of demographic parameters for fish and wildlife populations : past, present and future. *J. American Statist. Assoc.*, 86, 225–238.

Popper, K. (1983). *Postface to the Logic of Scientific Discovery, Realism and Science.* Hutchinson, London.

Press, J. (1989). *Bayesian Statistics.* John Wiley, New York.

Qian, W. et Titterington, D. (1991). Estimation of parameters in hidden Markov models. *Phil. Trans. Roy. Soc. London A*, 337, 407–428.

Racugno, W. (1999). *Model Selection.* Collana Atti di Congressi. Pitagora Editrice, Bologna.

Raftery, A. (1988). Inference for the binomial n parameter hierarchical Bayes approach. *Biometrika*, 75, 355–363.

Raftery, A. (1996). Hypothesis testing and model selection. In W.R. Gilks, S. R. et Spiegelhalter, D., éditeurs, *Markov chain Monte Carlo in Practice*, pages 163–188. Chapman and Hall, New York.

Raftery, A. et Lewis, S. (1992a). How many iterations in the Gibbs sampler ? In Bernardo, J., Berger, J., Dawid, A., et Smith, A., éditeurs, *Bayesian Statistics 4*, pages 763–773. Oxford University Press, Oxford.

Raftery, A. et Lewis, S. (1992b). The number of iterations, convergence diagnostics and generic Metropolis algorithms. Technical report, Department of Statistics, Univ. of Washington, Seattle.

Raftery, A., Madigan, D., et Volinsky, C. (1996). Accounting for model uncertainty in survival analysis improves predictive performance (with discussion). In Berger, J., Bernardo, J., Dawid, A., Lindley, D., et Smith, A., éditeurs, *Bayesian Statistics 5*, pages 323–349. Oxford University Press, Oxford.

Raftery, A. et Richardson, S. (1995). Model selection for generalized linear models via GLIB, with application to epidemiology). In Berry, D. et Stangl, D., éditeurs, *Bayesian Biostatistics*. Marcel Dekker, New York.

Raiffa, H. (1968). *Decision Analysis : Introductory Lectures on Choices under Uncertainty*. Addison-Wesley, Reading, Mass.

Raiffa, H. et Schlaifer, R. (1961). Applied Statistical decision theory. Technical report, Division of Research, Graduate School of Business Administration, Harvard Univ.

Rao, C. (1980). Discussion of J. Berkson's paper 'Minimum chi-square, not maximum likelihood'. *Ann. Statist.*, 8, 482–485.

Rao, C. (1981). Some comments on the minimum mean square error as criterion of estimation. In Csörgo, M., Dawson, D., Rao, J., et Saleh, A., éditeurs, *Statistics and Related Topics*, pages 123–143. North Holland, Amsterdam.

Rao, C., Keating, J., et Mason, R. (1986). The Pitman nearness criterion and its determination. *Comm. Statist.-Theory Methods*, 15, 3173–3191.

Redner, R. et Walker, H. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.*, 26, 195–239.

Richard, J. (1973). *Posterior and Predictive Densities for Simultaneous Equation Models*. Springer-Verlag, Berlin.

Richard, J. et Tompa, H. (1980). On the evaluation of poly-t density functions. *Econometrics*, 12, 335–351.

Richardson, S. et Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Royal Statist. Soc. Series B*, 59, 731–792.

Ripley, B. (1987). *Stochastic Simulation*. John Wiley, New York.

Ripley, B. (1992). Neural networks. In Barnorff-Nielsen, O., éditeur, *Networks and Chaos-Statistical and Probabilistic Aspects*. Chapman and Hall, New York.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Ann. Statist.*, 11, 416–431.

Rissanen, J. (1990). Complexity of models. In Zurek, W., éditeur, *Complexity, Entropy, and the Physics of Information*, volume 8. Addison-Wesley, Reading.

Robbins, H. (1951). Asymptotically subminimax solutions to compound statistical decision problems. In *Proc. Second Berkeley Symp. Math. Statist. Probab.*, volume 1. University of California Press.

Robbins, H. (1955). An empirical Bayes approach to statistics. In *Proc. Third Berkeley Symp. Math. Statist. Probab.*, volume 1. University of California Press.

Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. *Ann. Mathemat. Statist.*, 35, 1–20.

Robbins, H. (1983). Some thoughts on empirical Bayes estimation. *Ann. Statist.*, 11, 713–723.

Robert, C. (1990). Modified Bessel functions and their applications in probability and statistics. *Statist. Prob. Letters*, 9, 155–161.

Robert, C. (1991). Generalized inverse normal distributions. *Statist. Prob. Lett.*, 11, 37–41.

Robert, C. (1993a). A note on the Jeffreys-Lindley paradox. *Statist. Sinica*, 3, 601–608.

Robert, C. (1993b). Prior Feedback : A Bayesian approach to maximum likelihood estimation. *Comput. Statist.*, 8, 279–294.

Robert, C. (1995). Simulation of truncated Normal variables. *Statistics and Computing*, 5, 121–125.

Robert, C. (1996a). Inference in mixture models. In Gilks, W., Richardson, S., et Spiegelhalter, D., éditeurs, *Markov Chain Monte Carlo in Practice*, pages 441–464. Chapman and Hall, New York.

Robert, C. (1996b). Intrinsic loss functions. *Theory and Decision*, 40(2), 191–214.

Robert, C. (1998). Performances d'estimateurs à rétrécisseur en situation de multicolinéarité. *Ann. Eco. Stat.*, 10, 97–119.

Robert, C., Bock, M., et Casella, G. (1990). Bayes estimators associated with uniform distributions on spheres (ii) : the hierarchical Bayes approach. Technical Report Tech. Report BU-1002-M, Cornell University.

Robert, C. et Caron, N. (1996). Noninformative Bayesian testing and neutral Bayes factors. *TEST*, 5, 411–437.

Robert, C. et Casella, G. (1990). Improved confidence sets for spherically symmetric distributions. *J. Multivariate Anal.*, 32, 84–94.

Robert, C. et Casella, G. (1993). Improved confidence statements for the usual multivariate normal confidence set. In *Stat. Decision Theo. Rel. Topics V*, pages 351–368. Springer-Verlag, New York.

Robert, C. et Casella, G. (1994). Distance penalized losses for testing and confidence set evaluation. *Test*, 3(1), 163–182.

Robert, C. et Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer-Verlag, New York, première édition.

Robert, C. et Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer-Verlag, New York, seconde édition.

Robert, C., Celeux, G., et Diebolt, J. (1993a). Bayesian estimation of hidden Markov models : A stochastic implementation. *Statist. Prob. Letters*, 16, 77–83.

Robert, C. et Hwang, J. (1996). Maximum likelihood estimation under order constraints. *J. American Statist. Assoc.*, 91, 167–173.

Robert, C., Hwang, J., et Strawderman, W. (1993b). Is Pitman closeness a reasonable criterion ? (with discussion). *J. American Statist. Assoc.*, 88, 57–76.

Robert, C. et Mengersen, K. (1999). Reparametrization issues in mixture estimation and their bearings on the Gibbs sampler. *Comput. Statis. Data Ana.*, 29, 325–343.

Robert, C. et Reber, A. (1998). Bayesian modelling of a biopharmaceutical experiment with heterogeneous responses. *Sankhya B*, 60(1), 145–160.

Robert, C., Rydén, T., et Titterington, D. (1999a). Convergence controls for MCMC algorithms, with applications to hidden Markov chains. *J. Statist. Computat. Simulat.*, 64, 327–355.

Robert, C., Rydén, T., et Titterington, D. (1999b). Jump Markov chain Monte Carlo algorithms for Bayesian inference in hidden Markov models. *J. Royal Statist. Soc. Series B*, 62(1), 57–75.

Robert, C. et Soubiran, C. (1993). Estimation of a mixture model through Bayesian sampling and prior feedback. *TEST*, 2, 125–146.

Robert, C. et Titterington, M. (1998). Reparameterisation strategies for hidden Markov models and Bayesian approaches to maximum likelihood estimation. *Statistics and Computing*, 8(2), 145–158.

Roberts, G. et Rosenthal, J. (1998). Markov chain Monte Carlo : Some practical implications of theoretical results (with discussion). *Canadian J. Statist.*, 26, 5–32.

Roberts, G. et Sahu, S. (1997). Updating schemes, covariance structure, blocking and parametrisation for the Gibbs sampler. *J. Royal Statist. Soc. Series B*, 59, 291–318.

Roberts, G. et Tweedie, R. (2005). *Understanding MCMC*. Springer-Verlag, New York.

Robertson, T., Wright, F., et Dykstra, R. (1988). *Order Restricted Statistical Inference*. John Wiley, New York.

Robins, J. et Ritov, Y. (1997). A curse of dimensionality appropriate (coda) asymptotic for semiparametric models. *Statist. Medicine*, 16, 285–319.

Robins, J. et Wasserman, L. (2000). Conditioning, likelihood and concepts : A review of some foundational concepts. *J. American Statist. Assoc.*, 95, 1340–1346.

Robinson, G. (1979). Conditional properties of statistical procedures. *Ann. Statist.*, 7, 742–755.

Robinson, G. (1982). Behrens-Fisher problem. In Kotz, S. et Johnson, N., éditeurs, *Encyclopedia of Statistical Sciences*, volume 1, pages 205–209. wiley, New York.

Roeder, K. (1992). Density estimation with confidence sets exemplified by superclusters and voids in galaxies. *J. American Statist. Assoc.*, 85, 617–624.

Roeder, K. et Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *J. American Statist. Assoc.*, 92, 894–902.

Romano, J. et Siegel, A. (1986). *Counterexamples in Probability and Statistics*. Wadsworth, Belmont, CA.

Rousseau, J. (1997). *Performances fréquentistes des lois de référence et propriétés asymptotiques des procédures bayésiennes*. Thèse de doctorat, Université Paris VI.

Rousseau, J. (2000). Coverage properties of one-sided intervals in the discrete case and application to matching priors. *Ann. Inst. Statist. Math.*, 52(1), 28–42.

Rousseau, J. (2001). Asymptotic coverage of joint two-sided confidence intervals. *Scan. J. Statist.* (To appear.).

Rubin, D. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.*, 12, 1151–1172.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley, New York.

Rubin, D., Umbach, D., Shyu, S., et Castillo-Chavez, C. (1992). Using mark-recapture methodology to estimate the size of a population at risk for sexually transmitted diseases. *Statist. Medicine*, 11, 1533–1549.

Rubinstein, R. (1981). *Simulation and the Monte Carlo Method*. John Wiley, New York.

Rudin, W. (1976). *Principles of Real Analysis*. McGraw-Hill, New York.

Rue, H. (1995). New loss functions in Bayesian imaging. *J. American Statist. Assoc.*, 90, 900–908.

Rukhin, A. (1978). Universal Bayes estimators. *Ann. Statist.*, 6, 345–351.

Rukhin, A. (1988a). Estimated loss and admissible loss estimators. In Gupta, S. et Berger, J., éditeurs, *Statistical Decision Theory and Related Topics IV*, pages 409–420. Springer-Verlag, New York.

Rukhin, A. (1988b). Loss functions for loss estimations. *Ann. Statist.*, 16, 1262–1269.

Rukhin, A. (1995). Admissibility : Survey of a concept in progress. *International Statistical Review*, 63, 95–115.

Santner, T. et Duffy, D. (1989). *The Statistical Analysis of Discrete Data*. Springer-Verlag, New York.

Savage, L. (1954). *The Foundations of Statistical Inference*. John Wiley, New York.

Saxena, K. et Alam, K. (1982). Estimation of the non-centrality parameter of a chi-squared distribution. *Ann. Statist.*, 10, 1012–1016.

Schaafsma, W., Tolboom, J., et Van der Meulen, B. (1989). Discussing truth or falsity by computing a q-value. In Dodge, Y., éditeur, *Statistics, Data Analysis and Informatics*. North-Holland, Amsterdam.

Schervish, M. (1989). A general method for comparing probability assessors. *Ann. Statist.*, 17, 1856–1879.

Schervish, M. (1995). *Theory of Statistics*. Springer-Verlag, New York.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, 6, 461–464.

Seber, G. (1983). Capture-recapture methods. In Kotz, S. et Johnson, N., éditeurs, *Encyclopedia of Statistical Science*. John Wiley, New York.

Seber, G. (1986). A review of estimation of animal abundance. *Biometrika*, 42, 267–292.

Seidenfeld, T. (1987). Entropy and uncertainty. In MacNeill, I. et Umphrey, G., éditeurs, *Foundations of Statistical Inference*, pages 259–287. Reidel, Boston.

Seidenfeld, T. (1992). R.A. Fisher's fiducial argument and Bayes' theorem. *Statist. Science*, 7(3), 358–368.

Sen, P., Kubokawa, T., et Saleh, A. (1989). The Stein paradox in the sense of Pitman measure of closeness. *Ann. Statist.*, 17, 1375–1384.

Seneta, E. (1993). Lewis Carroll's pillow problems. *Statist. Science*, 8, 180–186.

Severini, T. (1991). On the relationship between Bayesian and non-Bayesian interval estimates. *J. Royal Statist. Soc. Series B*, 53, 611–618.

Shafer, G. (1996). *Art of Causal Conjecture*. MIT, Press, MIT, Cambridge.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Tech. J.*, 27, 379–423 et 623–656.

Shao, J. et Strawderman, W. (1996). Improving on the James-Stein positive-part estimator. *Statistica Sinica*, 6(1), 259–274.

Shinozaki, N. (1975). *Admissibility*. PhD thesis, Keio University.

Shinozaki, N. (1980). Estimation of a multivariate normal mean with a class of quadratic loss. *J. American Statist. Assoc.*, 75, 973–976.

Shinozaki, N. (1984). Simultaneous estimation of location parameters under quadratic loss. *Ann. Statist.*, 12, 322–335.

Shinozaki, N. (1990). Improved confidence sets for the mean of a multivariate normal distribution. *Ann. Inst. Statist. Math.*, 41, 331–346.

Shorrock, G. (1990). Improved confidence intervals for a normal variance. *Ann. Statist.*, 18, 972–980.

Sivaganesan, S. et Berger, J. (1989). Ranges of posterior measures for priors with unimodal contaminations. *Ann. Statist.*, 17, 868–889.

Smith, A. (1973). A general Bayesian linear model. *J. Royal Statist. Soc. Series B*, 35, 67–75.

Smith, A. (1984). Present position and potential developments : some personal view on Bayesian statistics. *J. Royal Statist. Soc. Series B*, 147, 245–259.

Smith, A. et Makov, U. (1978). A quasi-Bayes sequential procedure for mixtures. *J. Royal Statist. Soc. Series B*, 40, 106–112.

Smith, A., Sken, A., Shaw, J., Naylor, J., et Dransfield, M. (1985). The implementations of the Bayesian paradigm. *Comm. Statist.-Theory Methods*, 14, 1079–1102.

Smith, A. et Spiegelhalter, D. (1982). Bayes factors for linear and log-linear models with vague prior information. *J. Royal Statist. Soc. Series B*, 44, 377–387.

Smith, J. (1988). *Decision Analysis : A Bayesian Approach*. Chapman and Hall, New York.

Spiegelhalter, D., Best, N., et Carlin, B. (1998). Bayesian deviance, the effective number of parameters and the comparison of arbitrarily complex models. Technical report, MRC Biostatistics Unit.

Spiegelhalter, D. et Cowell, R. (1992). Learning in probabilistic expert systems. In Bernardo, J., Berger, J., Dawid, A., Lindley, D., et Smith, A., éditeurs, *Bayesian Statistics 4*, pages 447–460. Oxford University Press, Oxford.

Spiegelhalter, D., Dawid, A., Lauritzen, S., et Cowell, R. (1993). Bayesian analysis in expert systems (with discussion). *Statist. Science*, 8, 219–283.

Spiegelhalter, D. et Lauritzen, S. (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20, 579–605.

Spiegelhalter, D. et Smith, A. (1980). Bayes factors and choice criteria for linear models. *J. Royal Statist. Soc. Series B*, 42, 215–220.

Spiegelhalter, D., Thomas, A., Best, N., et Gilks, W. (1995a). BUGS : Bayesian inference using Gibbs sampling. Technical report, Medical Research Council Biostatistics Unit, Institute of Public Health, Cambridge Univ.

Spiegelhalter, D., Thomas, A., Best, N., et Gilks, W. (1995b). BUGS examples. Technical report, MRC Biostatistics Unit, Cambridge Univ.

Spiegelhalter, D., Thomas, A., Best, N., et Gilks, W. (1995c). BUGS examples. Technical report, MRC Biostatistics Unit, Cambridge Univ.

Spiegelhalter, D. J., Best, N., B.P., C., et Van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, 64, 583–640.

Srinivasan, C. (1981). Admissible generalized Bayes estimators and exterior boundary value problems. *Sankhya*, 43(Ser. A), 1–25.

Srivastava, M. et Bilodeau, M. (1988). Estimation of the MSE matrix of the Stein estimator. *Canadian J. Statist.*, 16, 153–159.

Stein, C. (1955a). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proc. Third Berkeley Symp. Math. Statist. Probab.*, volume 29, pages 197–206. University of California Press.

Stein, C. (1955b). A necessary and sufficient condition for admissibility. *Ann. Statist.*, 26, 518–522.

Stein, C. (1959). An examination of wide discrepancy between fiducial and confidence intervals. *Ann. Statist.*, 30, 877–880.

Stein, C. (1962a). Confidence sets for the mean of a multivariate normal distribution (with discussion). *J. Royal Statist. Soc. Series B*, 24, 573–610.

Stein, C. (1962b). A remark on the likelihood principle. *J. Royal Statist. Soc. Series B*, 125, 565–568.

Stein, C. (1965). Approximation of improper prior measures by prior probability measures. In *Bernoulli, Bayes, Laplace Anniversary Volume*. Springer-Verlag, New York.

Stein, C. (1973). Estimation of the mean of a multivariate distribution. In Hàjek, J., éditeur, *Proceedings of the Prague Symposium on Asymptotic Statistics*, pages 345–81. Charles University.

Stein, C. (1981). Estimation of the mean of a multivariate distribution. *Ann. Statist.*, 9, 1135–1151.

Stephens, M. (1997). *Bayesian methods for mixtures of normal distributions*. PhD thesis, Oxford University.

Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods. *Ann. Statist.*, 28, 40–74.

Steward, G. (1987). Collinearity and least-squares regression. *Statist. Science*, 2, 68–100.

Stigler, S. (1986). *The History of Statistics*. Belknap, Cambridge.

Stone, M. (1976). Strong inconsistency from uniform priors (with discussion). *J. American Statist. Assoc.*, 71, 114–125.

Strasser, H. (1985). *Mathematical Theory of Statistics*. W. de Gruyter, Berlin.

Strawderman, W. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *Ann. Mathemat. Statist.*, 42, 385–388.

Strawderman, W. (1974). Minimax estimation of location parameters for certain spherically symmetric distributions. *Multivariate Anal.*, 42, 255–264.

Strawderman, W. (2000). Minimaxity. *J. American Statist. Assoc.*, 95, 1364–1368.

Sweeting, T. (1985). Consistent prior distributions for transformed models. In Bernardo, J., DeGroot, M., Lindley, D., et Smith, A., éditeurs, *Bayesian Statistics 2*, pages 755–762. Elsevier Science Publishers, Amsterdam.

Tanner, M. et Wong, W. (1987). The calculation of posterior distributions by data augmentation. *J. American Statist. Assoc.*, 82, 528–550.

Thatcher, A. (1964). Relationships between Bayesian and confidence limits in prediction. *J. Royal Statist. Soc. Series B*, 26, 176–210.

Thisted, R. et Efron, B. (1987). Did Shakespeare write a newly-discovered poem ? *Biometrika*, 74, 445–468.

Thompson, P. (1989). *Admissibility of p-value rules*. PhD thesis, Dept. Statistics.

Tibshirani, R. (1989). Noninformative priors for one parameter of many. *Biometrika*, 76, 604–608.

Tierney, L. et Kadane, J. (1986). Accurate approximations for posterior moments and marginal densities. *J. American Statist. Assoc.*, 81, 82–86.

Tierney, L., Kass, R., et Kadane, J. (1989). Fully exponential Laplace approximations to expectations and variances of non-positive functions. *J. American Statist. Assoc.*, 84, 710–716.

Tierney, L. et Mira, A. (1998). Some adaptive Monte Carlo methods for Bayesian inference. *Statistics in Medicine*, 18, 2507–2515.

Titterington, D., Smith, A., et Makov, U. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley, New York.

Tong (1991). *Non-linear Time Series : a Dynamical Systems Approach*. Oxford Press University, Oxford.

Torrie, G. et Valleau, J. (1977). Nonphysical sampling distributions in Monte Carlo free-energy estimation : Umbrella sampling. *J. Comp. Phys.*, 23, 187–199.

Van Eeden, C. et Zidek, J. (1993). Group Bayes estimation of the exponential mean : a retrospective view of the Wald theory. In *Stat. Decision Theo. Rel. Topics V*, pages 35–50. Springer-Verlag, New York.

Venn, J. (1886). *The Logic of Chance*. Macmillan, London.

Verdinelli, I. et Wasserman, L. (1992). Bayesian analysis of outliers problems using the Gibbs sampler. *Statist. Comput.*, 1, 105–117.

Verdinelli, I. et Wasserman, L. (1998). Bayesian goodness-of-fit testing using infinite-dimensional exponential families. *Ann. Statist.*, 26, 1215–1241.

Villegas, C. (1977). On the representation of ignorance. *J. American Statist. Assoc.*, 72, 651–654.

Villegas, C. (1990). Bayesian inference in models with Euclidian structure. *J. American Statist. Assoc.*, 85, 1159–1164.

von Neumann, J. (1951). Various techniques used in connection with random digits. *J. Resources of the National Bureau of Standards–Applied Mathematics Series*, 12, 36–38.

von Neumann, J. et Morgenstern, O. (1947). *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, seconde édition.

Wakefield, J., Gelfand, A., et Smith, A. (1991). Efficient generation of random variates via the ratio-of-uniforms method. *Statistics and Computing*, 1, 129–133.

Wald, A. (1950). *Statistical Decision Functions*. John Wiley, New York.

Wallace, C. et Boulton, D. (1975). An invariant Bayes method for point estimation. *Classification Society Bulletin*, 3, 11–34.

Walley, P. (1991). *Statistical Reasoning with Imprecise Probability*. Chapman and Hall, New York.

Wasserman, L. (1992). Recent methodological advances in robust Bayesian inference. In Bernardo, J., Berger, J., Dawid, A., Lindley, D., et Smith, A., éditeurs, *Bayesian Statistics 4*, pages 483–490. Oxford University Press, Oxford.

Wasserman, L. (1999). Asymptotic inference for mixture models by using data-dependent priors. *J. Royal Statist. Soc. Series B*, 61(1), 159–180.

Welch, B. (1965). On comparisons between confidence point procedures in the case of a single parameter. *J. Royal Statist. Soc. Series B*, 27, 1–8.

Welch, B. et Peers, H. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *J. Royal Statist. Soc. Series B*, 25, 318–329.

Wells, M. (1992). Private communication.

West, M. et Harrison, J. (1998). *Bayesian Forecasting and Dynamic Models*. Springer-Verlag, New York, seconde édition.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. John Wiley, Chichester.

Wijsman, R. (1990). Invariant measures on groups and their use in statistics. In *IMS lecture notes-Monographs Series*. Hayward, California.

Wilkinson, G. (1977). On resolving the controversy in statistical inference. *J. Royal Statist. Soc. Series B*, 39, 119–171.

Wolter, W. (1986). Some coverage error models for census data. *J. American Statist. Assoc.*, 81, 338–346.

Yao, J. et Attali, J. (2000). On stability of nonlinear AR processes with Markov switching. *Applied Probability*, 32, 394–407.

Zabell, S. (1989). Fisher on the history of inverse probability. *Statist. Science*, 4, 247–263.

Zabell, S. (1992). Fisher and the fiducial argument. *Statist. Science*, 7, 369–387.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. John Wiley, New York.

Zellner, A. (1984). *Basic Issues in Econometrics*. University of Chicago Press, Chicago.

Zellner, A. (1986a). Bayesian estimation and prediction using asymmetric loss functions. *J. American Statist. Assoc.*, 81, 446–451.

Zellner, A. (1986b). On assessing prior distributions and Bayesian regression analysis with G–priors distributions. In Goel, P. et Zellner, A., éditeurs, *Bayesian Inference and Decision Techniques*, pages 233–243. Elsevier North-Holland, Amsterdam.

Zidek, J. (1965). A representation of Bayes invariant procedures in terms of Haar measure. *Ann. Inst. Statist. Math.*, 21, 291–308.

Zidek, J. (1970). Sufficient conditions for the admissibility under squared error loss of formal Bayes estimators. *Ann. Mathemat. Statist.*, 41, 1444–1447.

Zucchini, W. (1999). Frequentist model choice. Technical report, Cagliari, Sardinia. Summer school on Model Choice.

Index des noms

Abraham, C., 92, 152, 153
Abramovich, F., 52
Abramowitz, M., 161, 165, 314, 547
Adams, M., 398
Aitkin, M., 249, 250, 252, 288, 380
Akaike, H., 22, 176, 380
Alam, K., 24, 107, 108, 146, 165, 189, 307, 445, 492, 518, 543
Albert, J.H., 502
Anderson, T.W., 109, 203
Andrieu, C., 549
Angers, J.F., 154, 199, 498, 512
Arrow, K.S., 63, 91
Attali, J.G., 209

Balakrishnan, N., 563
Bar-Lev, S., 130
Baranchick, A.J., 107
Barbieri, M., 408
Barnard, G.A., 17
Barnett, G., 211, 214
Barron, A., 54, 406, 407
Bartlett, M.S., 149, 173
Basford, K., 364
Basu, D., 40, 179
Baum, L.E., 358
Bauwens, L., 33, 44, 154, 206, 210, 228, 230, 232, 234, 235, 549
Bayarri, M.J., 70
Bayes, T., 10–14, 50
Bechofer, R.E., 230
Bensmail, H., 364
Beran, R., 493
Bergé, P., 2
Berger, J.O., 8, 17, 20, 22, 31, 37, 38, 51, 63, 71, 76, 79, 86, 87, 100, 102, 107–109, 117, 120, 136, 141, 143, 145–147, 151–155, 166, 177–179, 182, 186, 191, 201, 216, 230, 232, 246, 250–254, 257, 258, 265, 267–271, 273, 277, 280, 282, 283, 286, 290, 291, 297, 298, 301, 302, 313, 342, 370, 376, 378, 381, 398, 409, 426, 430, 433, 436, 445, 446, 448, 453, 456, 465, 472, 475, 479, 482, 483, 485, 490, 498, 502, 503, 506, 509, 511, 515, 517, 519, 525, 529, 531, 533–535, 539, 550, 557, 561, 563
Berger, R., 84, 220, 245, 271–273, 276, 290, 291, 295, 299
Bergman, N., 182
Berliner, L.M., 152, 153, 155, 525
Bernardo, J.M., 50, 65, 91, 116, 141, 143, 145–147, 170, 171, 227, 250, 401, 503, 557, 561
Berry, D.A., 549
Berry, S.M., 48
Bertrand, J., 50, 114
Besag, J., 338, 339, 384, 549
Best, N.G., 329, 360, 361, 381–384, 410, 411, 549
Bhattacharya, R.N., 149
Bickel, P.J., 80, 149, 173

Billingsley, P., 38, 353
Billio, M., 214, 364
Bilodeau, M., 108
Binder, D., 357
Birnbaum, A., 17, 20
Bjørnstad, J., 22
Blackwell, D., 76, 435, 484
Blattberg, R.C., 206, 529
Blyth, C.R., 63, 111, 279, 436
Bobée, B., 549
Bock, M.E., 80, 107–109, 186, 187, 223, 225, 446, 493, 518, 540
Böhning, D., 520
Bondar, J.V., 108, 429, 480–482
Boole, G., 114
Bose, S., 111
Boukai, B., 268, 302
Boulton, D.M., 90
Box, G.E.P., 22, 208, 211, 215, 346, 549
Brandwein, A.C., 107–109, 187
Brewster, J.F., 527
Brockwell, P.J., 208, 213, 215, 228, 229
Broniatowski, M., 365
Brown, L.D., 71, 73, 76, 79, 92, 93, 102, 106, 108, 109, 125, 127, 131, 136, 158, 171, 187, 191, 226, 275, 281, 297, 298, 301, 302, 423, 429–432, 435, 436, 440–447, 450, 455, 512, 515, 531, 558
Buehler, R.J., 300, 492

Cappé, O., 213, 319, 342, 418, 420
Carlin, B.P., 122, 150, 313, 381–384, 391–393, 410, 411, 519, 520, 524, 550, 557
Carlin, J.B., 388, 498, 503, 507, 557
Caron, N., 291, 299
Carota, C., 377
Carriquiry, A., 550
Carter, G., 525
Casella, G., 2, 8, 15, 16, 23, 34, 41, 45, 47, 53, 63, 76, 80, 84, 86, 95, 108, 109, 179, 191, 206, 220, 234, 245, 266, 271–277, 280, 282–284, 288, 290–293, 295, 296, 299, 300, 306, 313–316, 319, 322–366, 380, 388, 394, 426, 455, 457, 458, 460, 507, 513, 518, 523, 527, 529–531, 537, 542, 544
Castillo-Chavez, C., 354
Castledine, B., 195, 222
Castro, I., 406, 415
Celeux, G., 45, 344, 357, 364, 365, 382
Cellier, D., 108, 109, 450, 518
Chalmers, T.C., 28, 496
Chamberlain, G., 58
Chen, J., 282, 319
Chen, M.H., 385, 387, 411, 412
Chernoff, H., 230
Chib, S., 214, 371, 390–393
Chickering, D.M., 549
Chow, M.S., 146, 171, 445
Chrystal, G., 114, 221
Chuang, D., 190
Clarke, B., 147
Clayton, D.G., 549
Clevenson, M., 431, 454
Clifford, M.S., 352
Clyde, M., 398, 399, 413
Cohen, A., 282, 283
Congdon, P., 550
Conigliani, C., 406, 415
Cornfield, J., 203, 204
Cowell, R.G., 341, 375, 544, 545, 549
Cowles, M.K., 361
Cox, D.R., 6, 20, 49, 199
Crawford, S.L., 342
Cressie, N., 375, 384

Dacunha-Castelle, D., 381
Dalal, S.R., 137, 169, 172, 442, 496
Dale, A.I., 50
Damien, P., 339
Darroch, J., 196
Das Gupta, A., 108, 152, 431, 448
Datta, G.S., 149
Daurés, J.P., 92, 152, 153
Davis, P.A., 208, 213, 215, 228, 229
Dawid, A.P., 116, 151, 163–165, 341, 375, 382, 492, 544–546, 549
Deely, J.J., 229, 230, 498, 519, 525
DeGroot, M.H., 58, 60, 63, 70, 94, 116, 170, 248, 342
Delampady, M., 152, 246, 268, 270, 271, 286, 290, 488
Dellaportas, P., 546

Dempster, A.P., 23, 342, 365, 496
Denison, D.G.T., 550
DeRobertis, L., 153
Dette, H., 27, 42
Devroye, L., 2, 316, 520
Dey, D., 9, 149, 381, 386
Diaconis, P., 47, 48, 54, 124, 130, 131, 133, 163, 173, 176, 193, 314, 442, 496
DiCiccio, T.J., 149, 173, 174
Dickey, J.M., 129, 185, 308, 361, 504
Diebolt, J., 331, 352, 357, 365, 379, 460
Doucet, A., 182, 306, 319, 549
Dransfield, M., 314, 322
Draper, D., 498
Drèze, J.H., 154, 234
Dudewicz, E.J., 230
Duffy, D., 217
Dumouchel, W.M., 498
Dupuis, J.A., 115, 116, 156, 196, 400–402, 404, 421
Dykstra, R.L., 23, 39, 179
Dynkin, E.B., 158

Eaton, M.L., 108, 109, 203, 432, 433, 460, 461, 465, 472, 473, 475–479, 487, 489
Eberly, L., 280
Eco, U., 398
Efron, B., 15, 101, 112, 198, 528
Eichenauer, J., 81
Engle, R.F., 233
Enis, P., 130
Escobar, M.D., 9, 367, 371
Evans, M., 21

Fabius, J., 52
Fang, K.T., 109
Farrel, R., 277
Farrell, R.H., 86, 274–277, 292, 441, 446, 457
Feller, W., 129, 217, 431
Ferguson, T.S., 9, 52, 58, 76, 220, 435
Fernandez, C., 44
Feyerabend, P., 553
Field, A., 6
Fieller, E.C., 47, 295
Fienberg, S., 70, 116
Finetti, B., de, 51, 124, 171
Fishburn, P.C., 58, 62
Fisher, R.A., 9, 16, 17, 22, 36, 50, 51, 140, 266
Fishman, G.S., 316
Fitzgerald, W.J., 549
Forbes, F., 382
Forster, J.J., 546, 557
Foster, D.P., 226, 413, 414
Fouley, J.L., 498
Fourdrinier, D., 82, 86, 108, 109, 148, 450, 518, 546, 547
Fraisse, A.M., 108, 445, 449, 456, 457
Francq, C., 209
Fraser, D.A.S., 21, 177
Freedman, D.A., 54, 176
Freitas, N., de, 319

Gassiat, E., 381
Gatsonis, C., 80, 220, 550
Gauss, C.F., 14, 57, 85, 198
Geisser, S., 203, 204
Gelfand, A.E., 237, 332, 333, 336, 337, 339, 342, 372, 373, 381, 386, 389, 390, 524
Gelman, A., 325, 361, 388, 417, 418, 498, 503, 507, 550, 557
Geman, D., 329, 549
Geman, S., 329, 549
Genest, C., 91
Gentle, J.E., 2, 316
George, E.I., 107–109, 186, 195, 196, 206, 226, 282, 310, 330, 333, 338, 399, 413, 414, 498, 529–531, 540
Geweke, J., 354, 361, 549
Geyer, C.J., 179, 306, 324, 333
Ghosh, J.K., 149
Ghosh, M., 108, 111, 147–150, 173, 206, 224, 529
Gianola, D., 498
Gibbons, J.D., 230
Gilks, W.R., 323, 329, 347, 360, 549, 550
Gill, J., 550
Gill, R.D., 182
Gini, C., 51
Giri, N., 479
Girshick, M.A., 76, 435, 484
Giudici, P., 546

Givens, G.H., 28, 496
Gleick, J., 2
Gleser, L.J., 295, 409, 479
Godsill, S.J., 306
Godwill, S., 549
Goel, P.K., 229, 231
Goldstein, M., 513
Good, I.J., 117, 152, 155, 242, 245, 501, 538, 559
Gordon, N.J., 182, 319
Gouriéroux, C., 36, 173, 226, 233, 380, 388, 403
Goutis, C., 121, 154, 266, 282, 300, 399, 400, 402, 403
Green, P.J., 53, 339, 365, 371, 379, 393, 394, 546
Greenberg, E., 214
Grenander, U., 53, 395, 418
Gruet, M.A., 342, 394–396
Grünwald, P., 549
Guihenneuc-Jouyaux, C., 235, 496, 498
Guillin, A., 319
Gupta, S.S., 229, 230
Gutmann, S., 108
Györfi, L., 520

Hadjicostas, P., 48
Haff, L., 108
Hàjek, J., 6
Hald, A., 50
Haldane, J., 32, 177, 280
Hall, P., 15
Hall, W.J., 32, 137, 169, 172, 442, 496
Hamilton, J.D., 209
Hammersley, J.M., 352
Hansen, M., 151
Harris, J.E., 498
Harrison, J., 210, 213, 550
Hartigan, J.A., 31, 153, 265, 280, 294, 536–538
Has'minskii, R., 51, 54, 176, 555
Hastings, W.K., 325, 327
Healy, J.D., 479
Heath, D., 8
Heckerman, D., 549
Heidelberger, P., 361
Heitjan, D.F., 359
Helland, I.S., 492, 493
Hesterberg, T., 386
Hills, S., 314
Hinkley, D., 49, 302
Hjort, N.L., 52
Hoaglin, D., 416
Hobert, J.P., 34, 45, 53, 325, 363, 459, 460, 498
Hodges, J.S., 550
Hoerl, A., 513
Holmes, C.C., 550
Hora, R.B., 492
Huber, P.J., 88, 153
Hubert, P., 549
Huerta, G., 211
Hui, S., 525
Hurn, M.A., 45, 344, 365
Hutchinson, D., 279
Hwang, J.T.G., 23, 85, 86, 92, 93, 103, 104, 108, 111, 112, 146, 179, 191, 274–277, 282–284, 292, 295, 296, 409, 423, 429, 430, 436, 447, 450, 455, 457, 458, 529, 542

Ibragimov, I., 51, 54, 176, 555
Ibrahim, J., 319, 385, 387, 412
Im, S., 498

Jacquier, E., 233, 549
James, W., 107, 110, 129, 447
Jaynes, E.T., 118, 163, 164, 560
Jefferys, W., 398
Jeffreys, H., 50, 51, 67, 114, 116, 117, 125, 139, 141, 193, 221, 224, 241, 242, 250, 264, 265, 376, 398, 557
Jenkins, G.M., 208, 211, 215
Johnson, B.M., 451, 457
Johnson, N.L., 563
Johnstone, D.J., 289
Johnstone, I.M., 80, 102, 108, 191, 430, 432
Johnstone, R.W., 108
Jones, M.C., 214
Jordan, R., 1, 55, 113, 175, 237, 305, 369, 423, 463, 495, 549
Joshi, V.M., 177, 281, 282, 542
Judge, G., 107, 446

Kadane, J.B., 190, 320–322, 342, 349

Kariya, T., 479
Karlin, S., 260, 426, 455
Kass, R.E., 119, 137–139, 143, 147, 151, 156, 166, 173, 242, 243, 320–322, 349, 378, 381, 529, 550
Keating, J.P., 111
Keeney, R.L., 64
Kelker, D., 35, 109
Kemp, A.W., 563
Kemperman, J., 47, 48
Kempthorne, P.J., 80, 81, 155, 442, 561
Kendall, M., 523
Kennard, R., 513
Keynes, J.M., 50
Khattree, R., 111
Kiefer, J., 84, 267, 285, 300, 301, 479
Kiiveri, H., 544
King, A., 333
Kirby, A.J., 497, 549
Kleibergen, F., 233
Kohn, R., 211, 214
Kokaram, A.C., 549
Kolmogorov, A., 50
Kong, A., 337
Kontkanen, P., 549
Koo, J.O., 230
Koopman, B., 125
Kotz, S., 563
Krishnan, T., 23, 365
Kubokawa, T., 47, 108, 111, 112, 227, 517, 527, 539, 541

Lad, F., 50
Laird, N.M., 23, 342, 365, 524
Laplace, P.S., 10–14, 22, 32, 50, 57, 64, 87, 137, 177, 192, 193, 198, 319
Lasserre, V., 496, 498
Lauritzen, S.L., 158, 341, 375, 497, 533, 544–546, 549
Lavielle, M., 365
Lavine, M., 52
Lawley, D.N., 173
Le Cam, L., 76, 136, 177, 482
Lee, T.M., 237
Legendre, A., 14, 85
Lehmann, E.L., 6, 8, 15, 16, 23, 34, 47, 76, 107, 109, 258, 260–263, 281, 298, 300, 301, 380, 425, 426, 450, 473, 474, 479, 482, 484, 486, 487, 555
Lehn, J., 81
Lenk, P., 4, 5, 371, 406
Leonard, T., 322
Letac, G., 127, 130
Levit, B.Y., 182
Lewis, S., 325, 361
Lindley, D.V., 11, 15, 31, 32, 51, 57, 85, 114, 116, 148, 155, 203, 204, 224, 265, 274, 281, 289, 297, 298, 320, 498, 501, 504, 509, 512, 513, 519, 525, 534, 557
Liseo, B., 149, 408
Liu, J.S., 333, 337, 363
Louis, T.A., 122, 302, 313, 519, 520, 524, 550, 557
Lu, K., 86, 108, 109, 191, 282, 531
Lubrano, M., 206, 210, 228, 230, 232, 234, 235, 549
Lwin, T., 519, 521, 524, 525

Maatta, J., 300, 527
MacGibbon, B.K., 80, 154, 220, 498
Machina, G., 58
MacLachlan, G., 23, 364, 365
Madigan, D., 396–398, 413, 544, 546
Makita, S., 541
Makov, U.E., 345, 364
Mallick, B.K., 550
Marin, J.M., 53, 319, 365
Maritz, J.S., 519, 521, 524, 525
Marsaglia, G., 360
Mason, R., 111
McCullagh, P., 382, 388, 408, 420, 502
McCulloch, R.E., 399, 401, 550
McNeil, A.J., 549
Meng, X.L., 363, 365, 387, 388, 417, 418
Mengersen, K.L., 28, 45, 53, 235, 329, 356, 365, 366, 371, 379, 401, 496
Metropolis, N., 315, 323, 325
Meyn, S.P., 207, 208, 323, 324, 431
Miller, M.I., 53, 395, 418
Milnes, P., 480–482
Miquel, J., 549
Mira, A., 340
Møller, J., 340

Monahan, J.F., 211
Monette, G., 21, 177
Monfort, A., 36, 173, 214, 226, 364, 380, 388, 403
Moors, J.J.A., 457
Mora, M., 127
Morales, J.A., 154
Morgenstern, O., 58
Morisson, D., 523
Morita, S., 541
Morris, C., 127, 159, 314, 345, 519, 521, 527, 528, 538
Mortera, J., 268
Mosteller, F., 28, 193, 198, 416, 496
Moulines, E., 213, 319, 365
Mukerjee, R., 147–149
Müller, P., 9, 52, 314
Muller, M., 346
Murphy, A.H., 70
Musio, M., 44
Myllymäki, T., 549

Nachbin, L., 475, 479
Nagakura, K., 541
Naylor, J.C., 314, 322
Neal, R.M., 549
Nelder, J., 382, 388, 408, 420, 502
Nelson, D.B., 233
Neumann, J., von, 58, 315
Newton, M.A., 386
Neyman, J., 18, 51, 71, 89, 192, 405, 415
Ng, K.W., 177
Novick, M.R., 32

O'Hagan, A., 152, 153, 255, 256, 406, 415, 557
Occam, W., d', 398
Olkin, I., 40, 230
Olver, F.W.J., 320
Osborne, C., 47, 227
Owen, A., 386

Panchapakesan, S., 230
Parent, E., 549
Parmigiani, G., 377
Pathak, P.K., 111
Pearl, J., 286
Pearson, E.S., 18, 36, 71, 89
Pearson, K., 50, 365
Peddada, S., 111
Peers, H.W., 148, 149, 225
Pemantle, R., 85, 274, 277
Pericchi, L.R., 250–254, 257, 258, 370, 376, 378, 381, 409
Perk, W., 157
Perron, F., 479
Petkau, A.J., 40
Petrella, L., 408
Petrie, T., 358
Petrone, S., 406
Pettit, L.I., 250
Pfangzagl, J., 260
Philippe, A., 151, 297, 342, 394–396
Phillips, D.B., 52, 53, 232, 371, 395, 418
Pierce, D., 79, 300
Pitman, E.J.G., 111, 125, 158, 467
Pitt, M.K., 549
Plessis, B., 5, 6, 311
Poincaré, H., 553, 559
Poirier, D.J., 549
Pollock, K., 194
Polson, N.G., 233, 377, 549
Pommeau, Y., 2
Popper, K., 193, 551, 553
Press, J.S., 204
Price, R., 50

Qian, W., 365

Racugno, W., 44, 370
Raftery, A.E., 195, 242, 243, 325, 361, 364, 378, 381, 382, 385, 386, 396–398, 400, 413
Raiffa, H., 20, 38, 44, 63, 64, 95, 97, 124, 195
Ralescu, S., 107
Rao, C.R., 111
Rao, R., 149
Raoult, J.P., 445, 456, 457
Reber, A., 500, 502, 503, 508, 509
Redner, R., 365
Reid, N., 199
Richard, J.F., 33, 154, 206, 210, 228, 230, 232, 234, 235, 549
Richardson, S., 53, 323, 365, 371, 379, 394, 400, 496, 498, 550
Ripley, B.D., 360, 395, 418, 549
Rissanen, J., 22, 151

Ritov, Y., 49, 54
Robbins, H., 155, 519, 521
Robert, C.P., 2, 23, 41, 45, 47, 53, 71, 80, 86, 91, 101, 103, 104, 108, 109, 111, 112, 151, 161, 162, 179, 182, 186, 191, 195, 196, 214, 222, 223, 227, 234, 235, 249, 250, 274–277, 282–284, 288, 292, 293, 296, 297, 299, 306, 310, 313–316, 319, 320, 322–366, 371, 379, 382, 388, 394–396, 399–404, 418, 420, 421, 433, 442, 445, 449, 450, 455–460, 478, 493, 500, 502, 503, 506–509, 513, 515, 518, 525, 527, 533, 534, 540, 544, 546, 547, 561
Roberts, G.O., 324, 337, 340
Robertson, T., 23, 39, 179
Robins, J., 8, 49, 54, 375
Robinson, G.K., 267, 285, 295, 300, 301
Roeder, K., 365, 366, 371, 372
Rolph, J., 525
Romano, J.P., 36, 40, 41
Ronchetti, E., 6
Rosenbluth, A.W., 323, 325
Rosenbluth, M.N., 323, 325
Rosenthal, J.S., 340
Rossi, P.E., 233, 401, 549
Rousseau, J., 149, 151, 322
Roy, M., 108, 445, 449, 456, 457
Rubin, D.B., 23, 62, 325, 342, 359, 361, 365, 388, 498, 503, 507, 557
Rubin, G., 354
Rubin, H., 63, 91, 147, 229, 231, 260, 455
Rubinstein, R.Y., 348
Rudin, W., 474
Rue, H., 105, 106
Rukhin, A.L., 86, 93, 101, 108, 191, 285, 424, 443, 515
Rydén, T., 213, 319, 344, 358, 394, 418, 420

Sackrowitz, H., 283
Sahu, S.K., 337
Saleh, A.K.Md.E., 108, 111, 112, 206, 527, 529
San Cristobal, M., 498
Santner, T.J., 217
Savage, L.J., 8, 51, 177
Saxena, K., 24, 146, 165, 189, 307, 445, 492
Schaafsma, W., 274
Schervish, M.J., 54, 70, 85, 96, 274, 320
Schlaifer, R., 20, 38, 44, 63, 97, 124, 195
Schwarz, G., 381
Seber, G.A.F., 194
Seidenfeld, T., 50, 119, 156
Sellke, T., 152, 254, 268–270, 273, 277, 291
Sen, P.K., 111, 112, 206
Seneta, E., 221
Severini, T.A., 148
Shafer, G., 375
Shannon, C., 118, 151
Shao, J., 107, 319, 385, 387, 411, 412
Sharples, L.D., 549
Shaw, J., 314, 322
Sheather, S., 211, 214
Shephard, N., 549
Shinozaki, N., 86, 100, 108, 282, 431
Shorrock, G., 282
Shyu, S.F., 354
Sidàk, Z., 6
Siegel, A.F., 36, 40, 41
Silander, T., 549
Silverman, B.W., 52
Singpurwalla, N., 550
Sinha, B.K., 431
Sinha, D., 9, 108
Sivaganesan, S., 152, 153
Sken, A., 314, 322
Small, M.J., 342
Smith, A.F.M., 52, 53, 63, 65, 91, 116, 170, 171, 203, 204, 227, 237, 250, 299, 314, 322, 332, 333, 336, 337, 339, 345, 364, 371, 379, 395, 401, 418, 501, 503, 504, 509, 512, 513, 534, 550, 557
Smith, D.D., 28, 496
Smith, J.Q., 64, 65, 70, 95, 116, 190, 215, 218, 558
Sobel, M., 230
Soubiran, C., 357
Spatinas, T., 52
Speed, T.P., 544

Spiegelhalter, D.J., 250, 299, 323, 341, 360, 375, 379, 381–384, 410, 411, 497, 533, 544, 545, 549, 550
Srinivasan, C., 108, 120, 431, 445, 446, 456
Srivastava, M.S., 108, 150
Stangl, D.K., 549
Stark, J.A., 549
Steel, M., 44
Steffey, D., 173, 322, 529
Stegun, I., 161, 165, 314, 547
Stein, C., 51, 102, 106–108, 110, 129, 143, 177, 179, 186, 216, 281, 430, 436, 438, 441, 447, 477, 480, 515, 535
Stephens, D., 418, 419
Stephens, M., 53, 365, 395
Stern, H.S., 388, 498, 503, 507, 557
Stern, S.E., 149, 173, 174
Steward, G., 206
Stigler, S., 9, 12–14, 37, 50, 198, 314
Stone, M., 138, 163, 177, 179, 492
Stone, N., 163–165
Strasser, H., 76, 473, 482
Strawderman, W.E., 71, 73, 79, 80, 82, 103, 104, 107–109, 111, 112, 148, 187, 220, 282, 430, 517, 537, 539
Stuart, A., 523
Studden, W.J., 27, 42, 152
Sudderth, W.J., 8
Sweeting, T.J., 148

Tan, K.K.C., 329, 360
Tanner, M., 331, 333, 334
Teicher, H., 171
Teller, A.H., 323, 325
Teller, E., 323, 325
Thatcher, A.R., 294
Thisted, R.A., 198
Thomas, A., 360
Thompson, E.A., 179, 306
Thompson, P.M., 266
Tiao, G.C., 22, 549
Tibshirani, R., 143, 151
Tierney, L., 320–322, 340, 349
Tirri, H., 549
Titterington, D.M., 344, 358, 364–366, 382, 394, 561
Tolboom, J., 274
Tompa, H., 154, 234
Tong, H., 208
Torrie, G.M., 412
Tsui, K., 108
Tukey, J.W., 416
Tweedie, R.L., 28, 207, 208, 323, 324, 329, 431, 496

Ulam, S., 315
Ullah, A., 108, 282
Umbach, D., 354

Valleau, J.P., 412
Van der Linde, A., 382–384
Van der Meulen, B., 85, 274
Van Dijk, H.K., 233
Van Dyk, D.A., 363, 365
Van Eeden, C., 91
Venn, J., 50, 114
Verdinelli, I., 314, 364, 406, 407, 417, 550
Vidakovic, B., 9, 52, 314
Vidal, C., 2
Villegas, C., 32, 491
Vines, K., 361
Volinsky, C., 396

Wakefield, J.C., 339
Wald, A., 18, 51, 58, 71, 301, 444
Walker, H., 365
Walker, S., 339
Wallace, C.S., 90
Wallace, D.L., 198
Walley, P., 152, 167, 296, 297, 496
Wang, Y., 268, 302
Wasserman, L., 8, 49, 54, 119, 137–139, 143, 147, 151, 152, 154, 156, 166, 314, 364–366, 371, 375, 379, 406, 407, 417
Welch, B.L., 148, 149, 225
Welch, P.D., 361
Wells, M.T., 63, 82, 86, 101, 108, 148, 266, 274–277, 292, 342, 356, 457
West, M., 9, 210, 211, 213, 367, 371, 550
Whittaker, J., 544

Wijsman, R.A., 465, 473
Wild, P., 347
Wilkinson, G., 51
Winkler, R.L., 70
Wolpert, R., 17, 20, 22, 37, 38, 87, 177–179, 216, 267, 298, 301, 302, 557
Wolter, W., 196
Wong, W.H., 331, 333, 334, 337, 387
Wright, F.T., 23, 39, 179
Wu, Y.N., 363

Yahav, J.A., 230
Yang, M.C., 224
Yang, R., 232
Yao, Y.C., 209
Ylvisaker, D., 124, 130, 131, 133, 163, 173, 442, 496
York, J., 544, 546
Yu, B., 151

Zabell, S.L., 50, 51, 221, 314
Zakoïan, J.M., 209
Zaman, A., 360
Zellner, A., 100, 154, 204–206, 409, 549
Zhou, Y., 386
Zidek, J.V., 40, 91, 163–165, 430, 431, 451, 454, 477, 492, 527
Zucchini, W., 153

Index des matières

a posteriori, 10, 18, 25
  construction, 175
  coût
    approché, 319
    moyen, 68
  erreur quadratique, 181
  impropre, 362, 363
  information, 146
  marginal, 234
  médiane, 14, 88, 190
  moyenne, 85
  probabilité, 148, 241
    p-value comme, 301
    variation de la, 268
  propre, 31, 342
  pseudo-, 256, 521
  région crédible, 148
a priori, 10
  arbitraire, 114, 122
  à symétrie sphérique, 442
  biais, 80
  choix d'un, 15, 113, 155, 361, 397, 558
    automatique, 114
    exact, 152
    paramétré, 120
    subjectif, 114, 116, 117
  classe
    à moments déterminés, 152
    de voisinages, 153
    ϵ-contaminée, 153
    rapport de densités, 153
    sous-spécifiée, 153
cognitif, 553
comme outil, 552, 555
conjuguée, voir conjuguée
construction de l', 116
d'entropie maximale, 154, 156
de coïncidence, 148, 149, 225
de Dirichlet, 366, 406, 546
de Haldane, 491
de Jeffreys, voir non informatif, a priori de Jeffreys
  controversé, 210
de référence, 137, 143, 144, 146, 147, 176, 227, 230
  construction, 145
du maximum d'entropie, 121
dualité entre coût et, 147, 242
existence d'un, 169, 170
fermé par échantillonnage, 123
fondements axiomatiques d'un, 170
G, 204–206, 409
hiérarchique, voir Bayes hiérarchique
impropre, voir impropre a priori
  et hypothèse ponctuelle, 248
  et test, 403
incertitude sur l', 152
inconnu, 519
influence de l', 114
information, 113, 137, 152, 155, 206, 311
  diffuse, 247
intrinsèque, 254, 256, 289
intuition, 553

invariante par changement d'échelle, 139
Jeffreys
alternative à l', 483
mélange de, 133
modélisation, 134
effet de la, 342
modification de l', 244, 273
non informatif, voir non informatif
objectif, 123
par arbres de Pólya, 52
par défaut, 137
paramétré, 114, 115, 119
poly-t, 206, 234, 235
pour les tests, 151
probabilité d'un modèle, 397
pseudo, 391
relativement invariant, 487
robuste, 154
sélection d'un, 71
uniforme, 137, 138, 146
vague et propre, 31
absence de décision
alternative par, 301
réponse, 302
acceptation
niveau d', 240, 241, 259
conventionnel, 243, 249, 266
rapport d', 326, 327
acceptation-rejet, 346
accidents dans le Michigan, 4, 371
adéquation, 374, 405
problème d', 406
admissibilité, 81, 82, 86, 91, 148
condition nécessaire d', 275
d'un unique estimateur, 81
de l'estimateur de Bayes, 82
de l'estimateur des moindres carrés, 106
et récurrence, 432
AIC, voir critère d'information d'Akaike
ajustement
contre erreurs d'estimation, 373
meilleur, 374
aléatoire
générateur, voir générateur pseudo-aléatoire
algorithme, 146
ARMS, 329, 360
d'acceptation-rejet, 346
avec enveloppe, 347
de Box-Muller, 346
de Durbin-Levinson, voir récurrence de Durbin-Levinson
de Gibbs, voir échantillonnage, de Gibbs
de Metropolis-Hastings, 325, 327, 350
à marche aléatoire, 328, 329
EM, voir EM
MCMC, 323
α
niveau, 266
amiante, 239
analyse
conditionnelle, 300
de la variance, 147, 230, 286
de robustesse, 152, 155
de sensibilité, 152, 558
animal
biologie, 192, 194
élevage, 498
approche
bayésienne empirique, 27
conditionnelle, 285
fiduciaire, 22
fréquentiste, 18, 66
non paramétrique, 6, 364, 370
paramétrique, 6, 8
approximation, 136
de Bartlett, 149
de la densité marginale, 385, 390
de Laplace, 54, 319–322, 342, 349, 380
du second ordre par χ^2_k, 173
numérique, 313
par point-selle, 329
AR(1), voir modèle, AR(1)
arbre
classement par, 40
de décision, 65
de modèles possibles, 374
élagage, 402
orangers, 372
pins, 392
ARCH, 233
argument limite, 33

aucun modèle n'est vrai, 238
augmentation de données, 331
auteur
identification d', 198
autocorrélation
partielle, voir partiel
autocovariance, 212
autorégressive, voir modèle, AR
axiomes, 116
bayésiens, 11, 30, 35, 169, 550–561
choix des, 472
de cohérence, 168
de pari, 167
des probabilités, 50
statistiques, 15, 35

base fonctionnelle, 124
bayésien
calcul, 313, 341
contre fréquentiste, 238
critère d'information (BIC), 381, 382, 414
logiciel, 561
meilleur centre, 293
modèle statistique, 10
non paramétrique, 9, 406
paradigme, 9–15, 22, 25, 50, 114, 251, 371, 405, 495, 500, 531
principe d'actualisation, 156
software, 313
test, 237, 241
UPP, 263
bayésienne
approche
cohérence de l', 550, 554
critique de l', 557
non informative, 560
approximation, 345
imagerie, 70
inférence, 136
minimaxité, 561
réponse la moins favorable, 268
robustesse, 155
Bayes
empirique, 122, 518
de Bayes, 546
et effet Stein, 525–531
inconvénient du, 529, 531
modélisation, 311
non paramétrique, 519
paramétrique, 521
test, 524
estimateur de, 69
admissible, 424
analytique, 186
calcul d'un, 305, 556, 557
du coût, 190
et admissibilité, 82
généralisé, 69, 79, 141, 275, 428, 429, 443, 446
hiérarchique, 308
inadmissible, 424
inconsistant, 54
limite d', 442
linéaire, 124
meilleur équivariant, 476
minimax, 77, 80, 537
pour une mesure invariante, 473
propre, 438
pseudo-, 109
randomisé, 78
représentation différentielle, 429
universel, 93
hiérarchique, 122, 154
décomposition, 504–506
estimateur, 514
et robustesse, 496
inconvénient du modèle, 507
modèle, 499
motivations, 501–504
problèmes numériques, 507–509
règle de, 184
risque de, 69
infini, 73
intégré, 68
théorème de, 9–10, 375, 551
Behrens-Fisher, problème de, 294
Berger
paradoxe de, 296
phénomène de, 108
réconciliation au sens de, 301–303
biais, 373
BIC, voir bayésien, critère d'information
Biostatistiques, 549
bootstrap, 15
borné
coût, 62, 190, 277, 428, 434, 441, 452

ensemble de risque, 443, 456
espace des paramètres, 179
paramètre, 86
bornes inférieures de demi-queue (STUB), 447
bruit blanc, 211
BUGS, 360, 378
bus, 3

C, 360
calcul
de l'incertain, 3
difficultés de, 306
calibration, 226, 295, 328
d'expert, 70
linéaire, 47, 143, 150, 408, 409
intervalle de confiance, 409
cancer de la lèvre, 384
capture-recapture, 3, 115, 309
calcul de l'a posteriori en, 310
modèle, 194–198
de Darroch, 196
temporel, 331
carte d'allocation, 234, 395
censure, 38
cercle unité, 208, 211
cerf, 194
chaos, 2
choix de modèle, 342, 369
espace des paramètres, 375
et estimation, 370
et test, 369
termes de pénalisation, 377
classe complète, 260, 274, 276, 435, 443–446
essentiellement, 276
minimale, 455
classement, 230
classification, 364
taux d'erreur, 105
clique, 506, 544
ordre parfait pour une, 544
coïncidences, 193
CODA, 324, 325, 360, 361
coefficient
de normalisation, 28, 140, 223
de réflexion, 211
multinomial, 195
cohérence, 8, 26, 34, 97, 98, 138, 157, 168, 551
des facteurs de Bayes, 256
coin, 133
colinéarité, 513
comité, 91
complet
chaîne, 358
classe, 34
distribution, 111
espace, 72
forme exponentielle, 320
ignorance, 32
statistique, 37
variable aléatoire, 359
condition
d'équilibre ponctuel, 326, 350, 394, 418
de positivité, 331
nécessaire et suffisante
de Stein, 426, 441
STUB, 543
suffisante d'admissibilité, 426
d'Eaton, 460–461
de Blyth, 436, 441
conditionnement, 504
confiance
index de, 285
intervalle de, 216
niveau de, 29, 238
rapporté, 530
région, 19, 24, 28, 108, 191, 277
comme région crédible, 279
de niveau α, 281
efficace, 282
non connexe, 280
recentrée, 282, 296, 529, 530
conjugué, 341
a priori, 123–125, 127, 128, 131, 133, 154, 187
classe de, 152
et a priori d'entropie maximale, 160
et loi de Jeffreys, 142
et BUGS, 360
hyperparamètre, 137
mélange de, 135, 157, 172, 442, 524
naturel, 171

normal, 199
famille, 123
minimale, 123, 157
contrôle, 325
contrainte
de positivité, 337
convergence, 23, 144, 176, 407, 463
diagnostic de, voir diagnostic
géométrique, 329, 331, 344, 353, 358
indicateur de, 508
quadratique, 440
vitesse de, 324
convexité, 152
coordonnées polaires, 145
Cornell University, 354
correction de Bartlett, 149, 173
pour un a posteriori, 173
coût, 56, 66, 554
absolu, 274
aléatoire, 91
asymétrique en erreur quadratique, 101
bidimensionnel, 259
borné, 62
classique, 85, 184
construction d'un a priori à partir du, 147
contrainte de, 555
convexe, 70, 76, 81, 85, 88, 99
strictement, 274
d'information, 264
d'opportunité, 97
de classification, 105
de Hellinger, 90, 101, 166
de prévision, 183
d'une p-value, 268
en erreur absolue, 54, 87, 100
entropique, 90, 91, 101, 469, 526, 527, 543, 547
estimation de, 86, 190, 191
global, 285
intrinsèque, 23, 89, 101
invariant, 224, 469
par changement d'échelle, 189
par reparamétrisation, 90
linéaire, 283, 284
LINEX, 100
maximal en potentiel explicatif, 401
multiple, 93
pour images, 105–106
pour l'estimation
de mélanges, 365
ensembliste, 283, 284
propre, 274, 277
quadratique, 85, 86, 106, 109, 190
rationnel, 284, 296
robustesse sous, 108
$0-1$, 89, 105, 239, 258, 261, 274
covariables, 372
covariance
matrice de, 206
critère
d'information
bayésien, voir bayésien, critère d'information
d'Akaike (AIC), 380, 382, 414
de déviance (DIC), 382
de proximité de Pitman, 103
de Schwarz, 380, 381
inefficace, 381
minimax, 73
croyance, 552

DAG, voir graphe acyclique orienté
décision
arbre de, 65
dans l'incertain, 63
erreur, 57
espace de, 66, 267
compact, 76
optimale, 97
unique, 259, 554
pour le choix de modèle, 377
décomposition de Wold, 211, 228
densité prédictive, 367, 376, 377, 389, 396
déplacement
fusion, 395
naissance et mort, 395, 417
séparation, 395
développement d'Edgeworth, 149
déviance, 382, 408
bayésienne, 383
associée, 410
pénalisée, 382

diagnostic
d'autocorrélation, 361
de convergence, 324
software, 361
DIC, voir critère d'information de déviance
dichotomie unilatérale/bilatérale, 273
distance
de Hellinger, 54, 90
de Kullback-Leibler, voir divergence, Kullback-Leibler
de Prohorov, 135
en variation totale, 172
entre distributions, 370
entropique, 90
distribution de probabilité, voir loi
divergence, 400
de Kullback-Leibler, 90, 118, 146, 153, 377, 401, 402, 414
domination
stochastique, 92, 111
universelle, 92
donnée
collecte de, 1
grossière, 359
manquante, 160, 313, 335
représentation par, 342
dualité, 86
principe de, 331

échangeabilité, 170, 528
échantillon d'apprentissage, 251
minimal, 251
pour un mélange, 379
échantillonnage
d'importance, 317, 329, 342, 385, 388, 411
choix de la fonction, 318
défensif, 386
et variance infinie, 319
pour le choix de modèle, 385
de Gibbs, 329–341
bivarié, 331, 333, 337, 338, 351–353
pour les mélanges, 344
processus de Dirichlet pour, 366
de substitution, 337
exact, 324
hybride, 339
par chemin, 388, 417, 418
par paquets, 325
par parapluie, 412
par passerelle, 387–388
par tranche, 339–341, 344, 354, 389
séquentiel, 253
échelle, 153
de cohérence, 116
Économétrie, 226, 376, 549
Écosse, 384
Edgeworth, voir développement
effet Stein, voir Stein, 479, 483, 484
effets aléatoires, 499, 503, 550
a priori de Jeffreys pour, 363
modèle à, 204
propriété a posteriori pour, 45, 363
efficacité, 23
dans l'inférence, 553
élevage, 499
EM, 365, 382
étapes, 365
encompassing, voir imbrication
ensemble
de troncature, 275
entropie
coût, voir coût
définition, 118
maximale, 117, 118
Environnemétrie, 549
équation
d'état, 213
d'observation, 213
différentielle, 150
équations simultanées, 154
équiprobabilité des événements élémentaires, 13, 138
ergodicité, 323, 324
uniforme, 324
erreur
de mesure, 497
de type I, 89, 258
de type II, 89, 258, 267
espérance morale, 64
espace
d'état, 213
représentation, voir représentation, espace d'état
dénombrable, 117

de dimension
infinie, 52
variable, 370
des paramètres, 7, 30, 125
compact, 138
naturel, 127
restreint, 179
fonctionnel, 9
estimateur, 66
admissible
avec risque de Bayes infini, 425
comme limite d'estimateurs de Bayes, 275
et minimax, 148
à rétrécisseur, 92, 108, 493, 511
matriciel, 446
multiple, 108, 541
à risque continu, 434
construction optimale d', 71
de Bayes, voir Bayes, estimateur de
de James–Stein, 129
de James-Stein, 104, 107, 282, 294, 525, 530
à partie positive, 74, 107
tronqué, 186, 225, 445, 526
de la régression inverse, 227
de Pitman, 468
de Rao-Blackwell, 508
des moindres carrés, 36, 106, 203
du maximum a posteriori (MAP), 102, 105, 174, 176, 376
du maximum de vraisemblance, 22, 131, 521
pénalisé, 176
équivariant, 469
inadmissible, 81
inconsistant, 54
meilleur équivariant, 141, 465
minimax, 73, 76
équivariant, 482
monotone, 435
performances d'un, 284
pseudo-Bayes, 225
randomisé, 71, 72, 81, 85, 88
ridge, 513
sans biais, 36
unique, minimax, 83, 292
estimation, 67
ensembliste, 283, 284
comme forme d'estimation, 284
comme forme de test, 284
comme indicateur de performance, 284
et test, 7
et évaluation, 8
mélange, 153, 245
non paramétrique, 520
sans biais, 191
évaluation
ex post, 238
évidence, 20
expérience mixte, 20
expert
calibration d', 70
ordre d', 95
système, 375, 544, 549
explicative
variable, 206

facteur de Bayes, 241, 243, 244, 246, 250, 378
approximatif, 243
arithmétique intrinsèque, 253, 254, 256, 289, 409
fractionnaire, 255, 256, 289, 409
géométrique intrinsèque, 254, 289
médian intrinsèque, 254
pour un a priori impropre, 378
pseudo-, 252, 256, 257, 370, 379, 392
calcul de, 258
choix de, 258
cohérence de, 378
facteur de pénalisation, 382
famille
conjuguée, voir conjuguée, famille
de position, 148
exponentielle, voir famille exponentielle
famille exponentielle, 16, 91, 125, 130, 131, 157, 160, 185, 521
courbe, 129
de variance quadratique, 160, 314, 538
naturelle, 345
estimateur admissible pour, 445

et rapport de vraisemblance monotone, 260
extension, 405
minimale, 127
naturelle, 158, 274
forme de, 125
paramètre naturel d'une, 522
restreinte, 159
position et, 158
pseudo-, 157
quasi, 125
régulière, 127, 158
variance d'une, 159
Federalist Papers, 198
fermeture, 30
file d'attente, 74
filtrage, 213
filtre de Kalman, 213
Finance, 549
Fisher
loi de, 36
fléau de la dimension, 314
fonction
analytique, 80, 322
coût, voir coût
confluente hypergéométrique, 161, 165, 199, 318, 518, 547
cumulante des moments, 127
d'importance, 348, 385, 386
choix de la, 316, 318, 385
optimale, 411
d'utilité, 61, 64
convexe, 64
de Bessel modifiée, 165, 222, 223, 308, 410, 442, 493
de lien, 387, 502
de vraisemblance, voir vraisemblance
estimable invariablement, 492
gamma, 132
génératrice des moments, 321, 349
régulière, 322
surharmonique, 516
variance quadratique, 159
forage pétrolier, 74
forme
complètement exponentielle, 320
standard, 320
formule de réflexion, 160
fractile, 153
fréquentiste
approche, 66
cadre décisionnel, 67, 71
conditionnel, 267
couverture, 148
méthode, 555
notion d'optimalité, 423
paradigme, 67
propriété, 147
de long terme, 151
risque, 66
test, 72, 259
validité, 191, 271, 455

Γ-minimax
regret, 155
risque, 155
générateur
infinitésimal, 496
pseudo-aléatoire, 315, 316, 360
Gibbs
champ de, 329
échantillonnage, voir échantillonnage, de Gibbs
Glasgow, 194
graphe, 544
non orienté, 544
orienté, 544
acyclique, 497, 532, 544, 545
Green
complétion de, 393
groupe
action d'un, 139
moyennable, 480, 482
structure de, 118
transitif, 470

Haar
mesure de, voir mesure, de Haar
ondelette de, 52
Harris
récurrence au sens de, 323
hétérocédasticité, 233
hiérarchique
a priori, voir Bayes hiérarchique
modélisation, 154
histogramme, 117, 120, 153
hot hand, 243, 244

Hubble
constante de, 245
hyper a priori, 122, 154, 503
hyperbolic secant, 160
hyperparamètre, 122, 128, 155, 202, 206, 499
conjugué, 219
estimé, 521
poly-t, 235
hypothèse
alternative, 239
contiguë, 259
nulle, 239, 240
ponctuelle, 237, 238, 245, 289
unilatérale, 271

identifiabilité, 26, 212, 215, 561
image, 70, 105
noir et blanc, 105
radiologique, 4
traitement, 549
imbrication, 376, 403
impropre
a priori, 30–34, 138, 140, 365
comme a priori usuel, 32
et échantillon d'apprentissage, 252
et hypothèse nulle ponctuelle, 238
et régions crédibles, 279
et test, 248, 249, 254
inégalité
de Jensen, 76
de Van Trees, 182
inadmissibilité, 107
de la p-value, 277
du maximum de vraisemblance, 106
incohérence, 116, 141
inconsistance
d'estimateurs de Bayes, 54, 144
de la loi de Jeffreys, 143
inégalité
de Cauchy-Schwarz, 452
de Cramér-Rao, 426
inférence, 1, 3, 7
causale, 375
quantitative, 552
infinie divisibilité, 129
information, 17, 27
a priori, 30, 35, 113
décomposition de l', 498
et loi a priori, 557
insuffisante, 496
justifications subjectives, 142
résumé de, 175
de Fisher, 139, 141, 144, 149, 150, 165
équivalent échantillon, 116
limitée, 114, 120, 124
manquante, 146
pour un problème de test, 267
structurelle, 502
supplémentaire, 245
vague, 121, 181
intégrabilité, 34
intégrale
approximation, 314
développement de Laplace d'une, voir approximation, de Laplace
invariante à gauche, 475
invariante à droite, 475
rapport d', 319
intégration
analytique, 305
numérique, 307
interface
bayésien-fréquentiste, 109, 268, 301
interprétation contre explication, 2, 550
invariance, 24, 57, 90, 124, 463
et mesure de Haar, 142
groupe, 468
par reparamétrisation, 118, 138, 140, 147
par translation, 139
structure d', 139
translation, 465
inverse généralisée, 417
inversibilité, 212, 229
inversion, 551
causale, 50
de la Statistique, 22
des probabilités, 9, 10, 25
entre causes et effets, 9
irréductibilité, 323, 326

Jeffreys

a priori, voir non informatif, a priori de Jeffreys
échelle de, 242
Jensen
inégalité de, 76, 359

Kalman
filtre de, 213
Kepler
problème de, 265
χ^2, voir loi, du khi deux

Lagrange, voir multiplicateur
Laplace
approximation, voir approximation, de Laplace
développement de, voir approximation, de Laplace
règle d'équiprobabilité de, 138
règle de succession de, 192
transformée de, 186
lemme
de Jensen, 76
de Neyman-Pearson, 259
de Pitman-Koopman, 125, 157, 158, 260
de Schur, 209, 211
de Stein, 102
le problème a une valeur, 77
lézard, 115
linéarisation, 61
linéarité, 131
lissage, 213
loi
a posteriori, voir a posteriori
a priori, voir a priori, 11
à support fini, 152
à symétrie sphérique, 35, 109, 158, 282
bêta, 35, 115, 120, 156, 564
bêta-binomiale, 330, 332, 507, 523
bêta-Pascal, 195
binomiale, 565
binomiale négative, 219, 566
généralisée, 459
cible, 326, 329
classe de, 152
conjuguée, 154
d'Erlang, 563
de Bernoulli, 171, 192
de Cauchy, 23, 104, 120, 136, 199, 307, 564
de Dirichlet, 52, 126, 565
de Fisher, 564
de Haldane, 216
de Jeffreys, voir non informatif, a priori de Jeffreys
de l'arcsinus, 217
de Laplace, 14, 44
de Pareto, 125, 128, 142, 157, 565
de Poisson, 129, 484, 519, 565
de probabilité invariante, 474
de Student, 93, 128, 154, 155, 199, 201, 512, 564
décomposition de Dickey pour la, 129, 308
mélange de, 235
de Weibull, 23, 39, 327, 340
de Wishart, 158, 202, 204, 357
double exponentielle, 14
du khi deux (χ^2), 173, 427, 563
décentré, 24, 104, 143, 223, 308, 317, 358, 361, 459, 488, 565
du logarithme itéré, 38
du maximum d'entropie, 118
du rapport de vraisemblance monotone, 272
exponentielle, 563
F, voir Fisher, loi de
géométrique, 220
gamma, 33, 129, 563
inverse, 155, 200, 210, 564
gaussienne inverse, 158
généralisée, 14
hypergéométrique, 3, 194, 566
impropre, 31
jointe, 24
la moins favorable, 31, 77, 79, 80, 263
pour les tests, 262
log-normale, 563
logistique, 218
mélange, voir mélange
marginale, 24, 117, 154, 521
estimation de la, 520
multimodale, 307
multinomiale, 194, 565
naturelle conjuguée, 130

non centrée, 129
non informative, voir non informatif
approximation, 54
normale, 116, 119, 126, 198, 210, 563
inverse généralisée, 128, 161, 478
tronquée, 354, 415
prédictive, 25
stationnaire, 208, 227, 323, 325, 326, 331
symétrique, 157
t, voir Student
unimodale, 167, 270
Loi des Grands Nombres, 67, 315, 324

`Maple`, 244
marche aléatoire, 208, 431
marché boursier, 207
marginalisation, 176
paradoxe de, 138, 163
Markov
chaîne de, 207, 209, 210, 217, 323, 354
apériodique, 326
à temps continu, 496
cachée, 52, 209, 217, 341, 344, 357
ergodique, 331
génération d'une, 325
irréductible, 323, 352
Monte Carlo par, voir MCMC
φ-mélangeance, 352
récurrente, 432
réversible, 350
transiente, 363, 431
Kakutani, 482
modèle de
caché, 341, 358, 394
`Mathematica`, 244
mauvaise spécification, 31
maximum de vraisemblance, voir estimateur du maximum de vraisemblance
estimation, 22
et estimation non paramétrique, 520
méthode du, 555
MCMC, 323, 339
à sauts réversibles, 394
et DIC, 383
et loi impropre, 34
impact des méthodes, 342
pour les modèles
à dimensions variables, 377, 391
de mélange, 342–344
dynamiques, 364
hiérarchiques, 507
non paramétriques, 407
simulation et stockage, 396
médicament, 509
meilleur estimateur équivariant, 465, 471, 472, 527, 557
comme estimateur de Bayes, 468
de Pitman, 467
existence d'un, 466
sous coût entropique, 527
mélange, 34, 44, 52, 129, 157, 163, 218, 270, 290, 312, 364, 401, 503, 550
bêta, 163, 406
caché, 129, 185, 339, 361
continu, 137
d'échelle, 291
de masses de Dirac, 136
de Student, 235
échantillonnage de Gibbs pour, 344
estimation de, 561
exponentiel, 356, 394
géométrique, 27
non-identifiabilité, 408
normal, 4, 23, 45, 257, 310, 342, 372, 410
uniforme, 167
mesure
de Haar, 31, 465, 479, 482
à droite, 118, 148, 473, 475, 477, 481
à gauche, 475
droite contre gauche, 477
finitude, 475
de Lebesgue, 14, 30
de Radon, 475
de référence, 118
invariante, 474
méta analyse, 28, 496, 501
méta modèle, 117, 405
météorologiste, 70, 96, 245

méthode
de filtrage particulaire, 319
de Monte Carlo, voir Monte Carlo
de Simpson, 314
des moments, 117, 120
ML-II, 117
métro de Londres, 416
minimax
analyse, 75
estimateur, 73, 76
randomisé, 76
règle de Bayes, 80
stratégie, 76
minimaxité, 73, 91, 148
et admissibilité, 81, 450
minimaxité, 518
condition nécessaire et suffisante de, 516
d'estimateurs bayésiens hiérarchiques, 512, 514
modèle
à variables latentes, 499
à facteurs, 230
AR(1), 207, 208, 232, 454
AR(p), 208, 210, 212, 213, 364
à sauts, 209
ARCH(p), 230, 233
ARIMA, 229
ARMA(p, q), 214
à variables latentes, 341, 460
à volatilité stochastique, voir volatilité
causal, 229
censuré, 357
choix de, 17, 239, 245
complet, 400, 403
projection du, 400
construction d'un, 6
de calibration linéaire, voir calibration
de capture-recapture, voir capture-recapture
de Darroch, 196, 222
de dimension variable, 391
de Markov, voir Markov, chaîne de, cachée
de Wolter, 196, 222
décomposable, 545
de mélange, voir mélange
discret, 191
dynamique, 206, 207
échangeable, 503
exploration de, 402
GARCH, 233
graphique, 341, 506, 544
hiérarchique, 48, 117, 334, 337, 353, 360, 497
échangeable, 510
et densité conditionnelle complète, 506
et robustesse, 504
imbrication, 370, 376, 403, 405
imbriqué, 299
linéaire, 501
additif, 500
à effets aléatoires, voir effets aléatoires
généralisé, 387, 399, 420, 502
logistique, 2
logit dichotomique, 46
MA(q), 212, 214
moyennage de, 342
moyenne de, 377, 395
multinomial, 415
normal, 198–206
avec a priori hiérarchique, 509
probit, 160, 355, 533
dichotomique, 46
qualitatif, 132
saturation de, 391, 410
spatial autorégressif, 384
statistique, 7, 115
invariant, 468
temporel, 222
tobit, 226
volatilité stochastique, 499
vrai, 115
module, 475, 489
moment
canonique, 42
méthode de, 365
Monte Carlo
approximation par, 179, 322
définition, 315
par fonction d'importance, 316
populationnel, 319
résolution par, 137

moving average, voir MA
moyenne
a posteriori, 85
et théorème de Hunt-Stein, 481
fonctionnelle, 481
géométrique, 403, 415
harmonique, 386, 387
invariante à droite, 482
la plus élevée, 230
mobile, 212
multicolinéarité, 206, 513, 535
minimaxité, 513
multiplicateur, 475, 487
de Lagrange, 118

naissance et mort, 394
processus à temps continu, 395
processus de saut, 418, 419
naissances mâles, 138
Neyman-Scott, problème de, 54
nœud, 544
Le Nom de la Rose, 398
nombre de composantes, 366, 367, 373
inconnu, 372
non informatif, 137
a priori, 50, 51, 123, 137–151
comme limite de conjugués, 202
de Haldane, 287
de Jeffreys, 114, 139, 141–143, 147, 150, 151, 166, 200, 232, 248, 250, 560
et loi conjuguée, 123
et reparamétrisation, 138
et structure d'invariance, 463
mesure de Lebesgue comme, 84, 139, 249, 300, 474, 492
hiérarchique, 503
loi, 22
modélisation, 30, 552
réponse
p-value comme, 273
non-centralité, 488
non-identifiabilité, 26
non-transitivité, 111
normalisation
constante de, 132, 133, 235, 376, 385, 390
rapports de, 387
notation, 7, 567–570
nul modèle n'est parfait, 373, 374
numérique
approximation, 313
calcul, 189
intégration, 307, 308, 314, 319
et méthodes de Monte Carlo, 319
méthode, 238
minimisation, 190, 305
outil, 51
technique, 29

objectivité, 132
observation, 6, 7
espace d', 66
imaginaire, 251
virtuelle, voir virtuelle, observation
Occam
fenêtre d', 398
rasoir d', 398
ondelette, 314, 406
base d', 52
de Haar, 52
optimale
décision, voir décision, optimale
optimalité
asymptotique, 147
classique, 555
orbite, 470
ordre, 94, 95, 111, 147
dans $\mathscr{P}$, 59
des préférences, 170
grossier, 170
partiel, 81, 92
relation d', 169
social, 63
total, 67, 73
orthogonal
base, 314
polynôme, 160
régresseur, 398

paradigme, 9
paradoxe, 147
de Borel, 286
de Condorcet, 63
de Jeffreys-Lindley, 34, 247, 249, 291, 299, 300, 378
de l'information, 313, 342

de marginalisation, 32, 138, 146, 163, 225, 400, 492, 493
de projection, 400
de Saint-Pétersbourg, 56, 64
de Simpson, 36, 63
de Stein, 102, 177
de vraisemblance, 555
des statistiques libres, 226
paramétrisation, 149, 401
choix de la, 57, 90
en moyenne, 171
influence, 57, 90
linéaire, 61
naturelle, 89
paramètre
aléatoire, 11
borné, voir borné
d'échelle, 139
d'interêt, 143
de non-centralité, 223, 445, 488, 539
de nuisance, 143, 144, 147, 148, 174, 179, 445, 561
de position, 30, 139, 465
espace de, 66
naturel, 91, 138
sous contrainte d'ordre, 23
parcimonie, 215, 231, 374, 396
pari, 167
désirable, 167
procédure de, 300
partiel
autocorrélation, 209, 211, 228, 364
inverse, 214
partition de l'échantillon, 312, 313
partitionnement, 138
performance, 67, 81
fréquentiste, 109
perspective conditionnelle fréquentiste, 302
pertinent
sous-ensemble
biaisé négativement, 300
biaisé positivement, 300
phénomène de dégénérescence, 214
Pillow Problems, 221
Pitman
admissibilité, 104
domination au sens de, 103
proximité de, 104, 111
pixel, 4, 105
placebo, 509
Pluralitas non est ponenda sine necessitate, 398
point-selle, 365
polynôme
d'Hermite, 314, 346
Legendre, 406
orthogonal, 160
quadrature par, 314
population
estimation de la taille d'une, 194
finie, 192
rare, 193
précision, 181
évaluation, 86
prévision, 8, 182, 213, 277
cohérente, 168
conjuguée supérieure, 168
densité de, 367
pour la régression linéaire, 226
supérieure et inférieure, 167
prévisionniste, 70
principe
d'exhaustivité, 16, 17, 20
d'invariance, 464
formelle, 465
de conditionnement, 20, 21
de dualité, 331, 353
de la raison insuffisante, 114, 138
de parcimonie, 215, 231, 364, 396, 398
de vraisemblance, 17, 21, 71, 151, 203, 216, 251, 267, 301, 484, 554
et a priori de Jeffreys, 142
et Théorie de la Décision, 88
implémentation, 23, 35
justification, 20
mise en œuvre, 177, 179
version bayésienne du, 175
des règles d'arrêt, 19, 264
des zéros séparés, 80
minimax, 74
probabilisation, 550
probabiliste
interprétation, 3
modélisation, 3, 9, 11, 12

probabilités
axiomes des, 50
imprécises, 496
inverses, 9
théorie des, 13
problème
de Fieller, 47
de Neyman-Scott, 144
inverse, 519
mal posé, 405
NP-complet, 313
processus
bêta, 52
de Dirichlet, 364
de saut, voir naissance et mort, processus de saut
de Lévy, 52
inférentiel
apport subjectif dans un, 559
non stationnaire, 208
stationnaire, 211
programme
d'optimisation, 557
informatique, 315
universel, 556
projection, 475, 510, 541
de Kullback-Leibler, 420
proposition, 326
choix d'une, 326, 327, 329
indépendante, 329
par marche aléatoire, 327, 329, 344
pseudo-a priori, 391
puissance
d'un test, 258
de calcul, 323
du continu, 36
loi, 340
p-value, 266, 267, 273, 439
admissible, 440
comportement conservateur de la, 273

quadrature, 314
quantile, 153
quantité pivotale, 148
queue, 117
épaisse, 136

radiographie, 4
randomisation, 150, 261, 264
et distributions discrètes, 279
Rao-Blackwellisation, 332, 333, 338, 353, 390
rapport des risques, 138
rasoir d'Occam, 398
rationalité, 63
des décideurs, 63
réalisation, 7
récompense monétaire, 62
réconciliation, 271, 301, 561
récurrence, 432
au sens de Harris, 323
d'une chaîne, 431, 433
de Durbin-Levinson, 211, 228
réduction, 68, 115
des principes, 17
et modélisation, 238
par la modélisation, 2–4, 8
référence
a priori
comme loi coïncidente, 150
d'ordre inverse, 151
pour la calibration linéaire, 47
région
à queues égales, 280
HPD, 28, 148, 149, 283, 284, 542
α-crédible, 278
non connexe, 280
uniformément plus précise, 281
règle
d'arrêt, 19
de score, 96
propre, 96
régression
et calibration, 47
inverse, 226
isotonique, 39
linéaire, 132, 179
logistique, 132, 160, 218, 239, 350, 355, 400, 533
modèle de, 203, 206
non paramétrique, 6
normale, 413
poissonienne, 4
regroupement, 337
rejet d'une hypothèse nulle, 264
renouvellement, 324

reparamétrisation, 129, 138, 211, 214, 215
invariance par, 101, 142
naturelle, 158
non paramétrique, 406
répétabilité des expériences, 67, 117, 170, 551, 552, 559
réponse la moins favorable, 268, 269
réponse non informative
pour des tests, 245
représentation
a priori, 553
de la matière, 553
espace d'état, 213, 214, 364
pour ARMA(p, q), 215
forward-backward, 358, 532
intégrale, 557
markovienne, 431
non paramétrique, 171
par mélange caché, 362
par polynôme retard, 228, 229
polynomiale orthogonale, 406
probabiliste, 551
stationnaire, 232
réseaux
bayésiens, 549
de neurones, 52, 549
restaurant chinois, 48
rétrécisseur
estimateur à, voir estimateur
rétroaction a priori, 162
RIC, 414
risque
amateurs de, 64, 85
aversion au, 64, 65
constant, 79, 82
continu, 434
ensemble de, 443
estimateur sans biais du, 102, 108, 515
fréquentiste, 66
intégré, 68
maximin, 77
minimax, 73
vecteurs de, 77
robustesse, 31, 91, 120, 124, 514, 518
dans la fonction de coût, 155
et analyse hiérarchique, 558
rumeur, 286
`R`, 360
sans biais, 15, 108
sauts réversibles, 397
Schwarz
critère de, 381
score de Brier, 95
sélection de variables, 372, 396, 400
descendante, 402
montante, 402
sélection, 230
séparateur, 545
séries temporelles, 206
Shakespeare
vocabulaire de, 198
SIDA, 496
signal
traitement du, 549
significativité
niveau de, 259, 261, 264, 266
usuel, 271
simulation, 306
acceptation-rejet, 346, 347
alternative à la, 319
bases, 315–319
d'une loi conditionnelle, 330, 339
itérative, 340
résultats, 53
software, 313
sommet, 544
`S-Plus`, 361
stationnarité, 208, 209, 215, 324
contrainte de, 208
Statistique
bayésienne
définition de la, 10
non paramétrique, 51
fiduciaire, 50, 551
linguistique, 198
mathématique, 51
perspective conditionnelle dans la, 14
statistique
complète, 37, 38
d'ordre, 37, 129
définition, 15
du rapport de vraisemblance, 173
exhaustive, 15, 125, 199
minimale, 16
invariante maximale, 471

libre, 37, 38, 298, 301, 302, 471
Stein, 15, 106, 108, 479
condition d'admissibilité, 143
de effet, 497
effet, 92, 102, 186, 264, 282, 428, 513, 523, 525, 527, 558
analyse fréquentiste de l', 107
et espace des paramètres fini, 108
robuste, 108
stochastique
complexité, 151
domination, 92, 111
STUB, voir bornes inférieures de demi-queue
subjectivité, 123
suite
échangeable, 171
infiniment échangeable, 171
surajustement, 373
surharmonicité, 516, 541

table de contingence, 217
technique de saturation, 393
test, 34, 239
bilatèral, 250
comme problème d'estimation, 237
de Neyman-Pearson, 18
de Student, 301
du khi deux, 264, 405
du rapport de vraisemblance séquentiel, 301
répété, 301
sans biais, 261
uniformément plus puissant (UPP), 259
uniformément plus puissant sans biais (UPPS), 262
minimax, 298
théorème
Central Limit, 4, 198, 199, 264
de Basu, 37
de Bayes, voir Bayes, théorème
de factorisation, 16, 21
de Fubini, 43, 69
de Hammersley-Clifford, 41, 352
de Hunt-Stein, 81, 479, 482, 483, 490
de la double projection, 421
de Markov-Kakutani, 482
de Rao-Blackwell, 16, 17, 76, 81, 98, 333, 480
de représentation de Riesz, 444
ergodique, 323, 324
hyperplan séparateur, 443
théorie
de l'information, 151
de la connaissance, 553
de la Décision, 8, 63, 112, 237
bayésienne, 184
fondements de la, 285
fréquentiste, 71
de Neyman-Pearson, 89, 239, 258, 274, 280
des jeux, 58, 69, 73, 75
des tests, 237
total
ignorance, 137
ordre, 63, 67, 68
traitement
analytique, 341
commodité du, 124
du signal, 118, 211, 358
tramway, 193
transformation d'échelle, 469
transience, voir Markov
transition
noyau de, 326
transitivité, 63, 169

utilité
construction de l', 170
existence de la fonction d', 63

vache laitière, 117
valeur, 80
aberrante, détection, 364
validation asymptotique
des estimateurs de Bayes, 54
des méthodes non paramétriques, 6
des méthodes bayésiennes empiriques, 525
du maximum de vraisemblance, 23
variable
auxiliaire, 308, 340, 341, 354
explicative, 372
latente, 313

variance
    sous-estimation de la, 373
    estimation d'une, normale, 282
    fonction, 159
    hétérogène, 233
    inconnue, 204
    sous-estimation de la, 528
VIH, 497
virtuel
    échantillon, 116
    observation, 124, 153, 171, 201, 299
vitesses de galaxies, 371, 419
volatilité, 233, 235, 341
vraisemblance, 9, 17, 21, 170
    définition récursive de la, 213
    explicite, 213
    observée, 233
    principe, voir principe de vraisemblance
    profilée, 179
    rapport de, 242, 262, 380

Wheel of Time, The, 1, 55, 113, 175, 237, 305, 369, 423, 463, 495, 549
within-between, 325

Achevé d'imprimer sur les presses de l'Imprimerie BARNÉOUD
B.P. 44 - 53960 BONCHAMP-LÈS-LAVAL
Dépôt légal : novembre 2005 - N° d'imprimeur : 511.076
Imprimé en France