MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE MOULOUD MAMMERI DE TIZI-OUZOU

FACULTE DES SCIENCES
DEPARTEMENT DE MATHEMATIQUES

MEMOIRE DE MAGISTER(ECOLE DOCTORALE)

SPECIALITE : MATHEMATIQUES
OPTION : STATISTIQUE

Présenté par

Djoweyda Ghouil

Sujet :

Aspects de la robustesse Bayésienne dans les Modèles AR(1)

Devant le jury d'examen composé de :

Hamadouche Djamel	Professeur	UMMTO	Président
Fellag Hocine	Professeur	UMMTO	Rapporteur
Boudiba Mohand Arezki	Maître de conférences A	UMMTO	Examinateur
Atil Lynda	Maître de conférences B	UMMTO	Examinatrice
Belkacem Cherifa	Maître Assistante A	UMMTO	Membre invité

Remerciements

Tout d'abord, je tiens à remercier le Bon Dieu, sans qui ce mémoire n'aurait pu exister.

Je présente mes sincères remerciements à mon directeur de mémoire, Monsieur Fellag Hocine , Professeur à l'UMMTO pour ses nombreux conseils, sa grande disponibilité, son aide précieuse et son soutien constant. Ses multiples compétences et sa rigueur ont beaucoup contribué à l'aboutissement de ce travail. Qu'il trouve ici toute ma profonde gratitude.

Je remercie vivement Monsieur Hamadouche Djamel, Professeur à l'UMMTO, pour l'honneur qu'il me fait en acceptant d'être président du jury.

J'adresse mes vifs remerciements à Monsieur Boudiba Mohand Arezki, Maître de conférences à l'UMMTO d'avoir bien voulu participer à ce jury, et d'avoir pris le temps de lire et juger ce travail.

Grand Merci à Atil Lynda, Maître de conférences à l'UMMTO pour m'avoir fait bénéficier de ses critiques objectives et ses conseils avisés et d'avoir pris le temps de lire, apprécier et juger ce travail.

Je remercie aussi Belkacem Cherifa, Maître assistante à l'UMMTO d'avoir pris le temps de lire et juger ce travail.

J'ai eu à effectuer un stage à l'univercité Rey Juan Carlos de Madrid auprès de Monsieur David Rios Insua. Je le remercie vivement pour m'avoir fait bénéficier de sa rigueur scientifique et pour l'accueil qu'il a su me réserver.

Je m'incline respectueusement devant les deux êtres à qui je dois mon existence, mon père et ma mère. Je leur exprime mes profonds signes de reconnaissance et d'obéissance pour tous les efforts qu'ils ont fournis et tous les sacrifices qu'ils ont généreusement faits, pour que je grandisse dans de parfaites conditions d'amour, de satisfaction et d'épanouissement. C'est à eux que je dédie le résultat de mon travail ainsi qu'à la mémoire de ma chère grande mère et tous mes frères et sœurs.

Je souhaite aussi remercier tous mes frères surtout Mohammed et Salah pour leurs compréhension et leurs soutien au long de ces années, ainsi que toutes mes sœurs surtout Saida et Rokia, qui ont vécu avec moi tous les états d'âme à travers lesquels je suis passé tout au long de ce travail.

Enfin je n'oublie pas de remercier chacune de mes adorables amies : Hadjila, Fadhila, Dalila, Houria, Amira, Samia et Naima, ainsi que toute personne qui m'a aimé, aidé, soutenu et cru en moi. Avec eux, j'ai partagé des moments agréables et inoubliables.

Table des matières

Introduction générale

Il subsiste, chez beaucoup, une représentation de la statistique Bayésienne fondée sur l'arbitraire de la loi a priori, point de référence qui ne saurait être remis en cause, tout en déterminant l'inférence résultante.

Si telle est la vision présente dans la littérature du XX^e siècle, cette approche de la statistique Bayésienne a totalement cédé la place à des perspectives beaucoup plus pragmatiques, tout d'abord dans les années 1990 avec l'apparition des méthodes de calcul spécialement adaptées qui permettent d'affiner le choix de la loi a priori et d'en étudier les conséquences. En sus, l'adoption de cette perspective par un nombre croissant de praticiens d'autres disciplines fait que l'utilisation de l'approche Bayésienne relève de moins en mois du dogme et de plus en plus de considérations pratiques de maniabilité et d'efficacité.

Durant ces dernières années et grâce à l'apparition des algorithmes de calcul et plus particulièrement les algorithmes MCMC, l'approche Bayésienne a considérablement accru sa visibilité dans divers domaines et pour divers modèles ; partant de la statistique médicale en passant par la biologie et l'agronomie, en allant au traitement du signal, à l'économie et à la finance.

Un de nos objectifs est de guider le lecteur à se familiariser un peu dans la découverte de l'inférence Bayésienne. Quatre idées doivent motiver cette découverte : L'inférence Bayésienne n'est pas récente ; elle apparaît supérieure sur le plan théorique ; elle est une inférence naturelle et flexible et enfin, elle est et elle va devenir de plus en plus facilement et largement utilisable. On peut pour s'en convaincre, consulter les tables de matière (et les éditoriaux) des grands journaux de statistique comme Annals of Statistics, journal of the American Statistical Association , journal of the Royal Statistical Society ou Biometrika, ainsi que le programme des conférences internationales de statistique, comme par exemple le congrès de Bernoulli ou encore les congrès de l'IMS, de l'ISI et de l'ASA.

Pour un modèle paramétrique bien défini, de densité $f(x|\theta)$ où θ est le paramètre indiçant la densité, à valeurs dans l'espace Θ. La démarche Bayésienne consiste à traiter le paramètre inconnu θ comme une variable aléatoire en lui associant une loi de probabilité sur l'espace Θ dite loi a priori et notée $\pi(\theta)$. Cette loi représente pour un statisticien Bayésien, l'ensemble des informations a priori disponibles sur le paramètre θ ainsi que les imprécisions qui s'y rattachent et dans un contexte pratique, elle regroupe aussi l'ensemble des opinions des experts.

Le choix de la loi a priori a constitué pendant longtemps le point le plus criticable et le plus critiqué de l'analyse Bayésienne par les non Bayésiens. Sur le plan pratique de l'approche Bayésienne, il n'existe jamais une unique loi a priori pour θ, mais plutôt un ensemble de lois compatibles avec les informations a priori disponibles et les opinions des modélisateurs. Tout au plus on peut chercher à réduire l'influence de ce choix en retenant

des lois a priori à faible contenu informatif, comme les lois de maximum d'entropie, ou des lois dites non informatives, comme les lois de Bernardo-Berger, qui minimisent l'information apportée par la loi a priori face à celle apportée par les données. Il existe différentes méthodes pour transformer l'information a priori en une loi a priori, certaines d'entre elles sont abordées dans le premier chapitre du mémoire.

Une fois cette loi a priori est construite, le théorème de Bayes rassemble l'information apportée par la loi a priori avec celle apportée par les données dans une nouvelle distribution dite la distribution a posteriori notée $\pi(\theta|x)$, et qui est le pendant de la vraisemblance dans l'approche classique, de fait que toute inférence au sens Bayésien est basée sur cette distribution a posteriori. En général, le calcul des quantités d'intérêt Bayésiennes comme les estimateurs et les fonctions prédictives prend la forme de l'évalution des intégrales

$$\int g(\theta)\pi(\theta|x)d\theta$$

d'une fonction $g(\theta)$ à l'égard de la distribution a posteriori. L'évaluation de ces intégrales n'est pas évidente et elle a été un inconvénient de l'utilisation de l'approche Bayésienne jusqu'à l'apparition des algorithmes MCMC ; qui ont pu répondre à ce problème et qui ont rendu l'approche Bayésienne trés flexible et appliquable à presque tout type de modèles. Une discution de ces algorithmes est donc trés nécessaire. C'est quoi une méthode MCMC ? pourquoi ? et comment l'appliquer ? est l'objet d'intérêt de la dernière partie du premier chapitre de notre mémoire.

Beaucoup de phénomènes rencontrés dans la vie pratique se modélisent par des modèles des séries temporelles, les modèles autoregressifs sont souvent les plus importants et les plus utilisés grâce à leur simplicité par rapport aux autres. L'objectif principal d'errière toute étude d'une série temporelle est de prévoir les valeurs futures de la série à partir de ses valeurs passées. Dans un cadre Bayésien, Nous visons par ce travail aussi à éclairer la démarche Bayésienne dans la mise en place d'une inférence d'un modèle autoregressif AR(1) comme un type d'une série temporelle.

Un autre objectif très important d'errière ce mémoire et qui est un objet de recherche jusqu'à présent chez tous les statisticiens est de construire des estimateurs robustes. Comme nous avons déjà dit, dans la pratique, il existe un ensemble de lois a priori qui sont compatibles avec les informations a priori disponibles. La robustesse Bayésienne consiste à mettre toutes les lois compatibles avec l'information a priori dans une classe et évaluer les changements effectués sur les quantités a posteriori quand la loi a priori varie dans cette classe.

Ce mémoire contient trois chapitres, afin d'aider le lecteur intéressé à avoir quelques idées sur la démarche Bayésienne qui lui permettent de se plonger dans ce vaste domaine.

Le premier chapitre est un aperçu sur quelques méthodes de la construction d'une loi a priori et de la démarche Bayésienne dans la mise en place d'une inférence sur un paramètre θ. Dans le deuxième chapitre nous avons rassemblé le bagage nécessaire pour effectuer une étude de la robustesse dans un sens Bayésien, et dont nous aurons besoin pour établir notre futur travail qui consiste à construire des estimateurs robustes pour un modèle AR(1). Le troixième chapitre aborde toujours le même sujet mais dans une structure de dépendance en étudiant un type de modèles des séries temporelles ; le modèle AR(1) avec une application.

Chapitre 1

L'analyse statistique Bayésienne

1.1 Introduction

Apprendre en observant, est l'objet principal de l'analyse statistique : Il s'agit de développer les outils et le cadre permettant de mieux comprendre un phénomène observé, en vue d'aider à une prise de décision ou parfois, plus simplement en vue de déceler des structures complexes du processus qui engendre les données. Face à l'incertitude de la complexité du phénomène observé, deux approches statistiques s'opposent. La première 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 régréssion, etc.). Cette approche est dite non paramétrique. En revanche, l'approche paramétrique représente la distribution des observations par une fonction de densité $f(x|\theta)$, où seul le paramètre θ est inconnu.

Nous allons nous intérésser dans notre travail qu'à l'approche paramétrique, nous supposons que les observations x_1, x_2,..., x_n sur lesquelles l'analyse statistique se fonde proviennent de lois de probabilités paramétriques. Donc que x_i $(1 \leq i \leq n)$ a une distribution de densité $f_i(x_i|\theta_i, x_1, ..., x_{i-1})$ sur $\mathbb{R}^n$, telle que le paramètre θ_i est inconnu, mais appartient à un espace Θ de dimension finie, que la littérature scientifique appelle souvent ensemble des états de la nature, et que la fonction f_i soit connue. Et nous allons nous intérésser à établir une inférence sur les θ_i.

Pour mieux comprendre considérons que x est la concentration d'une substance indésirable dans un milieu donné. On dit qu'une norme x_0 est respectée si la probabilité de dépassement est inférieur à une tolérance fixée. Un modèle statistique paramétrique trés souple est la loi gamma dont la densité $[x|\ \alpha,\ \beta]$ implique un paramètre de forme, $\alpha > 0$, et un paramètre d'échelle $\beta > 0$. La probabilité de dépasser la norme x_0 est conditionnelle

aux valeurs prises par ces paramètres :

$$P(x > x_o|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_{x_0}^{\infty} x^{\alpha-1} e^{-\beta x}\, dx$$

où $\Gamma(\alpha)$ est la fonction eurélienne gamma.

Certes, on ne peut pas calculer cette probabilité tant que le paramètre $\theta = (\alpha, \beta)$ reste inconnu. Bien sûr, on peut à l'aide d'une méthode appropriée estimer une partie du plan $\mathbb{R}^2$ dans laquelle la vraie valeur de θ a toutes les chances de se trouver. Évidemment, plus on restreint ce dommaine, plus le risque d'exclure la vraie valeur de θ est grand. A contrario, plus on l'agrandit, plus on a l'incertitude, car on a une valeur de la prbabilité de dépassement pour chaque valeur de θ. Certaines seront sous le seuil de tolérance admises, les autres seront au-dessus. Finalement, comment décider ?.

Il existe deux écoles statistiques opposées pour établir cette inférence sur le paramètre θ.

L'école classique qui attribue à θ une vraie valeur inconnue mais certaine. Pour un statisticien classique, le mode de raisonnement est toujours le même quel que soit le paramètre θ. Dans sa tête, θ a une valeur unique et pour l'estimer il construit une statistique dont les paramètres dépendent de θ. Les données disponibles permettent de calculer un intervalle de confiance correspondant à un risque α fixé. Le paramètre inconnu θ est ou n'est pas dans cet intervalle. Aussi pour décrire son incertitude sur θ, le statisticien classique imagine une collection d'échantillons recueillis dans les mêmes conditions et, pour chacun d'entre eux, il calcule un intervalle de confiance et conclut en disant que $1 - \alpha$ pour cent d'entre eux contiendraient θ, c'est la vision fréquentiste : tout est dans les données.

Toutefois, que faire alors avec tous les problèmes bien concrets où ces répétitions imaginaires n'ont pas de sens ? Comment accepter que plusieurs techniques d'estimation (méthode des moments, du maximum de vraisemblance, etc.) puissent produire des intervalles de confiances différents ? Pourquoi cette fiabilité est-elle quasi systématiquement donnée en situation asymptotique, alors que dans la majorité des problèmes la taille de l'échantillon est trés limitée ?.

La deuxième école est l'école Bayésienne, qui par opposition considère le paramètre du modèle statistique, $[x|\theta]$, incertain. Le statisticien Bayésien va donc chercher à quantifier son incertitude en mobilisant toutes les informations disponibles. C'est ce qui fait toute la différence puisque cela revient à conférer au paramètre θ le statut de variable aléatoire. Dès lors, il lui attribue une distribution de probabilité qui décrit le savoir actuel sur ce paramètre et qui quantifie l'état de connaissances d'un expert sur le problème en main. Cette distribution de probabilité est appelée la distribution a priori, et il est préférable que le savoir de l'expert encodé dans la loi a priori soit indépendant de l'échantillon en main.

Il faut bien comprendre que lorsque un Bayésien parle de probabilité, il ne la conçoit pas comme une fréquence limite dans une succession d'essais dans laquelle on rapporte le nombre de cas favorables au nombre d'essais effectivement réalisés. La probabilité Bayésienne est le résultat d'un pari, propre à l'individu, donc subjectif, mais pas arbitraire.

Dans ce premier chapitre introductif, nous présenterons superficiellement les notions et les outils sur lesquels se fonde une analyse Bayésienne, et dont nous aurons besoin pour établir les prochains chapitres de ce mémoire. Dans un premier temps et aprés avoir donné comment construire une loi a posteriori, sur laquelle l'approche Bayésienne se fonde, nous allons parler de la loi a priori qui est le moteur de l'analyse Bayésienne, et est au même temps, la source de sa difficulté. Par la suite, dans la section 3 du chapitre, nous allons voir comment estimer, tester et prévoir au sens Bayésien. Et enfin, la dernière section présente des méthodes de calcul Bayésien, qui sont globalement numériques et qui ont rendu l'approche Bayésienne plus régide et performante ; les méthodes MCMC.

1.2 Le Paradigme Bayésien

En modélisant des paramètres inconnus de la distribution d'échantillonnage à travers une structure probabiliste, donc en probabilisant l'inconnu, l'analyse statistique Bayésienne autorise un discours quantitatif sur ces paramètres. Elle vise à exploiter le plus efficacement possible l'information apportée par x sur le paramètre θ, pour ensuite construire des procédures d'inférence. Bien que x ne soit qu'une réalisation aléatoire d'une loi gouvernée par θ, elle apporte une actualisation aux informations préalablement recueillies par l'expérimentateur. Elle permet aussi l'incorporation de l'information a priori et de l'imprécision de cette information dans la procédure inférentielle, 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. le concepte fondamental du paradigme Bayésien est la distribution a posteriori qui est un résumé complet de l'information disponible sur le paramètre θ lui même, qui est contenue dans une loi dite loi a priori et de celle apportée par les observations x_i. Aprés avoir présentée la formule de Bayes, nous discuterons la spécification de la distribution a priori en analyse Bayésienne.

1.2.1 la formule de Bayes

Considérons un modèle statistique où la loi de probabilité $f(x_i|\theta)$ qui génère les observations est donnée par un modèle paramétrique particulier qui dépend d'un paramètre inconnu de dimension n, $\theta \in \mathbb{R}^n$.
On dispose d'un échantillon i.i.d $x = (x_1, ..., x_n)$ dont la fonction de vraisemblance s'écrit

$$l(\theta, x) = f(x|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \tag{1.1}$$

Conceptuellement, le statisticien peut exprimer ses connaissances a priori sur θ à travers une distribution a priori $\pi(\theta)$. Une fois ces deux distributions données, nous pouvons en construire plusieurs d'autres.
La distribution jointe de (x, θ) s'obtient par

$$f(x, \theta) = f(x|\theta)\ \pi(\theta) \tag{1.2}$$

La formule de Bayes est basée sur la décomposition inverse de (1.2)

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

On obtient donc la densité a posteriori de θ conditionnelle à x

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{m(x)} \tag{1.4}$$

avec m(x) ne dépend pas de θ, et est la densité prédictive de x, c'est la constante d'intégration de (1.2)

$$m(x) = \int_{\Theta} f(x|\theta)\pi(\theta)\ d\theta \tag{1.5}$$

la formule de Bayes dans (1.4) est approximative à

$$\pi(\theta|x) \propto f(x|\theta)\pi(\theta) \tag{1.6}$$

Ce qui est souvent utile pour un statisticien est d'étudier le comportement d'une valeur future de x notée y, réplication d'une future observation, étant donnée l'information déjà récoltée à l'observation de x. Sous l'hypothèse que conditionnellement à θ, y est indépendante de x, on obtient la densité prédictive de y

$$f(y|x) = \int_{\Theta} f(y|\theta)\pi(\theta|x)\ d\theta \tag{1.7}$$

Par ailleurs, le paramètre d'intérêt est souvent multidimensionnel

$$\theta = (\theta_1, ..., \theta_n) \in \mathbb{R}^n$$

Dans ce cas, à partir des densités jointes (a priori ou a posteriori) de θ, on peut obtenir toutes les densités marginales (a priori ou a posteriori) de chaque composante θ_i de θ par intégration des autres composantes. Par exemple, a posteriori

$$\pi(\theta_i|x) = \int \pi(\theta|x)\ d\theta_1, ..., d\theta_{i-1}, d\theta_{i+1}, ..., d\theta_n \tag{1.8}$$

D'un point de vue pratique, le choix de la loi a priori est souvent perçu comme une difficulté majeure de l'approche Bayésienne en ce que l'interprétation de l'information a

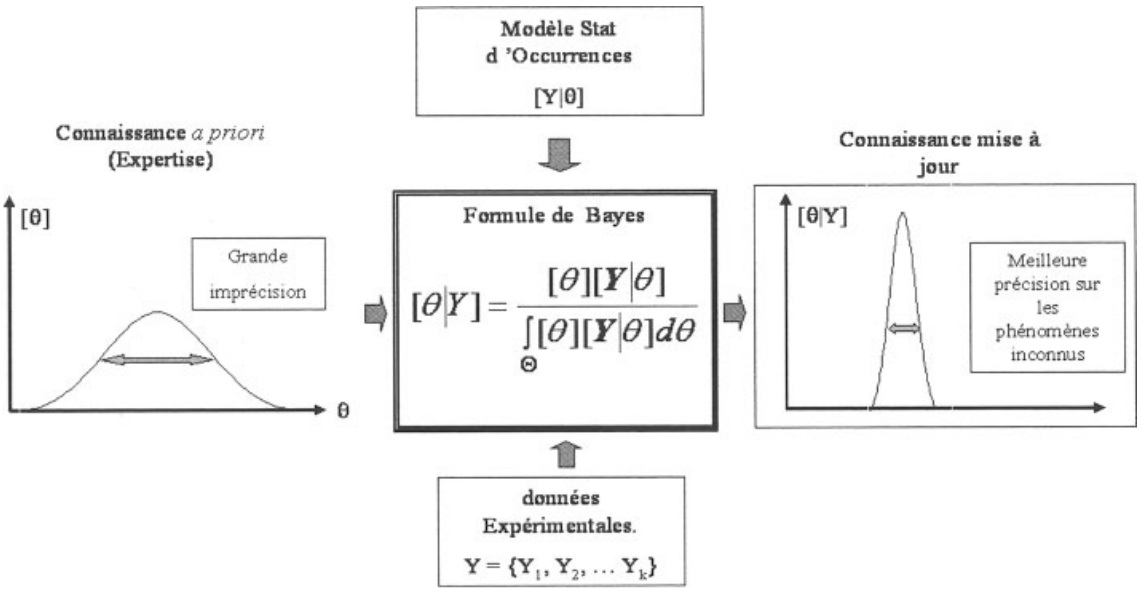

FIG. 1.1 – Le paradigme Byésien

priori disponible est rarement assez précise pour conduire à la détermination d'une seule et unique loi a priori. Il existe néanmoins des lois calibrées en fonction de la distribution des observations, dites lois conjuguées, et des lois à faible contenu informatif, dites lois non informatives, qui permettent d'évaluer l'influence d'une loi a priori donnée. Une autre difficulté de l'approche Bayésenne est le calcul des procédures Bayésiennes ; comme l'espérance a posteriori ou la distribution prédictive, cependant cette difficulté ne doit pas être considérée comme un inconvinient majeur, car la statistique numérique est actuellement en train de subir un dévelopement trés rapide et elle nous permet d'évaluer des telles procédures en utilisant des techniques de simulation ; comme mous le verrons dans la section 4 du présent chapitre.

Ce que nous devons constater est que la statistique Bayésienne repose sur la loi a posteriori, qui peut s'interpréter comme un résumé de l'information disponible sur θ, une fois x observé car une fois que cette loi 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 pour obtenir la distribution prédictive.

La figure (1.1) ci-dessus synthétise le mécanisme de l'approche Bayésienne, où deux modèles doivent être spécifiés, le modèle d'échantillonnage et la loi a prior.

1.2.2 la spécification de la distribution a priori dans l'analyse Bayésienne

La loi a priori est le moteur de l'inférence Bayésienne et sa détermination est donc l'étape la plus importante dans la mise en œuvre de cette inférence. La loi a priori est une transformation sous une distribution de probabilité du savoir de l'expert qu'on appelle encore l'expertise déjà connu sur le problème en main, en dehors des informations apportées par les résultats expérimentaux.

Comment passer des informations a priori à des lois a priori ? est la question fondamentale et légitime dans la mise en œuvre de toute analyse au sens Bayésien, et qui a constitué pendant longtemps la pierre d' achoppement entre l'école classique et l'école Bayésienne.

Effectivement, le statisticien classique pose le principe que seules les données doivent être utilisées pour l'inférence sur le paramètre θ, c'est-à-dire qu'il utilise l'information x pour améliorer sa connaissance de θ. Dès lors, il réfute l'introduction de l'expertise au nom d'une prétendue objectivité nécessaire à la procedure d'inférence sur θ. En fait, la subjectivité est inévitable dans la modélisation probabiliste, depuis la sélection des variables surveillées jusqu'aux conclusions-recommandations en passant par le choix du modèle de connaissance. La démarche scientifique ne consiste donc pas à nier la subjectivité mais bien à la contrôler.

A contrario, l'école Bayésienne a développé un cadre formel pour traduire de façon quantitative l'expertise via des distributions de probabilité a priori.

Évidemment, dans la pratique, il est rare que l'information a priori (l'expertise) 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. Et cela revient à plusieurs raisons ; Parfois le statisticien n'a pas le temps ni les ressources pour construire une loi a priori exacte et il doit compléter l'information partielle qu'il a rassemblé à l'aide des données subjectives, ce qui le pousse à l'interpréter comme une succession de paris sur les valeurs du paramètre, bien sûr sans mobiliser les données impliquées dans la vraisemblance. Mais il faut comprendre que ces paris ne sont pas arbitraires et des méthodes ont été dévéloppées pour les traduire du mieux possilble sous la forme d'une distribution de probabilité, et nous allons exposer quelques unes d'entre elles plus loin. Aussi, l'utilisation systématique de lois usuelles (normale, gamma, bêta, etc.) et la restriction plus forte encore aux lois conjuguées que nous allons définir plus loin, 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.

Toutes 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 ce choix a un impact sur l'inférence résultante qui peut être négligable, modéré ou

énorme, puisqu'il est toujours possible d'obtenir la loi a priori qui conduira à la réponse qu'on souhaite obtenir.

Dans la pratique, l'information a priori peut être codée selon une des façons suivantes :

1. Prendre une loi a priori vague, c'est-à-dire non informative ;
2. Choisir une loi a priori conjuguée à la vraisemblance (commodité mathématique)
3. Déterminer une loi a priori subjectivement.

Lois a priori non informatives

Les lois a prior non informatives représentent une ignorance sur le problème en main, mais ne signifient pas que l'on sache absolument rien sur la distribution statistique du paramètre. En effet, on connaît au moins son domaine de variation, c'est-à-dire l'ensemble des états de la nature, Θ, et le rôle de chaque composante du paramètre sur les observables (paramètre de localisation, d'échelle, etc). Ces lois doivent être donc particulièrement construites à partir de la distribution de l'échantillonnage, puisque, c'est le seul moyen disponible pour avoir des informations sur le paramètre θ. À cet égard, les lois a priori non informatives peuvent être considérées comme des lois de références, auxquelles chacun pourrait avoir recours quand toute information a priori sur θ est absente.

En résumé, quand on dit une loi a priori non informative, il faut comprendre que :

1. Le savoir de l'expert sur le problème en main ne lui permet pas de lier les paramètres

$$\theta_1 \perp \theta_2 \perp ... \perp \theta_n \Rightarrow [\theta_1, ..., \theta_n] = \prod_{j=1}^{n} [\theta_j]$$

2. Toutes les plages de valeurs de θ_j sont, aux yeux de l'expert, équiprobables, c'est-à-dire qu'il ne pariera pas davantage sur une valeurs que sur une autre.

Nous décrirons dans ce qui suit, quelques-unes des techniques les plus populaires dans la construction des lois a priori non informatives.

1. **Lois a priori invariantes**

   Le fait de formaliser l'absence d'information a priori par une propriété d'invariance est naturelle au sens où seuls les paramètres de la distribution de θ changent lorsqu'on effectue une transformation de θ. Par exemple, les distributions de θ et de $\theta - \theta_0$, en réalité, ne sont pas les mêmes, mais dire qu'elles sont les mêmes, c'est-à-dire

   $$\pi(\theta) = \pi(\theta - \theta_0)$$

   pour tout θ_0, exprime certainement l'ignorance sur la valeur de θ.
   On dit dans ce cas que la loi a priori π est invariante par translation, et $\pi(\theta) = c$, la

loi uniforme sur Θ.
Cette technique de construction des lois non informatives 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, ou être sans intérêt pour le décideur.

2. **Lois a priori de Jeffreys**

La spécification de la loi a priori non informative de Jeffreys consiste à assigner à un modèle d'échantillonnage caractérisé par sa vraisemblance $f(x|\theta)$.
Les lois a priori de jeffreys sont fondées sur l'information de Fisher, donnée par

$$I(\theta) = -E_\theta \left[\frac{\partial^2 \log f(x|\theta)}{\partial \theta^2} \right] \tag{1.9}$$

D'où la loi a priori de Jeffreys est

$$\pi(\theta) = I^{1/2}(\theta) \tag{1.10}$$

La loi de Jeffreys n'est pas invariante en général au sens de l'invariance par une famille de transformations, mais elle doit s'entendre comme une invariance par rapport au choix de la paramétrisation, puisque pour une transformation bijective donnée h qui transforme le paramètre θ en $h(\theta)$, nous avons la tansformation Jacobienne

$$I(\theta) = I(h(\theta))(h'(\theta))^2$$

Dans le cas où le paramètre θ est multidimensionnel, la matrice d'information de Fisher s'obtient par généralisation de (1.9). Pour $\theta \in \mathbb{R}^k$, $I(\theta)$ a les éléments suivants :

$$I_{ij}(\theta) = -E_\theta \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(x|\theta) \right], (i, j = 1, ..., k) \tag{1.11}$$

La technique de Jeffreys fournit une des meilleures techniques pour construire une loi a priori non informative ; et elle permet bien souvent de retrouver les estimateurs classiques surtout dans des cas unidimensionnels, mais de sa part, elle a été critiquée par certains Bayésiens comme étant un outil sans justification subjective en terme d'information a priori.

3. **Lois a priori de référence**

Une loi a priori de référence est tout simplement une loi a priori non informative (objective) construite d'une manière particulière. Mais d'une certaine sorte, toutes les lois a priori non informatives sont des lois de référence du fait que chaque loi a priori non informative peut être considérée comme un point de référence auquel chacun pourrait avoir recours quand toute information sur θ est absente.

Cette approche est une modification de l'approche de Jeffreys qui a été proposé par Bernardo (1979), elle repose sur le principe de faire la distinction entre l'importance des paramètres c'est-à-dire entre les paramètres de nuisance et les paramètres d'intérêt. Nous allons donner brièvement le principe de la construction de ces lois en demendant aux lecteurs de se référer à Berger et Bernardo (1989, 1992b, a), Bernardo et Smith (1994) et Kass et wasserman (1996).

Considérons tout d'abord le cas d'un paramètre à deux composantes, $\theta = (\theta_1, \theta_2)$, où θ_1 est le paramètre d'intérêt (de plus importance) et θ_2 est le paramètre de nuisance, et soit $x \sim f(x|\theta)$.

La stratégie introduite par Bernardo est la suivante : pour θ_1 fixé, on détermine tout d'abord la densité conditionnelle $\pi(\theta_2|\theta_1)$ comme la loi de Jeffreys associée à $f(x|\theta)$, puis on calcule $\pi(\theta_1)$ qui est la loi de Jeffreys associée à la loi marginale :

$$\widetilde{f}(x|\theta_1) = \int f(x|\theta_1, \theta_2)\ \pi(\theta_2|\theta_1)\ d\theta_2 \tag{1.12}$$

La loi de référence de θ est le produit des deux lois, c'est-à-dire : $\pi(\theta_1, \theta_2) = \pi(\theta_2|\theta_1)\pi(\theta_1)$. Cette manière de faire peut se généraliser si $\theta = (\theta_1, ..., \theta_n)$, et si l'on a ordonné sans perte de généralité les θ_i par intérêt croissant.

Il est clair que ce raisonnement n'est pas purement objectif parce que donner plus d'importance à un paramètre qu'à un autre relève une fois encore d'un choix.

Lois a priori conjuguées

Ce type de lois a priori est utilisé quand l'information a priori disponible sur le modèle est trop vague ou peu faible. Dans ce cas l'analyste regarde la forme de la fonction de vraisemblance et choisit une famille de lois qui se marie bien avec elle. Par exemple, pour la vraisemblance d'un n-échantillon i.i.d selon une distribution exponentielle de paramètre d'échelle $\rho > 0$ qui est donnée par $\rho^n \ \exp(-n\overline{x}\rho)$, la loi a priori conjuguée est une loi Gamma dont la forme fonctionnelle s'écrit $\rho^{a-1} \ \exp(-b\rho)$ et appliquant le théorème de Bayes, la distribution a posteriori suit encore une loi Gamma : $\rho|a, b, n, \overline{x} \ \sim \mathcal{G}(a \ + \ n, b \ + \ n\ \overline{x})$.

Rappelons ici qu'une famille $\mathcal{F}$ de distributions de probabilité sur Θ est dite conjuguée (ou fermée par échantillonnage) par une vraisemblance $f(x|\theta)$ si, pour toute loi a priori $\pi \in \mathcal{F}$, la distribution a posteriori $\pi(.|x)$ appartient également à $\mathcal{F}$.

L'avantage des familles conjuguées est avant tout la simplicité des calculs. Avant l'essor du calcul numérique, ces familles étaient pratiquement les seules qui permettaient de faire aboutir des calculs. L'intérêt principal du caractère conjugué se manifeste quand $\mathcal{F}$ est paramétrée. Effectivement le passage de la distribution a priori à la distribution a posteriori n'est dans ce cas qu'une mise à jour des paramètres correspondants, ce que nous pouvons le constater dans l'exemple ci-dessus. Et par conséquent, les distributions a posteriori sont toujours calculables dans ce cas.

Définition 1.2.1 ([29]). Familles exponentielles

La famille exponentielle regroupe les lois de probabilité qui admettent une densité de la forme :

$$f(x|\theta) = h(x)\ e^{\alpha(\theta)\ T(x)\ -\ \psi(\theta)},\ \ \theta \in \Theta$$

. T est une statistique exhaustive. Une telle famille est dite régulière si Θ est un ouvert tel que $\Theta = \left\{\theta|\ \int h(x)\ e^{\alpha(\theta)\ T(x)}\ d\mu(x) < \infty\right\}$.
En outre, on appelle paramétrisation canonique, l'écriture : $f(x|\theta) = h(x)\ e^{\theta\ T(x)\ -\ \psi(\theta)}$ et famille naturelle l'expression $f(x|\theta) = h(x)\ e^{\theta}\ T(x)$.

Théorème 1.2.1 ([29]). ***Famille exponentielles***

Si $x \sim f(x|\theta) = h(x)e^{\theta T(x) - \psi(theta)}$, alors la famille de lois a priori

$$\left\{\pi_{\lambda,\mu}(\theta)\ \propto\ h(x)\ e^{\theta\ \mu - \lambda\ \psi(\theta)}, \lambda, \mu\right\}$$

est conjuguée. On note que $\pi_{\lambda.\mu}$ est une densité de probabilité si et seulement si $\lambda > 0$ et $\mu/\lambda\ \in\ \Theta$. La loi a posteriori correspondante est $\pi(\theta|\lambda + 1,\ \mu + T(x))$.

En effet,

$$\begin{aligned} \pi_{\lambda,\mu}(\theta|x) &\propto h(x)e^{\theta T(x) - \psi(\theta)}e^{\theta\ \mu - \lambda\ \psi(\theta)} && (1.13)\\ &\propto h(x)e^{\theta(T(x)\ +\ \mu) - (\lambda\ +\ 1)\psi(\theta)} && (1.14)\\ &= \pi_{\lambda\ +\ 1, \mu\ +\ T(x)}. && (1.15) \end{aligned}$$

Le tableau ci-dessous présente quelques lois a priori conjuguées pour quelques familles exponentielles usuelles.

TAB. 1.1 – lois a priori conjuguées pour quelques familles exponentielles usuelles

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Normale $\mathcal{N}(\theta, \sigma^2)$	Normale $\mathcal{N}(\mu, \tau^2)$	$\mathcal{N}(\varrho(\sigma^2\mu + \tau^2 x), \varrho\sigma^2\tau^2)$ $\varrho = 1/(\sigma^2 + \tau^2)$
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 $\mathcal{M}_k(\theta_1, ..., \theta_k)$	Dirichlet $\mathcal{B}e(\alpha, \beta)$	$\mathcal{D}(\alpha_1 + x_1, ..., \alpha_k + x_k)$
Normale $\mathcal{N}(\mu, 1/\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	$\mathcal{G}(\alpha + 0.5, \beta + (\mu - x)^2/2)$

Lois a priori d'entropie maximale

Si on dispose de certaines caractéristiques de la loi a priori de type $E^\pi[g_k(\theta)] = \mu_k$ (moments, quantiles, etc.), où pour chaque $k = 1, ..., n$, g_k est une fonction donnée. On peut utiliser la méthode de l'entropie maximale développée par Jaynes (1980, 1983) pour déterminer une loi a priori sous ces contraintes.

Pour comparer le caractère informatif, il est nécessaire d'avoir recours à un critère d'information. L'entropie de Shannon permet de définir ce niveau d'informativité.

Dans un cadre fini et discret, cette entropie est définie comme suit :

Pour $\theta \in 1, ..., n$ et $\pi(\theta) = \pi_1, ..., \pi_n$ tel que $\pi_i > 0$ et $\sum_i \pi_i = 1$

$$Ent(\pi) = -\sum_i \pi_i \log(\pi_i)$$

Sans contraintes sur π la distribution d'entropie maximale est la distribution uniforme sur Θ. Une entropie petite s'interprète comme une loi concentrée et informative. La maximisation de l'entropie sous ces contraintes mène à une minimisation de l'information a priori apportée par π sur θ. Le principe à la base de cette méthode est donc de chercher à calculer :

$$\underset{\pi}{Argmax}\ Ent(\pi) \quad \text{sous la contrainte} \quad E^\pi[g_k(\theta)] = \mu_k.$$

La solution de ce problème est alors donnée par :

$$\pi^* \propto e^{\sum_{k=1}^{n} \lambda_k \, g_k(\theta)}$$

où les λ_k sont les multiplicateurs de Lagrange associés qui se déterminent dans la pratique par un système d'équations à partir des contraintes.

L'extension au cas continu est différente, ce n'est pas possible de définir l'entropie comme dans le cas discret puisqu'on ne peut pas dénombrer les états en l'absence d'une mesure de référence. Ceci exige donc le choix d'une mesure de référence π_0 qui peut être caractérisée comme la distribution complètement non informative. Une fois π_0 est choisie, l'entropie de π est donnée par

$$Ent(\pi/\pi_0) = \int_\theta \pi(\theta) \log\left(\frac{\pi(\theta)}{\pi_0(\theta)}\right) d\theta$$

qui est aussi la distance de Kulback entre π et π_0.
Là encore, l'objectif est de maximiser $Ent(\pi/\pi_0)$ sous les contraintes $E^\pi[g_k(\theta)] = \mu_k$ et la solution générale est connue :

$$\pi^*(\theta) \propto e^{\sum_{k=1}^{n} \lambda_k \, g_k(\theta)} \, \pi_0(\theta)$$

Un inconvénient de cette méthode est que la distribution d'entropie maximale dépend du choix de la mesure de référence π_0. Lorsque une structure de groupe est disponible, un choix raisonnable de π_0 est la mesure de Haar invariante à droite. En plus parfois les contraintes ne sont pas suffisantes pour obtenir une distribution sur θ. qui est le cas quands les contraintes sont liées aux quantiles, où les fonctions $g_k(\theta)$ sont de la forme $I_{(-\infty,a_k]}$ ou $I_{(b_k,\infty]}$.

Exemple 1. Soit θ un paramètre réel.
Si l'on choisit la mesure de référence est la mesure de Lebesgue sur $\mathbb{R}$, et si $E^\pi[\theta] = \mu$ alors la théorie donne $\pi(\theta) \propto e^{\lambda\,\theta}$ qui ne peut pas être normalisée comme une distribution de probabilité.
Si de plus on sait que $var(\theta) = \sigma^2$, la loi a priori d'entropie maximale dans ce cas est

$$\pi(\theta) \propto e^{\lambda_1\theta + \lambda_2\theta^2}$$

c'est donc la loi normale $\mathcal{N}(\theta, \sigma^2)$.

Lois a priori subjectives

Précisons tout d'abord que cette démarche n'est pas forcément facile dans la pratique. L'idée est d'utiliser les données antérieures. Par exemple dans un cadre paramétrique, cela revient à présenter des valeurs ponctuelles de θ à l'expert et pour chacune d'entre elles, de lui demander les chances qu'il lui accorde.

TAB. 1.2 – Information a priori sur les paramètres de capture et de suivre 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	y=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. 1.3 – Modèle a priori de capture et de suivre correspondant à l'information du tableau (1.2)(même source)

Épisode	2	3	4	5	6
Dist.	$\mathcal{B}e(6,14)$	$\mathcal{B}e(8,12)$	$\mathcal{B}e(12,12)$	$\mathcal{B}e(3.5,14)$	$\mathcal{B}e(3.5,14)$
Site	A			B	
Épisode	t=1,3,5	t=2,4	t=1,3,5	t=2,4	
Dist.	$\mathcal{B}e(6.0,2.5)$	$\mathcal{B}e(6.5,3.5)$	$\mathcal{B}e(6.0,2.5)$	$\mathcal{B}e(6.0,2.5)$	

Dans une expérience de capture-recapture de lézards, des biologistes s'intéressent aux migrations de ces lézards entre des zones de leur territoire. L'information disponible auprès des biologistes sur les probabilités de capture p_t, et de suivre q_{it}, où t et i représentent le temps et la région considérés, est représentée dans le tableau (1.2) ci-dessus par une moyenne a priori et un intervalle de confiance a priori de 95% pour ces probabilités. Plusieures distributions a priori sont compatibles avec cette information a priori, puisque la distribution $\mathcal{B}e(\alpha,\beta)$ peut être caractérisée par sa moyenne et un intervalle de confiance, le statisticien choisit la distribution a priori bêta présentée dans le tableau (1.3) ci-dessus.

Ces distributons sont dites subjectives parce qu'elles sont propre à l'expert. Elles doivent être interprétées comme un pari de l'expert.

1.2.3 Lois a priori impropres

Une loi impropre (ou généralisée) est une mesure σ-finie sur l'espace des paramètres Θ, c'est-à-dire une mesure π telle que

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

Ces lois sont obtenues lorsqu'on dispose des critères subjectifs ou théoriques sur la distribution a priori du paramètre, qui conduisent à une mesure σ-finie sur Θ plutôt qu'à une mesure de probabilité. Les lois a priori impropres sont utiles dans les modèles non-informatifs

cependant, elles ne peuvent être utilisées que si la condition suivante est vérifiée :

$$m_\pi(x) = \int_\Theta f(x|\theta)\pi(\theta)\ d\theta < \infty$$

En conclusion, l'usage de lois a priori impropres est justifié si la loi a posteriori est propre car elle ne dépend pas de la constante multiplicative de la loi a priori. Dans ce cas ces lois sont utiles du moins tant que la loi a posteriori existe car l'inférence Bayésienne se fonde sur la loi a posteriori $\pi(\theta|x)$.

Une difficulté pratique dans l'utilisation des lois impropres est de vérifier la condition d'intégrabilité

$$\int_\Theta f(x|\theta)\pi(\theta)\ d\theta < \infty$$

surtout dans des situations complexes, comme les modèles hiérarchiques. Cependant cette difficulté ne doit pas être considérée comme un inconvinient car les nouvelles techniques de calcul Bayésien comme les algorithmes MCMC ne nécessitent pas dans la pratique de vérifier cette condition.

1.3 L'inférence Bayésienne

Comme nous l'avons indiqué antérieurement, aucune inférence Bayésienne ne peut se faire sans le calcul de la loi a posteriori, qui est un résumé complet de l'information disponible sur θ. Dans le reste de ce chapitre, nous considérerons que cette loi a posteriori est disponible et nous présenterons dans quelques idées générales comment conduire une inférence sur θ (estimation, tests et prévision) en utilisant cette distribution a posteriori, c'est-à-dire une inférence au sens Bayésien. Et pour plus de détails, nous avons mis à la disposition des lecteurs quelques ouvrages et articles auxquels ils peuvent se référer.

1.3.1 La prédiction

Le contexte du problème de la prédiction est le suivant : à partir d'un échantillon de n tirages x_1, ..., x_n de fonction de densité $f(x|\theta)$, il s'agit de déterminer le plus précisément possible ce que pourrait être le tirage suivant x_{n+1}.

Dans un cadre fréquentiste, il s'agit de calculer $f(x_{n+1}|x_1, ..., x_n, \hat{\theta})$, où $\hat{\theta}$ est l'estimateur de θ. Il est claire que dans ce cas on utilise deux fois les données : une fois pour l'estimation de θ et l'autre fois pour la prédiction.

En revanche, et comme l'école Bayésienne traite le paramètre θ comme une variable aléatoire de loi π. La stratégie ici consiste à intégrer la distribution prédictive par rapport

à la loi a posteriori afin d'obtenir la meilleure prédiction.
La fonction prédictive s'écrit ainsi :

$$f^{\pi}(x_{n+1}|x_1,...,x_n) = \int_{\Theta} f(x_{n+1}|x_1,...,x_n,\theta)\pi(\theta|x,...,x_n)\ d\theta$$

. Notons que l'approche de la théorie de la décision développée dans la section suivante s'applique aussi à la prédiction. En fait, si un coût de prédiction $l(\delta,\theta)$ est disponible, un prédicteur peut être choisi qui mininmise l'erreur de prédiction moyenne (l'espérance étant calculée par rapport à la distribution prédictive). Par exemple, si $l(\delta,\theta) = (\theta-\delta)^2$ le coût quadratique, on peut proposer le prédicteur :

$$\hat{x}^{\pi}_{n+1} = E^{\pi}(x_{n+1}|x_1,...,x_n) = \int x_{n+1}\ f(x_{n+1}|x_1,...,x_n)\ dx_{n+1}$$

1.3.2 L'estimation ponctuelle

Avant de parler de l'estimation, un passage sur la théorie de la décision est trés utile, car comme nous allons le voir, déterminer un estimateur de Bayes revient à déterminer une règle de décision.

Introduction à la théorie de la décision Bayésienne

Un problème de décision en général est fondé sur les trois éléments suivants :
- Un ensemble des actions (décisions) $\mathcal{D}$
- Un espace des paramètres Θ
- Une fonction de coût (de perte) $l(\theta,\delta)$ qui décrit la perte de prendre la décision δ lorsque le paramètre est θ.

Fonctions de perte et risque

Définition 1.3.1. Soit $\delta \in \mathcal{D}$ une règle de décision.
Une fonction de perte (de coût) est une fonction mesurable de $(\Theta \times \mathcal{D})$ à valeurs dans $\mathbb{R}_+$ notée $l(\delta,\theta)$ et définie telle que,

1. $\forall$ (δ,θ) $l(\delta,\theta) > 0$
2. $\forall\theta$, $\exists\ \delta^*$ tels que : $l(\delta^*(x),\theta) = 0$

S'il faut faire un choix entre deux règles de décision, ce choix est impossible sans critère de coût, de sorte à définir correctement la notion de meilleur estimateur.

Définition 1.3.2. Le risque fréquentiste

Pour une fonction de perte donnée $l(\theta, \delta)$, la fonction de risque associée est

$$\begin{aligned} R(\delta, \theta) &= E_\theta[l(\theta, \delta(x))] \\ &= \int_X l(\theta, \delta(x))\ f(x|\theta)\ d\mu(x) \end{aligned}$$

C'est une fonction de θ et ne définit pas un ordre total sur $\mathcal{D}$ et ne permet donc pas de comparer toutes décisions et estimateurs. Il n'existe donc pas de meilleur estimateur dans un sens absolu. Ainsi, l'approche fréquentiste restreint l'espace d'estimation en préférant la classe des estimateurs sans biais dans laquelle il existe des estimateurs de risque uniformément minimal ; l'école Bayésienne ne perd pas en définissant un risque a posteriori. L'idée est d'intégrer sur l'espace des paramètres pour pallier cette difficulté.

Définition 1.3.3. Le risque a posteriori

Pour un Bayésien, θ est une variable aléatoire de distribution a priori $\pi(\theta)$, et aprés que les données seront disponibles, la distribution pertinente de θ sera donnée par la distribution a posteriori $\pi(\theta|x)$ et le risque pertinent sera le risque a posteriori ou bien le risque Bayésien.

$$\begin{aligned} \rho(\pi, \delta|x) &= E^\pi(l(\theta, \delta(x))|x) \\ &= \int_\Theta l(\theta, \delta(x))\ \pi(\theta|x)\ d\theta \end{aligned}$$

Ainsi, le problème change selon les données ; ceci dû à la non existence d'un ordre total sur les estimateurs.

Définition 1.3.4. Le risque intégré

Pour une foction de perte donnée, le risque intégré est défini par

$$\begin{aligned} r(\pi, \delta) &= E(R(\theta, \delta)|x) \\ &= \int_\Theta R(\theta, \delta)\pi(\theta)d\theta \end{aligned}$$

Une fois la loi a posteriori sur le paramètre est disponible, le problème de l'estimation Bayésienne ponctuelle peut être exprimé comme un problème de décision.

Définition 1.3.5. L'estimateur Bayésien

Un estimateur Bayésien est la règle de décision δ^π qui minimise $r(\pi, \delta)$. C'est-à-dire qui vérifie

$$r(\pi, \delta^\pi) = \inf_{\delta \in \mathcal{D}} r(\pi, \delta)$$

Pour obtenir la valeur de l'infimum, il faut en théorie minimiser une intégrale double. En effet, nous cherchons à minimiser la fonction de risque Bayésienne $r(\theta, \delta)$, nous pouvons écrire

$$\begin{aligned}
r(\theta, \delta) &= \int_\theta R(\delta, \theta)\ \pi(\theta)\ d\theta \\
&= \int_\theta \int_x l(\delta, \theta)\ f(x|\theta)\ dx\ \pi(\theta)\ d\theta \\
&= \int_\theta \int_x l(\delta, \theta)\ \frac{f(x|\theta)\ \pi(\theta)}{m_\pi(x)}\ m_\pi(x)\ dx\ d\theta \\
&= \int_x \int_\theta l(\delta, \theta)\ \pi(\theta|x)\ m_\pi(x)\ d\theta\ dx \\
&= \int_x \left\{ \int_\theta l(\delta, \theta)\ \pi(\theta|x)\ d\theta \right\} m_\pi(x)\ dx \\
&= \int_x \rho(\pi, \delta|x)\ m_\pi(x)\ dx
\end{aligned}$$

Et minimiser $r(\pi, \delta)$ pour toute valeur de x, sera donc équivalent à minimiser la fonction de risque a posteriori

$$\rho(\pi, \delta|x) = \int_\theta l(\delta, \theta)\ \pi(\theta|x)\ d\theta$$

La minimisation de cette dernière expression peut se faire analytiquement comme elle peut s'approcher numériquement (par des techniques de simulation) selon la complexité de la fonction de perte l et de la loi a posteriori, $\pi(\theta|x)$. Parfois il est impossible de calculer $\pi(\theta|x)$ et parfois même si elle est connue, l'integration analytique parait impossible, comme le cas des séries temporlles à cause de la complexité de la distribution de vraisemblane. Ce qui nécessite des approximations numériques comme les méthodes MCMC abordées dans la prochaine section.

Pour des fonctions de pertes classiques, les estimateurs de Bayes correspondant sont des caractéristiques usuelle de la distribution a posteriori (moyenne, médiane, fractiles, etc)

Exemple 2. La perte quadratique

Une fonction de perte quadratique est une fonction $l : (\Theta \times \mathcal{D}) \longrightarrow \mathbb{R}_+$ donnée par

$$l(\theta, \delta) = (\theta - \delta)^2$$

Ainsi, soit

$$\begin{aligned} f(\delta, x) &= \rho(\pi, \delta|x) \;=\; E(l(\delta, \theta)) \\ &= \int_\Theta (\theta - \delta)^2 \, \pi(\theta|x) \, d\theta \\ &= \int_\Theta \theta^2 \, \pi(\theta|x) \, d\theta \;-\; 2\delta \int_\Theta \theta \, \pi(\theta|x) \, d\theta \;+\; \delta^2 \int_\Theta \pi(\theta|x) \, d\theta \\ &= E(\theta^2|x) \;-\; 2\delta \, E(\theta|x) \;+\; \delta^2 \end{aligned}$$

La décision δ qui minimise $f(\delta, x)$ est celle qui vérifie

$$\frac{d}{d\delta} \, f(\delta, x) \;=\; 0$$

ce qui donne,

$$-2E(\theta|x) \;+\; 2\delta \;=\; 0$$

et donc,

$$\delta \;=\; E(\theta|x)$$

Donc pour la perte quadratique, l'estimateur de Bayes est la moyenne de la loi a posteriori.

Exemple 3. La perte absolute

De même,nous pouvons vérifier aisément que l'estimateur de Bayes utilisant un coût absolu

$$l(\delta, \theta) = |\theta - \delta|$$

est donné par la médiane a posteriori.
En remplaçant $l(\delta, \theta)$ dans l'expression de $\rho(\pi, \delta|x)$, nous obtenons

$$\begin{aligned} f(\delta, x) &= \int_\Theta |\theta - \delta| \pi(\theta|x) d\theta \\ &= \int_{\theta_1}^{\delta} (\theta - \delta) \pi(\theta|x) d\theta + \int_{\delta}^{\theta_2} (\delta - \theta) \pi(\theta|x) d\theta \end{aligned}$$

Nous cherchons à minimiser $f(\delta, x)$, donc nous résolvons

$$\frac{d}{d\delta} f(\delta, x) = 0$$

ça implique que

$$\int_{\theta_1}^{\delta} \pi(\theta|x) d\theta = \int_{\delta}^{\theta_2} \pi(\theta|x) d\theta$$

δ est bien entendue la mediane de la distribution a posteriori.

Définition 1.3.6. La perte $0-1$

Cette fonction de perte est utilisée dans le contexte des tests d'hypothèses et est un exemple typique d'une perte non quantitative, la pénalité associée à un estimateur δ est 1 si la réponse est correcte et 0 sinon.

Comme nous le savons, un test est la donnée d'une partition de Θ en $\Theta_0 \cup \Theta_1$. $\theta \in \Theta_i$ correspond à l'hypothèse H_i, H_0 est l'hypothèse nulle. Le principe décisionnel d'un test est le suivant :

$$\delta = \begin{cases} 1 & \text{si} \quad \theta \in \Theta_0 \\ 0 & \text{si} \quad \theta \in \Theta_1, \end{cases}$$

La fonction de perte correspondante est

$$L(\theta, \delta) = \mathbb{1}_{\theta \in \Theta_0} \times \mathbb{1}_{\delta=1} + \mathbb{1}_{\theta \in \Theta_1} \times \mathbb{1}_{\delta=0}$$

Le risque a posteriori est alors le suivant :

$$\rho(\pi, \delta|x) = \mathbb{1}_{\delta=1} P^{\pi}(\Theta_0|x) + \mathbb{1}_{\delta=0} P^{\pi}(\Theta_1|x)$$

ainsi l'estimateur de Bayes associé est

$$\delta^{\pi}(x) = \begin{cases} 1 & si \quad P^{\pi}(\Theta_0|x) < P^{\pi}(\Theta_1|x) \\ 0 & \text{sinon}, \end{cases}$$

c'est-à-dire que l'estimateur permet d'accepter H_0 si c'est l'hypothèse la plus probable a posteriori.

Mais il faut bien noter que le fait d'utiliser de telles fonctions de perte n'évite pas le recours à des approximations numériques surtout lorsque Θ est de grande dimension.

L'admissibilité

Définition 1.3.7. Une règle de décision δ_1 est dite meilleure que δ_2 si son risque associé est moins que celui associé à δ_2, c'est-à-dire si

$$\begin{cases} R(\delta_1, \theta) \leq R(\delta_2, \theta), & \forall \theta \in \Theta\,; \\ R(\delta_1, \theta) < R(\delta_2, \theta), & \text{pour au moins une valeur de } \theta. \end{cases}$$

Une décision δ est la meilleure de toutes les décisions si et seulement si sa fonction de risque est la plus petite.

Définition 1.3.8. Estimateur admissible

On dit que $\delta \in \mathcal{D}$ est inadmissible si et seulement si :

$$\exists \delta_0 \in \mathcal{D}, \forall \theta \in \Theta : R(\theta, \delta) \geq R(\theta, \delta_0) \text{ et } \exists \theta_0 \in \Theta : R(\theta_0, \delta) > R(\theta_0, \delta_0).$$

de ce fait, δ est dite admissible si elle n'est pas inadmissible et par conséquence, un estimateur est dit admissible si et seulement s'il n'est pas inadmissible.

Théorème 1.3.1 ([29]). *Estimateur Bayésien admissible*

Si l'estimateur Bayésien δ^π associé à une fonction de perte l et une loi a priori π est unique, alors il est admissible.

Démonstration. Supposons que δ^π est non admissible : $\exists \delta_0 \in \mathcal{D}, \forall \theta \in \Theta : R(\theta, \delta^\pi) \geq R(\theta, \delta_0)$ et $\exists \theta_0 \in \Theta : R(\theta_0, \delta^\pi) > R(\theta_0, \delta_0)$.

en intégrant la première inégalité, on obtient :

$$\int_\Theta R(\theta, \delta_0) \geq \int_\Theta R(\theta, \delta^\pi) = r(\pi)$$

donc δ_0 est aussi un estimateur Bayésien associé à l et π mais $\delta_0 \neq \delta^\pi$ d'aprés la seconde inégalité. Le théorème se déduit par contraposée.

L'unicité de l'estimateur Bayésien implique la finitude du risque $r(\pi) < \infty$

Définition 1.3.9. La π-admissibilité

Un estimateur δ_0 est π-admissible si et seulement si :

$$\forall (\theta, \delta_0), R(\theta, \delta) \geq R(\theta, \delta_0) \Rightarrow \pi\big\{\theta \in \Theta : R(\theta, \delta) < R(\theta, \delta_0)\big\} = 0$$

Autrement dit, cette définition implique que chaque estimateur non admissible est π-admissible

Proposition 1.3.1 ([29]). *Tout estimateur Bayésien tel que $r(\pi) < \infty$ est π-admissible.*

Nous pouvons maintenant énoncer une condition suffisante d'admissibilité des estimateurs Bayésiens.

Théorème 1.3.2 ([29]). *Si $\pi > 0$ sur Θ, $r(\pi) < \infty$ pour une fonction de perte l donnée, si δ^π l'estimateur Bayésien correspondant existe et si $\theta \longmapsto R(\theta, \delta)$ est continue. Alors δ^π est admissible.*

Démonstration. Supposons que δ^π est non admissible. D'aprés la définition précédente, δ^π est π-admissible. Ainsi, il existe δ_0 tel que $R(\theta, \delta_0) \geq R(\theta, \delta^\pi)$ et $\theta_0 \in \Theta : R(\theta_0, \delta_0) < R(\theta_0, \delta^\pi)$.
De tant que la fonction $\theta \longmapsto R(\theta, \delta)$ est continue, la fonction définie par $\theta \longmapsto R(\theta, \delta_0) - R(\theta, \delta^\pi)$ est aussi continue. Donc il existe un voisinage ouvert de θ_0, $\nu_0 \subset \Theta$ tel que $\forall \theta \in \nu_0, R(\theta, \delta_0) < R(\theta, \delta^\pi)$. En considérant $A = \left\{\theta \in \Theta : R(\theta, \delta) < R(\theta, \delta_0)\right\}$, il en résulte que $\pi(A) > \pi(\nu_0)$.(car π est supposée strictement positive et $\nu_0 \subset A$). Donc en prenant un modèle dominné par une mesure qui charge positivement les ouverts (la mesure de Lebesgue par exemple), $\pi(\nu_0) > 0$. A est donc non négligeable (de mesure non nulle), ce qui n'est pas conforme avec la π-admissibilité. En conclusion, δ^π est admissible.

L'estimateur MAP

On appelle estimateur MAP (estimateur de maximum a posteriori) tout estimateur $\delta^\pi(x)$ qui maximise l'information sur θ représentée par sa loi a posteriori, c'est-à-dire tout estimateur $\delta^\pi(x)$ tel que $\delta^\pi(x) \in Arg\max_\theta \pi(\theta|x)$. $\delta^\pi(x)$ doît donc être le mode de la distribution a posteriori.

Le grand avantage de cet estimateur est qu'il ne dépend pas d'une fonction de perte, et est utile pour les approches théoriques, sauf dans des procédures des tests où l'estimateur MAP s'intèrprète comme l'estimateur associé à la fonction de perte $0-1$ comme nous allons le constater dans la prochaîne section.

L'estimateur MAP est le pendant Bayésien de l'estimateur de maximum de vraisemblance, de ce fait ils partagent les mêmes inconvénients comme : la non unicité, l'instabilité (dus aux calculs d'optimisation) et la dépendance vis-à-vis de la mesure de référence (dominant Θ), seulement l'estimateur MAP ne vérifie pas la non invariance par reparamétrisation qui peut apparaître importante intuitivement.

1.3.3 Tests et intervalles de crédibilité

D'un point de vue statistique, un test soit au sens Bayésien ou au sens classique peut être considéré comme une des deux approches suivantes. Soit comme un procédé statistique, c'est-à-dire une fonction définie sur l'espace des observations à valeurs dans un espace à deux points que l'on arbitrairement appelle "accepter" et "rejeter" une hypothèse. Dans ce cas, on peut envisager un problème de test comme un problème de décision avec deux actions possibles. Sinon, il peut être considéré comme une façon pour le statisticien de gérer ses doutes relatifs à son modèle statistique .

Comme dans le cas de l'estimation, un test Bayésien se fait aprés avoir calculée la loi a posteriori.

Après avoir définies les régions de confiance au sens Bayésien, nous présenterons en quelques idées la notion des tests Bayésiens.

Intervalles de crédibilité

Définition 1.3.10. Région α-crédible

Pour $0 < \alpha < 1$, une région α-crédible de $100(1-\alpha)\%$ pour θ est un sous-ensemble $C \subset \Theta$ tel que

$$P^{\pi}\{\theta \in C | X = x\} = 1 - \alpha$$

Habituellement C est un intervalle.

Dans le cas où θ est une variable aléatoire continue, il suffit de déterminer deux quantiles $\theta^{(1)}$ et $\theta^{(2)}$ de $100\alpha_1\%$ et $100(1-\alpha_2)\%$ respectivement avec $\alpha_1 + \alpha_2 = \alpha$, de poser $C = [\theta^{(1)}, \theta^{(2)}]$, et de prendre par la suite $\alpha_1 = \alpha_2 = \alpha/2$ pour avoir deux régions α-crédibles à queues égales.

Mais ce n'est plus le cas si θ est discrète, il sera délicat de trouver une région C contenant θ avec une probabilité a posteriori égal exactement à $1-\alpha$. Pour cela, la condition a été généralisée à être

$$P^{\pi}\{\theta \in C | X = x\} > 1 - \alpha$$

et elle est utilisée même dans le cas continu s'il est difficile d'atteindre la probabilité a posteriori égal exactement à $1-\alpha$.

Dans la suite de ce travail, quand nous dirons une région α-crédible c'est la région $C \subset \Theta$ qui a une probabilité a posteriori supèrieur à $1-\alpha$ de contenir θ. i.e, une région C telle que $P^{\pi}\{\theta \in C | X = x\} > 1 - \alpha$, dans les deux cas continu et discrét.

Notons que le paradigme Bayésien permet une autre fois de s'affranchir d'un inconvénient de l'approche fréquentiste. En effet, au sens fréquentiste une région de confiance C est définie par $\forall\theta, P_\theta(\theta \in C) \geq 1-\alpha$. Il n'existe pas un certain x donné tel que la probabilité que x soit dans C est plus grande que $1-\alpha$, une région de confiance C n'a donc de sens que pour un trés grand nombre d'expériences.

En revanche, La définition Bayésienne exprime que la probabilité que θ soit dans C au vue de celles déjà réalisées est plus grande que $1-\alpha$. Il n'y a donc pas besoin d'avoir recours à un nombre infini d'expériences pour définir une région α-crédible.

Il existe une infinité des régions α-crédibles, il est logique de s'intéresser donc à celle qui a un volume minimal. Pour cela, nous avons besoin d'introduire la notion d'une région HPD (Highest Posterior Density).

Définition 1.3.11. Région HPD

Une région HPD est la région C_α^π définie par

$$C_\alpha^\pi = \{\theta, \pi(\theta|X=x) \geq h_\alpha\}$$

où h_α est donnée par : $h_\alpha = \sup\big\{h; P^\pi\big(\{\theta, \pi(\theta|x) \geq h\}\,|x\big) \geq 1-\alpha\big\}$.

C_α^π est parmi les régions α-crédibles sur lesquelles la densité a posteriori reste au dessus de la valeur la plus élevée possible (c'est-à-dire au dessus son mode).

Les régions HPD peuvent être calulées numériquement (par les calculs), ou approximativement comme elles peuvent être calculées par des méthodes de simulation comme les méthodes MCMC qui seront dévéloppées plus loin.

Les tests

Le facteur de Bayes

Supposons que nous avons deux hypothèses

$$\begin{aligned} H_0 &: \theta \in \Theta_0 \\ H_1 &: \theta \in \Theta_1 \end{aligned}$$

et nous devons choisir une parmi les deux dans un concept Bayésien. Pour le faire, il suffit de comparer les probabilités a posteriori des deux hypothèses

$$\begin{aligned} p_0 &= P(H_0|x) = P(\theta \in \Theta_0|x) \\ p_1 &= P(H_1|x) = P(\theta \in \Theta_1|x); \end{aligned}$$

la règle de Bayes consiste à choisir l'hypothèse de plus grande probabilité a posteriori. Supposons aussi que nous disposons des probabilités a priori des hypothèses

$$\begin{aligned} \pi_0 &= P(\theta \in \Theta_0) = P(H_0) \\ \pi_1 &= P(\theta \in \Theta_1) = P(H_1) = 1 - \pi_0 \end{aligned}$$

L'odds a priori (the odds prior) est le rapport des probabilités a priori de H_0 relativement à H_1

$$\frac{\pi_0}{\pi_1}$$

Ce rapport égal à 1 signifie que les hypothèses sont les mêmes avant d'observer les données. De même nous pouvons définir l'odds a posteriori (the odds posterior) comme

$$\frac{p_0}{p_1}$$

et le facteur de Bayes en faveur de H_0 relativement à H_1 est le rapport des deux odds

$$B_F = \frac{\textit{odds a posteriori}}{\textit{odds a priori}} = \frac{p_0\ \pi_1}{\pi_1\ \pi_0} \tag{1.16}$$

Pour mieux comprendre , laissons voir comment calculer les probabilités π_i, p_i et B_F en appliquant la formule de Bayes donnée dans la première section du chapitre.
Soit $g_i(\theta)$ la fonction de densité a priori de θ sous Θ_i

$$\int_{\Theta_i} g_i(\theta)\ d\theta = 1$$

ainsi, la probabilité a priori de θ sous $\Theta = \Theta_0 \cup \Theta_1$ est donnée par

$$\pi(\theta) = \pi_0\ g_0(\theta)\ I\{\theta \in \Theta_0\}\ +\ \pi_1\ g_1(\theta)\ I\{\theta \in \Theta_1\}$$

nous pouvons maintenant produire les probabilités a posteriori p_0 et p_1, notons que la densité marginale de x sous $\pi(\theta)$ peut être exprimée par

$$\begin{aligned} m_\pi(x) &= \int_{\Theta} f(x|\theta)\pi(\theta)\ d\theta \\ &= \pi_0 \int_{\Theta_0} f(x|\theta) g_0(\theta)\ d\theta + (1-\pi_0) \int_{\Theta_1} f(x|\theta) g_1(\theta)\ d\theta \\ &= \pi_0 P(X|H_0) + (1-\pi_0) P(X|H_1) \end{aligned}$$

d'où la distribution a posteriori de θ sachant $X = x$ est donnée selon la formule de Bayes par

$$\begin{aligned}\pi(\theta|x) &= \frac{f(x|\theta)\ \pi(\theta)}{m_\pi(x)} \\ &= \begin{cases} \frac{\pi_0\ f(x|\theta)\ g_0(\theta)}{m_\pi(x)} & si \quad \theta \in \Theta_0 \\ \frac{(1-\pi_0)\ f(x|\theta)\ g_1(\theta)}{m_\pi(x)} & si \quad \theta \in \Theta_1. \end{cases}\end{aligned}$$

et on en déduit que :

$$\begin{aligned}P(H_0|X) &= P(\theta \in \Theta_0|X) \\ &= \frac{\pi_0 \int\limits_{\Theta_0} f(x|\theta)\ g_0(\theta)\ d\theta}{m_\pi(x)} \\ &= \frac{\pi_0 \int\limits_{\Theta_0} f(x|\theta)\ g_0(\theta)\ d\theta}{\pi_0 \int\limits_{\Theta_0} f(x|\theta)\ g_0(\theta)\ d\theta\ +\ (1-\pi_0) \int\limits_{\Theta_1} f(x|\theta)\ g_1(\theta)\ d\theta} \\ &= \frac{\pi_0\ P(X|H_0)}{\pi_0\ P(X|H_0)\ +\ (1-\pi_0)\ P(X|H_1)}\end{aligned}$$

$$\begin{aligned}P(H_1|X) &= P(\theta \in \Theta_1|X) \\ &= \frac{(1-\pi_1) \int\limits_{\Theta_1} f(x|\theta)\ g_1(\theta)\ d\theta}{m_\pi(x)} \\ &= \frac{(1-\pi_1) \int\limits_{\Theta_1} f(x|\theta)\ g_1(\theta)\ d(\theta}{\pi_0 \int\limits_{\Theta_0} f(x|\theta) g_0(\theta)\ d\theta\ +\ (1-\pi_0) \int\limits_{\Theta_1} f(x|\theta)\ g_1(\theta)\ d\theta} \\ &= \frac{(1-\pi_0)\ P(X|H_1)}{\pi_0\ P(X|H_0)\ +\ (1-\pi_0)\ P(X|H_1)}\end{aligned}$$

en rempalcant ces probabilités dans la formule (1.16), le facteur de Bayes est définit comme suit

$$\begin{aligned}B_F &= \frac{P(X|H_0)}{P(X|H_1)} \\ &= \frac{\int\limits_{\Theta_0} f(x|\theta)\ g_0(\theta)\ d\theta}{\int\limits_{\Theta_1} f(x|\theta)\ g_1(\theta)\ d\theta}\end{aligned}$$

Ce rapport est une analogie Bayésienne du rapport de vraisemblances des tests classiques, il évalue la modification de la vraisemblance de l'ensemble Θ_0 par rapport à celle

de l'ensemble Θ_1.

Les facteurs de Bayes sont très flexibles pour la comparaison des hypothèses multiples et des modèles.
D'aprés la formule (1.16), nous pouvons écrire

$$\begin{aligned} [odds\ a\ posteriori] &= [odds\ a\ priori] \times [facteurdeBayes] \\ \frac{p_0}{p_1} &= \frac{\pi_0}{1-\pi_0} \times B_F \end{aligned}$$

généralement, on prend $\pi_0 = 1$ d'où le odds a priori vaut 1 et travailler avec le facteur de Bayes sera équivalent à travailler avec le odds a posteriori.

Comme nous l'avons déjà mentionné, le facteur de Bayes peut être utilisé pour comparer deux modèles

$$\begin{aligned} M_0 &= \{f_0(x|\theta_0),\ g_0(\theta_0)\} \\ M_1 &= \{f_1(x|\theta_1),\ g_1(\theta_1)\} \end{aligned}$$

le facteur de Bayes dans ce cas est défini comme suit :

$$\begin{aligned} B_F &= \frac{P(X|M_0)}{P(X|M_1)} \\ &= \frac{\int\limits_{\Theta_0} f_0(x|\theta_0)\ g_0(\theta_0)\ d\theta_0}{\int\limits_{\Theta_1} f_1(x|\theta_1)\ g_1(\theta_1)\ d\theta_1} \end{aligned}$$

et le odds a posteriori est défini de la même façon précédente

$$\frac{P(M_0|X)}{P(M_1|X)} = \frac{P(M_0)}{P(M_1)}\ B_F$$

Dans la pratique, on prend souvent

$$P(M_0) = P(M_1) = 1/2$$

c'est-à-dire que le odds a posterior sera égal au facteur de Bayes.

Interprétation du facteur de Bayes

Comme la règle de Bayes consiste à choisir l'hypothèse ou le modèle de plus grande probabilité a posteriori, le facteur de Bayes peut être interprété comme suit

$B_F \geq 1$	nous favorisons H_o
$10^{-1/2} \leq B_F < 1$	la certitude que H_0 est fausse est minimale
$10^{-1} \leq B_F < 10^{-1/2}$	cette certitude est substantielle
$10^{-2} \leq B_F < 10_{-1}$	la certitude est forte
$B_F < 10^{-2}$	la certitude est décisive et nous devons rejeter H_0

On constate ,ici aussi, un autre point de différence entre l'approche classique et l'approche Bayésienne. Un test Bayésien ne consiste pas à accepter une hypothèse et rejeter l'autre, mais bien à comparer les deux hypothèses et prendre la meilleure.

Exemple 4. Considérons un test pour controler le taux du sucre dans le sang d'une personne diabétique deux heures après son petit déjeuner. Il est d'intérêt de savoir si les médicaments qu'elle a pris ont controlé ce taux du sucre.

Assumons que le résultat de ce test X suit une loi normale avec une moyenne θ et une variance égale à 100, c'est-à-dire $X \sim \mathcal{N}(\theta, 100)$ où θ représente la vraie valeur de ce taux qui est inconnue. Dans la population (diabétiques qui suivent ce traitement) le paramètre θ est distribué selon une loi normale de $moyenne = 100$ et $variance = 900$, i.e. $\pi(\theta) = \mathcal{N}(100, 900)$. La distribution marginale de X est donc une loi normale $\mathcal{N}(100, 1000)$. Ainsi, la loi a posteriori de θ est encore normale avec
une $moyenne = \frac{900}{100}\, x + \frac{100}{1000}100 = 0.9\, x + 10$ et une $variance = \frac{100\times 900}{1000} = 90$.
donc $\pi(\theta|x) = \mathcal{N}(0.9\, x + 10,\ 90)$.

Supposons qu'on veut tester $H_0 : \theta \leq 130$ contre $H_1 : \theta > 130$.
Si le test du taux du sucre dans le sang du cette personne donne un résultat $x = 130$, que peut on conclure ? notons que dans ce cas la loi a posteriori de θ devient $\pi(\theta|x) = \mathcal{N}(127, 90)$.
alors, on obtient :

$$\begin{aligned} P(\theta \leq 130|X = 130) &= \Phi\left(\frac{130 - 127}{\sqrt{90}}\right) = \Phi(0.316) = 0.624 \\ P(\theta > 130|X = 130) &= 0.376 \end{aligned}$$

Le posterior odds donne dans ce cas :

$$\frac{P_0}{P_1} = \frac{0.624}{0.376} = 1.66$$

et comme $\pi_0 = P^{\pi}(\theta \leq 130) = \Phi(\frac{130-100}{30}) = \Phi(1)$ le prior odds est donc égale à
$\frac{\pi_0}{\pi_1} = \frac{\Phi(1)}{1-\Phi(1)} = \frac{0.8413}{0.1587} = 5.3$. ainsi, le facteur de Bayes est $B_F = \frac{1.66}{5.3} = 0.313$.
D'après cet exemple, on remarque que lorsque la loi a priori de notre paramètre est continue, on peut ne pas l'exprimer sous la forme $\pi(\theta) = \pi_0\ g_0(\theta)\ I\{\theta \in \Theta_0\} + \pi_1\ g_1(\theta)\ I\{\theta \in \Theta_1\}$ quand on fait les calculs.

1.4 Méthodes de Monte Carlo par chaînes de Markov

Cette section présente un aspect très important qui a offert une nouvelle vie à la statistique Bayésienne, et lui en a fait un moyen inévitable pour résoudre les problèmes des calculs pour divers modèles et dans des divers domaines ; les algorithmes de simulation MCMC. Notre objectif est de comprendre le mécanisme de fonctionnement de ces méthodes afin de les maîtriser et pouvoir les appliquer plus tard. Notons que ces algorithmes ne peuvent pas s'appliquer sans ordinateur. Le language de programmation R est le mieux placé et est le plus performant pour les statisticiens.

Une méthode MCMC est une technique de simulation qui consiste à générer un échantillon afin de mettre en place des chaînes de Markov avec des distributions érgodiques. Deux algorithmes MCMC sont les plus importants conçus pour créer des chaînes de Markov de loi stationnaire donnée : Le premier a été proposé par Metropolis et al.(1953) et Hastings (1970), où la prochaine valeur de la chaîne de Markov est générée à partir d'une loi dite loi de proposition (the jumping distribution or the proposal density), qui génère un nouveau candidat et sur la base de la probabilité d'acceptation, nous acceptons ou rejetons ce candidat. La deuxième méthode est l'échantillonneur de Gibbs introduite par Geman et Geman (1984), et qui a été developée plus tard par Tanner et Wang (1987), et ensuite par Gelfand et Smith (1990) dans laquelle la prochaîne valeur est générée en utilisant les nouvelles et dernières valeurs simulées.
Les méthodes MCMC fondent les algorithmes de calcul les plus efficaces, leur avantage sur les méthodes classiques de Monte carlo est qu'elles ne demandent pas de connaître la constante de normalisation, ce qui est en pratique le cas de la distribution a posteriori de la statistique Bayésienne. Au sens Bayesien toute inférence peut être déduite à partir de la distribution a posteriori par des rapports de synthèse appropriés. Cela prend généralement la forme de l'évaluation des intégrales

$$J \;=\; \int f(\theta)\;\pi(\theta|x)\;dx \tag{1.17}$$

d'une fonction $f(\theta)$ á l'égard de la distribution a posteriori. Par example les estimations ponctuelles d'un paramètre inconnu sont données par les moyennes a posteriori, i.e : $f(\theta) = \theta$, la préduction d'une valeur future $\widetilde{x}$ est fondée sur la distribution prédictive a posteriori

$$f(\widetilde{x}|x) \;=\; \int f(\widetilde{x}|\theta,x)\;\pi(\theta|x)\;d\theta \tag{1.18}$$

c'est-à-dire $f(\theta) = f(\widetilde{x}|\theta,x)$.

Le problème est que ces intégrales sont généralement difficilles à évaluer d'une manière analytique, et lorsque le paramètre est multidimentionnel, même les méthodes numériques peuvent échouer. Les méthodes MCMC sont les mieux placées pour évaluer de telles

intégrales .

Cette section est organisée comme suit : Dans la deuxième partie, nous allons présenter quelques notions sur les chaînes de Markov, afin de comprendre leurs intérêt dans lea méthodes MCMC. Ensuite, nous présenterons les deux téchniques de simulation les plus utilisées par les statisticiens, l'algorithme de Metropolis-Hastings et l'échantillonneur de Gibbs en donnant un exemple pour chaque méthode.

1.4.1 les chaînes de Markov

La propriété des chaînes de Markov que nous allons utiliser est que certaines d'entre elles convergent vers une unique et invariante distribution, permettant ainsi d'estimer les espérances. La théorie des chaînes de Markov est complexe et nous n'allons ici ne donner que les bases nécessaires à notre méthodes.

Définition 1.4.1. Une chaîne de Markov est une collection de variables aléatoires $(X_i)_{i \in N}$. L'évolution de cette chaîne sur un espace Ω est régie par le noyau de transition qui est un mécanisme décrivant le mouvement de la rpobabilité d'un état à un autre basant sur l'état actuel, et qui correspond à la distribution conditionelle de X_{i+1} sachant tout le passé,

$$P(x, A) \quad = \quad P(X_{i+1} \in A | X_i = x, X_j, j < i), x \in \Omega, A \subset \Omega \tag{1.19}$$

Pour que cette collection de variables aléatoires soit une chaîne de Markov, elle doit vérifier la proprieté d'absence de mémoire

$$P(X_{i+1} \in A | X_i = x, X_j, j < i) \quad = \quad P(X_{i+1} \in A | X_i = x) \tag{1.20}$$

L'hypothèse que la distribution de probabilité de l'élément suivant dans la séquence donnant le présent et tout le passé ne dépend que de l'état actuel est appellée aussi propriété de Markov.

Les définitions suivantes précisent les propriétés nécessaires pour la convergence des chaînes de Markov produites par les algorithmes MCMC de la section suivante.

Définition 1.4.2. Chaîne irréductible

Une chaîne de Markov est dite irréductible si tous les états communiquent entre eux, c'est-à-dire : si $\forall \theta, \theta' \in \Theta$ il y a une probabilité non nulle que partant de θ, on aboutisse à θ' en un nombre fini d'étapes. En terme de classe d'équivalance, une chaîne est irréductible s'il n'y a qu'une seule classe d'équivqlance.

Définition 1.4.3. Chaîne récurrente

Une chaîne de Markov irréductible est récurrente si l'espérance du nombre de visites qu'elle accorde à chaque état est infini

$$\forall(\theta,\theta'),\ E(\theta \longrightarrow \theta') = \sum_{r=1}^{\infty} \pi^r(\theta,\theta') = \infty$$

- Dans le cas où l'espace d'états est fini toute chaîne irréductible est récurrente, en effet, le nombre d'états étant fini, il existe donc au moins un état qui est visité infiniment souvent qu'on itère la chaîne à l'infini. Cet état étant connecté à tous les autres, chacun des états étant visité infiniment souvent. La question de la récurrence de la chaîne ne se pose donc réellement que lorsque l'espace d'états est infini.
- Une chaîne récurrente est positive si la fréquence de visites de tout sous- ensemble A à partir d'un état de départ θ est strictement positive, et nous avons alors un condidat a une distribution invariante de la chaîne.
- Une chaîne irréductible récurrente est positive si le temps de retour est fini pour chaque couple d'états

$$T_{\theta \longrightarrow \theta'} = E(T_{\theta \longrightarrow \theta'}) < \infty$$

Définition 1.4.4. Chaîne apériodique

Une chaîne de Markov est dite apériodique si elle est irréductible et tous les états sont de période 1.
On appelle une période T d'un état θ appartenant à une chaîne discète et on note $d(\theta)$, le plus grand commun diviseur des valeurs de $r \geq 1$ telles que les probabilités de transition $\pi^r(\theta,\theta)$ en r étape sont positives.

$$d(\theta) = PGCD\left\{r \in N^*, \pi^r(\theta,\theta) > 0\right\}$$

Lorsque la chaîne de Markov vérifie toutes ces propriétés, c'est-à-dire qu'elle est appériodique, irréductible et récurrente positive, elle sera dite chaîne ergodique.

Nous pouvons maintenant introduire le théorème fondamental de l'utilisation des chînes de Markov dans les méthodes de Monte Carlo.

Théorème 1.4.1 ([28]). ***(Théorème ergodique)***

Soient $\theta^{(1)}$,$\theta^{(2)}$,..., $\theta^{(T)}$ T valeurs d'une chaîne de Markov ergodique de distribution invariante (stationnaire), et tel que $E^{\pi}[g(\theta)] < \infty$. Avec une probabilité égale à 1,

$$\frac{1}{T}\sum_{t=1}^{T} g(\theta^{(t)}) \quad \underset{T \mapsto \infty}{\longrightarrow} \quad \int_{\Theta} g(\theta)\pi(\theta|x)d\theta = E^{\pi}[g(\theta)] \tag{1.21}$$

où π est la distribution stationnaire.

Le théorème ergodique répend aux problèmes de convergence rencontrés dans la simulation par chaînes de Markov, car il étend la loi des grands nombres à des suites dépendantes de variables aléatoires et supprime le besoin de construire un échantillon i.i.d.

1.4.2 Chaînes de Markov et méthodes de Monte Carlo

La méthode MCMC tire son nom de l'idée que, pour produire des approximations acceptables d'intégrales et d'autres fonctions dépendant d'une loi d'intérêt (La loi a posteriori dans notre cas), il suffit de générer une chaîne de Markov $(\theta^{(t)})_t$ de loi limite la loi d'intérêt.

Les chaînes de Markov $(\theta^t)_t$ produites par les algorithmes MCMC sont bénéficies par construction de propriétés de stabilité forte, à savoir l'existence d'une distribution stationnaire ou invariante, soit une distribution π telle que, si $x_n \sim \pi$, $x_{n+1} \sim \pi$. Cette propriété signifie dans la dynamique des chaînes de Markov que lorsqu'on injecte un point de départ θ tiré au hasard selon la densité de probabilité π, on retrouve généré par le noyau un point de sortie θ' qui suit lui même cette même loi de probabilité π. Ces chaînes sont aussi irréductibles. Grâce à cette stabilité, ces chaînes sont récurrentes de loi stationnaire $\pi(\theta|x)$ c'est-à-dire que le nombre moyen de visites dans un ensemble arbitraire A de mesure positive est infini, ou même Harris récurrentes, c'est-à-dire telle que la probabilité d'un nombre infini de visites dans A est 1, ce qui assure que la chaîne possède les mêmes propriétés limites quelle que soit la valeur initiale $\theta^{(0)}$ (cette propriété correspond à l'érgodicité de la chaîne). La récurrence au sens de Harris est donc nécessaire pour garantir la convergence à partir de tout point de départ.

Par conséquent, pour un nombre de simulation, k, suffisamment grand, le $\theta^{(k)}$ résultant est distribué approximativement selon la loi $\pi(\theta|x)$, quelle que soit la valeur initiale $\theta^{(0)}$, dans la pratique, le problème est de déterminer que signifie un grand k ?

le taux de décroissance de la différence entre la loi de $\theta^{(k)}$ et sa limite (la vitesse de convergence) peut apporter une réponse à ce problème. ce taux de convergence dépend souvent du point de départ (sauf si la chaîne est uniformement ergodique) et un nombre d'itération k donné ne fournit pas la même qualité d'approximation pour différentes valeurs de $\theta^{(0)}$. C'est pour cela qu'il est préférable de prendre la valeur actuelle comme nouvelle valeur initiale même si cela introduit de la dépendance entre les valeurs car nous nous intéressons à des fonctionnelles de $\pi(\theta|x)$, de plus, le théorème ergodique ne necessite pas l'indépendance et garantit la convergence.

Une fois $\theta_1 = \theta^{(k)}$ généré, une façon naîve de construire un échantillon indépendant et identiquement distribué suivant $\pi(\theta|x)$ est d'utiliser le même algorithme avec une autre valeur initiale $\theta_2^{(0)}$ et une autre séquence de transition de Markov afin d'obtenir θ_2 et ainsi de suite.

Donc, MCMC est une classe de méthodes qui consiste à simuler des tirages dépendants à partir de notre distribution d'intérêt (la distribution a posteriori), et les utiliser pour calculer les quantités d'intérêt de la loi a posteriori.

La partie suivante aborde les deux algorithmes MCMC les plus utilisés par les Bayésiens.

Algorithme de Metropolis-Hastings

La technique de Metropolis-Hastings est historiquement la première des méthodes MCMC, elle a été développée par Metropolis et al. (1953), au départ pour la physique particulaire, et généralisée par Hastings (1970) dans un cadre plus statistique. Elle est fondée sur la construction d'une distribution de proposition J qui génère un candidat θ^*, et sur la base de la probabilité d'acceptation ; nous acceptons ou rejetons ce candidat mais conservant la valeur précédemment simulée en cas de rejet comme nous le verons prochaînement. L'algorithme de Metropolis-Hastings est une généralisation de l'algorithme d'acceptation-rejet, l'idée est parvenue de fait que, les réalisations successives de θ_t dans l'algorithme d'acceptation-rejet sont indépendantes ce qui implique que la décision de considérer une réalisation issue de la loi à simuler comme réalisation candidate de la loi d'intérêt ne peut être prise que sur les données présentes, ce qui nécissite une décision plus nuancée qui peut s'appuyer sur l'utilisation des informations antérieurs sur la chaîne $(\theta^{(t)})_t$ qui doit avoir alors une structure de mémoire, la structure la plus simple et la mieux placée que l'on puisse envisager pour répondre à ce problème est celle de chaînes de Markov.

Pour une distribution a posteriori donnée $\pi(\theta|x)$, la procédure itérative de Metropolis-Hastings, génère à partir d'une valeur θ_i, la valeur suivante θ_{i+1} sur la base d'un algorithme en deux temps :

1. D'abord, on choisit une valeur candidate θ^* tirée aléatoirement d'une distribution de probabilité $J(\theta^*|\theta_i)$ éventuellement dépendante de θ_i. Cette loi est dite loi de proposition, mais aussi appelée "fonction de saut" parce qu'elle permet à la chaîne de bouger dans Θ à partir d'un point donné.
2. Ensuite le candidat θ^* est accepté avec une probabilité :

$$\alpha(\theta_i, \theta^*) = \min\left(1, \frac{\pi(\theta^*|x)\ J(\theta^*|\theta^{(i)})}{\pi(\theta^{(i)}|x)\ J(\theta^{(i)}|\theta^*)}\right) \tag{1.22}$$

"Accepter" le candidat signifie le choisir comme valeur suivante de la chaîne : $\theta_{i+1} = \theta^*$. A contrario, si le candidat est refusé, alors la chaîne ne bouge pas de θ_i : $\theta_{i+1} = \theta_i$. En pratique, si le rapport dans la formule (1.22) est supèrieur à 1, on accepte le candidat. S'il est inférieur à 1, on tire une valeur u d'une loi uniforme $\mathcal{U}(0,1)$ et on définit :

$$\begin{aligned} \theta_{i+1} &= \theta^* \ si\ u \ \leq\ \alpha(\theta_i, \theta^*) \\ \theta_{i+1} &= \theta_i \ si\ u \ >\ \alpha(\theta_i, \theta^*) \end{aligned}$$

Notons que $\theta^{(0)}$ doit avoir une probabilité positive

$$P(\theta^{(0)}|x) > 0$$

Si la densité de proposition est symétrique,i.e., $J(\theta, x) = J(x, \theta)$, la probabilité d'acceptation dans la formule (1.22) se réduit à $\pi(\theta^*|x)/\pi(\theta^{(i)}|x)$ qui est la formulation originale de Metropolis et al.(1953).

Lemme 1.4.1 ([28]). *Lorsque le support de $J(.|\theta)$ contient le support de $\pi(.|x)$,i.e., $supp\pi(.|x) \subset suppJ(.|\theta)$. La chaîne de Markov $(\theta^{(i)})_i$ produite par l'algorithme de Metropolis-Hastings est irréduvtible.*

L'irréductibilité de la chaîne découle de la condition sur le support de J, qui n'est cependant pas nécessaire pour assurer la validité de l'algorithme.

Théorème 1.4.2 ([29]). *Si la chaîne $(\theta^{(i)})_i$ est irréductible, c'est-à-dire si pour tout sous-ensemble A tel que $\pi(A) > 0$, il existe I tel que $P_{\theta^{(0)}}(\theta^{(I)} \in A) > 0$, 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*

$$\lim_{i \longrightarrow \infty} \sup_A |P_{\theta^{(0)}}(\theta^{(T)} \in A) - \pi(A) > 0| = 0$$

L'ergodicité de la chaîne est garantie par la condition d'équilibre ponctuel, c'est-à-dire le fait que le noyau de transition de la chaîne de Markov associé à l'algorithme ci-dessus, noté $K(\theta'|\theta)$, satisfasse

$$\pi(\theta|x)K(\theta'|\theta) = \pi(\theta'|x)K(\theta|\theta')$$

qui se vérifie aisément en écrivant le noyau de transition de l'algorithme de Metropolis-Hastings

$$K(\theta'|\theta) \quad = \quad \alpha(\theta, \theta')\ J(\theta'|\theta) + \left[1 - \int_\Theta \alpha(\theta, \theta^*)\ J(\theta^*|\theta)\ d\theta^*\right]\ \delta_\theta(\theta') \tag{1.23}$$

où δ est la masse de Dirac et, la probabilité de transition de θ à θ' est donnée par

$$P_{MH}(\theta, \theta') = \alpha(\theta, \theta')\ J(\theta'|\theta)$$

en effet,

$$\begin{aligned}
K(\theta'|\theta)\pi(\theta|x) &= \min\left(1, \frac{\pi(\theta'|x)\ J(\theta|\theta')}{\pi(\theta'|x)\ J(\theta'|\theta)}\pi(\theta|x)\ J(\theta'|\theta) + \int_\Theta [1-\alpha(\theta,\theta^*)\ J(\theta^*|\theta)]\ d\theta^*\ \delta_\theta(\theta')\ \pi(\theta|x)\right) \\
&= \min\left(\pi(\theta|x)\ J(\theta'|\theta), \pi(\theta'|x)\ J(\theta|\theta') + \int_\Theta [1-\alpha(\theta,\theta^*)\ J(\theta^*|\theta)]\ d\theta^*\ \pi(\theta|x)\ I_{\theta=\theta'}\right) \\
&= \min\left(\pi(\theta'|x), \frac{\pi(\theta|x)\ J(\theta'|\theta)}{J(\theta|\theta')}\ J(\theta|\theta') + \int_\Theta [1-\alpha(\theta,\theta^*)\ J(\theta^*|\theta)]\ d\theta^*\ \pi(\theta|x)\ I_{\theta=\theta'}\right) \\
&= \min\left(1, \frac{\pi(\theta|x)\ J(\theta'|\theta)}{\pi(\theta'|x)\ J(\theta|\theta')}\ J(\theta|\theta')\ pi(\theta'|x) + \int_\Theta [1-\alpha(\theta,\theta^*)\ J(\theta^*|\theta)]\ d\theta^*\ \pi(\theta'|x)\ I_{\theta=\theta'}\right) \\
&= K(\theta \mid \theta'\ \pi(\theta'|x))
\end{aligned}$$

Proposition 1.4.1 ([28]). *Si $\pi(\theta|x)$,$J(\theta,\theta')$ sont positives continues alors, l'algorithme de Metropolis-Hastings satisfait les conditions du théorème (1.4.1)*

D'autres conditions de convergence de la chaîne produite par l'algorithme de Metropolis-Hastings sont discutées par Smith et Robert(1993), Tierney (1994) et Tweeedie (1994).

L'importance de l'algorithme de Metropolis-Hastings, réside dans sa généralité. Sous conditions de régularité peu restrictives, par exemple la positivité du noyau de transition, la convergence de la chaîne vers la distribution visée $\pi(\theta|x)$ est assurée indépendamment du choix de la loi de proposition.

Cependant, même si la convergence théorique de la chaîne n'est pas mise en cause, en pratique, le choix des lois de proposition est un moment décisif dans la mise en place de la méthode. Des lois de proposition trop disperssées qui génèrent des candidats souvent refusés, peuvent laisser la chaîne bloquée sur une valeur et demander un nombre d'itérations pratiquement incompatible avec les exigences du modélisateur. Inversement, avec des lois de proposition peu disperssées, la chaîne peut bouger trop lentement. L'effet est le même : Trop d'itérations sont nécessaires pour atteindre la convergence de la chaîne.

Il convient de souligner qu'une fois la densité de proposition est déterminée, l'algorithme de Metropolis-Hastings est un moyen simple de simuler pratiquement n'importe quelle densité a posteriori.

Cette technique MCMC de type Metropolis-Hastings a été utilisée dans plusieurs cas réels. Elle a été utilisée par exemple pour reconstruire des images dégradées afin de pouvoir trouver l'image d'origine. L'idée est de proposer un ensemble d'images améliorées d'une image observée, afin de retenir l'image la plus proche de l'image originale.

Nous décrivons maintenant brièvement, un des aspects d'application de la technique de Metropolis-Hastings qui concerne la dégradation des compteur et qui a sucité beaucoup d'intérêt depuis les années 50.

Exemple 5 ([75]). La dégradation des compteurs

La dégradation des compteurs est décrite avec un modèle markovien. Normalement l'estimation de ces modèles est réalisée à partir des données sous la forme d'observations répétées d'un certain nombre d'individus. Dans le cas présent, puisque l'essai d'un compteur peut être considéré comme une mesure destructive de l'appareil, les techniques d'estimation usuelles ne sont pas utilisables et le problème est mathématiquement plus complexe. De ce fait, l'algorithme de Metropolis-Hastings a été utilisé pour mener les calculs d'inférence dans un cadre Bayésien.

Soit $X(t)$ une chaîne de Markov discrète et homogène, avec un nombre s d'états. Ce modèle est paramétré par une matrice carrée de transition θ :

$$\theta = \begin{pmatrix} \theta_{11} & \theta_{12} & ... & \theta_{1s} \\ \theta_{21} & \theta_{22} & ... & \theta_{2s} \\ ... & & & \\ \theta_{s1} & \theta_{s2} & ... & \theta_{ss} \end{pmatrix} \tag{1.24}$$

dont l'élément θ_{ij} est la probabilité que le système soit dans l'état j au temps t, sachant qu'il se trouvait dans l'état i au temps t-1 :

$$\theta_{ij} = [x(t) = j | x(t-1) = i] \tag{1.25}$$

ces paramètres sont indépendants de t, en vertu de l'hypothèse d'homogénéité, et vérifient la relation :

$$0 \leq \theta_{ij} \leq 1 \quad \sum_{j=1}^{s} \theta_{ij} = 1 \tag{1.26}$$

Si on dénote avec $[\underline{x}(t)]$ le vecteur ligne (à s composantes) des probabilités conditionnelles d'appartenance à chacun des s états :

$$[\underline{x}(t)] = \{[X(t) = 1], ..., [X(t) = s]\} \tag{1.27}$$

l'évolution du système est décrite par l'équation dynamique :

$$[\underline{x}(t)] = [\underline{x}(t-1)].\theta \tag{1.28}$$

qui peut s'écrire, en fonction des probabilités de l'état initial du système $[\underline{x}(0)]$:

$$[\underline{x}(t)] = [\underline{x}(0)].\theta^t \tag{1.29}$$

Dans de nombreux problèmes réels, les données de disposition ne sont pas des séries temporelles des états, mais l'individu k est observé une seule fois à un instant μ_k.

Pour fixer les idées on peut imaginer que les données sont issues de mesures destructives réalisées sur un certain nombre z' d'individus.

Individu	Instant de mesure	Etat observé
1	2	1
2	10	4
3	5	2
...		
z'	7	3

Ces données sont résumées de manière exaustive par les statistiques :

$$\{y_1(1), y_2(t), ..., y_s(t)\} \quad t = 1, 2, ..., t_{obs} \tag{1.30}$$

$y_i(t)$ étant le nombre d'individus observés à l'instant t et qui s'y trouvent dans l'état i :

$$y_i(t) = \sum_{k=1}^{z'} 1_{(\mu_k = t, x_k = i)} \tag{1.31}$$

et $t_{obs} = max\ (\mu_1, ..., \mu_{z'})$.

Dans la suite on dénotera $\underline{y}(t)$ le vecteur ligne des observations au temps t :

$$\underline{y}(t) = \{y_1(1), y_2(t), ..., y_s(t)\} \tag{1.32}$$

Les premières méthodes utilisées, dès les années 50, pour résoudre le problème de l'estimation de la matrice de transition θ dans ce cas "peu usuel" étaient basées sur le remplacement des probabilités conditionnelles $[\underline{x}(t)]$ par les proportions des individus observés de même âge, qui se présentent comme leurs estimateurs naturels :

$$[\widehat{\underline{x}}(t)] = \left\{ \frac{y_1(t)}{n(t)}, \frac{y_2(t)}{n(t)}, ..., \frac{y_s(t)}{n(t)} \right\} \tag{1.33}$$

où $n(t) = \sum_{j=1}^{s} y_i(t)$ est la taille de l'échantillon d'âge t. Il n'est pas superflu d'observer que le nombre d'observations $n(t)$ n'est pas forcement le même pour toutes les valeurs de t. On peut imaginer, par exemple d'étalonner 10 individus de 5 ans, aucun individu de 10 ans et 20 individus de 15 ans.

Le remplacement de $[\underline{x}(t)]$ par $[\underline{\widehat{x}}(t)]$ dans les équations (1.28) donne lieu à un système de t_{obs}. s équations en les s^2 inconnus θ_{ij}, qui, si $t_{obs} > s$, peut être résolu avec des techniques de moindres carrés (Miller, 1952), (Goodman, 1953), (Madansky, 1959). La difficulté principale dans la résolution de ce système est le respect des contraintes spécifiées par les équations (1.26). Cette méthode est sensible à la présence des données manquantes, parce que le manque d'observations pour un âge donné en un temps t^* entraîne une suppression de 2s équations (celles de t^* et $t^* - 1$).

Cette approche a été récement revisitée dans un cadre Bayésien par Congdon (2001).

Le problème a été reformulé successivement, dans les années 60, par Lee et al.(1968) qui ont proposé une technique d'estimation Maximum de vraisemblance. Ils ont obtenu un estimateur de θ en maximisant la vraisemblance donnée par :

$$[\underline{y}|\theta] = \prod_{t=0}^{t_{obs}} \frac{n(t)!}{\prod_{j=1}^{s} y_j(t)!} \prod_{j=1}^{s} [x_j(t)]^{y_j(t)} \tag{1.34}$$

sous les contraintes (1.26) avec des techniques de programation quadratiques.

Pour estimer ce modèle dans un cadre Bayésien, il faut choisir préalablement une loi a priori pour θ. Une distribution de probabilité adaptée pour θ est le produit de s distributions de Dirichlet indépendantes, une pour chaque ligne $\underline{\theta_i}$ de la matrice, chacune paramétrée par s constantes positives $\{\alpha_{i1}, ..., \alpha_{is}\}$:

$$\underline{\theta_i} \sim \mathcal{D}(\underline{\alpha_i}) \tag{1.35}$$

avec $\underline{\alpha_i} = \{\alpha_{i1}, ..., \alpha_{is}\}$.

Les distributions de Dirichlet, normalement utilisées en statistique Bayésienne (Good, 1965 ; Gelman et al., 1995) comme des lois a priori conjuguées des lois multinomiales sont des généralisations multidimensionnelles de lois Bêta et ont la propriété que leurs réalisations vérifient automatiquement les conditions (1.26).
La distribution a priori de θ s'écrit alors :

$$[\theta] = {}^{s}\prod_{i=1} \frac{\Gamma(\alpha_{i1}, ..., \alpha_{is})}{\Gamma(\alpha_{i1})\Gamma(\alpha_{is})} \, \theta_{i1}^{\alpha_{i1}-1} ..... \, \theta_{is}^{\alpha_{is}-1} \tag{1.36}$$

Lee et al. (1986) se servent de la méthode utilisée pour la maximisation de la vraisemblance pour obtenir un estimateur Bayésien ponctuel de θ en maximisant le produit de la vraisemblance (1.34) et la loi a priori (1.36). Cet estimateur puisque dans la formule de Bayes le dénominateur ne joue aucun rôle, maximise aussi la loi a posteriori de θ et représente le mode a posteriori.

L'expression exacte de la loi a posteriori est pratiquement incalculable. Et c'est à ce niveau là que la mise en place d'un algorithme MCMC de type Metropolis-Hastings est nécessaire. La mise en place d'un algorithme de Metropolis-Hastings nécessite la détermination d'une loi de proposition pour générer une matrice candidate θ^*.

Dans le cas présent, un choix commode est de tirer indépendament chaque ligne $\underline{\theta_i^*}$ de la matrice θ^* selon une loi de Dirichlit dont l'espérance est le vecteur ligne $\underline{\theta_i^{k-1}}$ de la matrice θ^{k-1}, où k est le nombre d'itérations.

$$\underline{\theta_i^*} \ \sim \ D(h_i, \underline{\theta_i^{k-1}})$$

La constante h_i peut être interprétée comme un paramètre de forme de la distribution de probabilité.

Dans la pratique technique, le blocage d'un compteur est détecté par les releveurs au moment de la lecture périodique de l'index. Si, par rapport à la dernière valeur connue, ce dernier n'a pas bougé. Alors le compteur est jugé bloqué et une demande de remplacement est immédiatement établie.

Les données

Le tableau (1.4) ci-dessous représente des données qui concernent des compteurs volumétriques domestiques (DN15 et 20 mm), d'âge compris entre 1 et 20 ans. On imagine que l'âge est le seul facteur explicatif de la dégradation.

Concernant les blocages, les données ICBC (de confiance aux informations de la base "Branchemenr et Compteurs") sont relatives à un échantillon significatif de l'effectif total de compteurs installés et on fait l'hypothèse que le taux de renseigement de l'information, au sein de la population examinée est de 100% : en pratique tous les blocages sont inclus dans les observations.

L'estimation des paramètres

Les distributions a priori des probabilités de transition θ_{ij} sont des distributions de Dirichlet dont les paramètres sont tous égaux à 1. Ce choix donne lieu à des lois uniformes pour les 3 vecteurs des probabilités de transition à partir des états 1, 2 et 3 :

$$\underline{\theta_1} = \{\theta_{11}, \theta_{12}, \theta_{13}, \theta_{14}\} \sim \mathcal{D}(1,1,1,1) \quad (1.37)$$
$$\underline{\theta_2} = \{\theta_{22}, \theta_{23}, \theta_{24}\} \sim \mathcal{D}(1,1,1) \quad (1.38)$$
$$\underline{\theta_3} = \{\theta_{33}, \theta_{34}\} \sim \mathcal{D}(1,1) \quad (1.39)$$

Le choix d'une loi uniforme traduit le manque de connaissance préliminaire sur les probabilités de transition : toutes les valeurs possibles sont également probables a priori.

TAB. 1.4 – Les données

	Données métrologiques			Données de facturation	
Âge	Y_1	Y_2	Y_3	Nombre de blocages	Taille de la pop.observée
1	43	5	2	136	63741
2	25	6	3	67	61983
3	78	16	7	82	41595
4	157	33	9	116	66475
5	237	50	7	149	82744
6	235	37	14	116	106898
7	243	72	11	64	45928
8	170	63	29	84	34452
9	174	61	21	95	33880
10	187	80	35	97	29209
11	179	94	27	134	32109
12	149	86	32	127	33132
13	150	125	51	176	26832
14	152	146	41	146	19203
15	157	137	50	394	64403
16	114	81	89	129	23155
17	80	65	66	85	10149
18	70	50	22	48	4676
19	79	62	25	107	10484
20	63	63	46	113	12010

TAB. 1.5 – Les résultats de la simulation

					Percentiles	
Param.	Moyenne	Mode	Mediane	Ec.type	2.5%	97.5%
θ_{11}	0.9483	0.9485	0.9484	0.0011	0.9461	0.9604
θ_{12}	0.0492	0.0489	0.0491	0.0012	0.0469	0.0515
θ_{13}	0.0011	0.0008	0.0010	0.0005	0.0003	0.0022
θ_{14}	0.0013	0.0014	0.0013	0.0001	0.0012	0.0015
θ_{22}	0.9389	0.9395	0.9389	0.0026	0.9337	0.9444
θ_{23}	0.0610	0.0603	0.0611	0.0026	0.0556	0.0662
θ_{24}	0.0001	0.0000	0.0001	0.0001	0.0000	0.0004
θ_{33}	0.9667	0.9677	0.9669	0.0017	0.9633	0.9698
θ_{34}	0.0332	0.0322	0.0331	0.0017	0.0302	0.0367

Pour obtenir des tirages aléatoires dans la loi a posteriori des θ_{ij}, on a réalisé avec l'algorithme de Metropolis-Hastings 10 000 itérations de 5 chaînes différentes, à partir de points initiaux trés dispersés. L'utilisation de plusieurs chaînes permet de vérifier la convergence vers la loi visée. Les dernières 5000 valeurs de chaque chaînes sont retenues pour simuler les lois a posteriori.

Le tableau (1.5) ci-dessus montre les principales caractéristiques de ces lois, obtenues empiriquement, à partir des valeurs retenues.

Les percentiles d'ordre 2.5% et 97.5% représentent les bornes de l'intervalle de crédibilité a posteriori de niveau 95% interprété par les Bayésiens comme l'intervalle où se trouvent 95% des valeurs possibles des paramètres.

L'échantillonnage de Gibbs

Nous allons à présent mettre en place une autre méthode MCMC qui permet elle aussi de générer des variables aléatoires suivant approximativement la loi a posteriori $\pi(\theta|x)$ dans laquelle, le vecteur aléatoire est partitioné en plusieurs blocs et la densité de transition est définie comme le produit des densités complètes. Cette méthode tire son nom des champs aléatoires de Gibbs où elle a été utilisée pour la première fois par Geman et Geman (1984). Elle repose sur une perspective différente que l'algorithme de Metropolis-Hastings de sorte qu'elle ne demande pas de mettre en place une fonction de proposition, mais de sa part, elle est fondée sur la loi cible (la loi a posteriori) $\pi(\theta|x)$ et elle est éssentiellement basée sur les distributions conditionnelles complètes. De plus, l'algorithme de Gibbs pour l´estimation et construction de modèles par conditionnement probabiliste donne souvent de meilleurs résultats .

Supposons d'abord que le vecteur θ ait deux coordonnées et supposons aussi que l'on connaisse les deux densités conditionnelles $[\theta_1|\theta_2]$ et $[\theta_2|\theta_1]$.
En donnant des valeurs initiales $\theta_1^{(t)}$ et $\theta_2^{(t)}$ à l'étape t, décrivons l'algorithme de simulation itératif de Gibbs à l'étape t+1

1. On génère $\theta_1^{(t+1)}$ en simulant selon la loi $[\theta_1|\theta_2^{(t)}]$
2. On génère $\theta_2^{(t+1)}$ en simulant selon la loi $[\theta_2|\theta_1^{(t+1)}]$

Remarquons que c´est la nouvelle et dernière valeur simulée $\theta_1^{(t+1)}$ qui est utilisée pour générer la seconde ccomposante $\theta_2^{(t+1)}$ associée à la même étape.
La suite des couples $\begin{pmatrix}\theta_1^{(t+1)}\\ \theta_2^{(t+1)}\end{pmatrix}$ ainsi générés est une chaîne de Markov et son noyau est

$$K\left(\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix}, \begin{pmatrix}\theta_1'\\ \theta_2'\end{pmatrix}\right) = [\theta_1'|\theta_2][\theta_2'|\theta_1'] \tag{1.40}$$

La formulation générale de cet algorithme pour un vecteur à k composantes peut être décrite comme suit : En supposant connu le vecteur $\theta^{(t)} = \theta_1^{(t)}, \theta_2^{(t)}, ..., \theta_j^{(t)}, \theta_k^{(t)}$ à l'etape t

1. On génère $\theta_1^{(t+1)}$ en simulant selon la loi $[\theta_1|\theta_2^{(t)}, ..., \theta_j^{(t)}, \theta_k^{(t)}]$
2. On génère $\theta_2^{(t+1)}$ en simulant selon la loi $[\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)}...., \theta_j^{(t)}, \theta_k^{(t)}]$
3. On génère $\theta_j^{(t+1)}$ en simulant selon la loi $[\theta_j|\theta_1^{(t+1)}, \theta_2^{(t+1)}, ..., \theta_{j-1}^{(t+1)}, \theta_{j+1}^{(t)}, ..., \theta_k^{(t)}]$
4. On génère $\theta_k^{(t+1)}$ en simulant selon la loi $[\theta_j|\theta_1^{(t+1)}, \theta_2^{(t+1)}, ..., \theta_j^{(t+1)}, ..., \theta_{k-1}^{(t+1)}]$

L'étape $t+1$ de l'algorithme est scindée en k sous-étape, chaque sous-étape est un tirage au sort dans la distribution conditionnelle complète de chaque θ_j connaissant l'ensemble des autres variables. À l'origine de chaque étape, les valeurs simulées à l'étape précédente sont utilisées pour spécifier la première conditionnelle complète, celle de θ_1. Ensuite, les valeurs de conditionnement sont progressivement remplacées par les nouvelles valeurs générées à l'étape t+1.
La transition de la chaîne entre deux points différents $\theta^{(t)}$ et $\theta^{(t+1)}$ est donnée par

$$P_G(\theta^{(t)}, \theta^{(t+1)}) = \prod_{j=1}^{k} \pi(\theta_j^{(t+1)}|\theta_1^{(t+1)}, ..., \theta_{j-1}^{(t+1)}, \theta_{j+1}^{(t)}, ..., \theta_k^{(t)}) \tag{1.41}$$

Nous passons maintenant à quelques questions qui se posent dans l'application de l'algorithme de Gibbs. Prmièrement, dans la conception des blocs, les composantes hautement corrélées devraient être regroupées, autrement dit la chaîne est susceptible d'afficher les autocorrélations qui se décomposent lentement ce qui entraine une matière de convergence lente de la densité cible.

Deuxièment, si quelques densités conditionnelles complètes sont diffciles à simuler par l'algorithme de Gibbs, il est préférable d'utiliser l'algorithme de Metropolis-Hastings.
Il existe plusieurs conditions suffisantes pour que la chaîne de Markov générée par l'algorithme de Gibbs soit convergente.

Lemme 1.4.2 ([29]). *Si les densités conditionnelles sont presque partout positives, c'est-à-dire* $\pi(\theta_j^{(t+1)}|\theta_1^{(t+1)}, ..., \theta_{j-1}^{(t+1)}, \theta_{j+1}^{(t)}, ..., \theta_k^{(t)}) > 0, j = 1, ..., k$ *sur* Θ*, la suite* $(\theta^{(t)})$ *est une chaîne de Markov ergodique de loi invariante* $\pi(\theta|x)$*. Si de plus ces densités sont continues, la chaîne est Harris récurrente.*

L'intuition de ces conditions (et leur connexion à l'irréductibilité et l'apperiodicité) devrait être notée. Ces conditions assurent que chaque densité conditionnelle complète est bien définie et son support n'est pas séparé en régions disjointes de sorte qu'une fois la chaîne se déplace dans une région elle n'en sort pas, il faut noter que le support de la loi invariante doit être le produit cartésien des supports des π_j.

Bien que ces conditions soient strictement faibles, elles sont satisfaisantes pour la convergence de l'échantillonneur de Gibbs dans la plupart des applications économétriques.

Nous allons maintenant introduire un exemple dont nous allons établir l'algorithme de Gibbs étape par étape.

Exemple 6. : (Robert et Casella,10.17)

Dans une centrale nucléaire , on s'intéresse au nombre de pannes (y_i) pour chaque 10 pompes, où chaque pompe est observée au temps t_i. Et considérant que les y_i sont modélisés par une loi de poisson $\mathcal{P}(\lambda_i t_i)$, on veut éffectuer une estimation Bayésienne des paramètres λ_i. Pour cela, on doit traiter les λ_i comme des variables aléatoires et par conséquence il faut déterminer une loi a priori pour les λ_i.

Une loi a priori pour λ_i est donnée par une $gamma(\alpha, \beta)$ avec $\alpha = 1.8$ et β inconnu. Donc le modèle de cet exemple contient 11 paramètres inconnus $(10\lambda_i et \beta)$.
Soit une loi a priori $gamma(\gamma, \delta)$ pour β avec $\gamma = 0.01$ et $\delta = 1$. Comme nous l'avons déjá dit, au sens Bayésien toute inférence se faite aprés avoir calculée la loi a posteriori.
Nous avons :

$y_i \sim \mathcal{P}(\lambda_i t_i) \Rightarrow P(y) = \prod_{i=1}^{10} \frac{e^{-\lambda_i t_i}(\lambda_i t_i)^{y_i}}{y_i!}$

$\lambda_i \sim gamma(\alpha, \beta)$

$\beta \sim gamma(\gamma, \delta)$

En appliquant le théorème de Bayes, nous obtenons :

$$\begin{aligned}
\pi(\lambda, \beta | y, t) &\propto P(y) \times \pi(\lambda) \times \pi(\beta) \\
&= \left(\prod_{i=1}^{10} \frac{e^{-\lambda_i t_i}(\lambda_i t_i)^{y_i}}{y_i!} \times \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda_i^{\alpha-1} e^{-\beta\lambda_i} \right) \times \frac{\delta^\gamma}{\Gamma(\gamma)} \beta^{\gamma-1} e^{-\delta\beta} \\
&\propto \left(\prod_{i=1}^{10} e^{-\lambda_i t_i}(\lambda_i t_i)^{y_i} \times \beta^\alpha \lambda_i^{\alpha-1} e^{-\beta\lambda_i} \right) \times \beta^{\gamma-1} e^{-\delta\beta} \\
&= \left(\prod_{i=1}^{10} \lambda^{y_i+\alpha-1} e^{-(t_i+\beta)\lambda_i} \right) \beta^{10\alpha+\gamma-1} e^{-\delta\beta}
\end{aligned}$$

Comme il est déjà mentionné dans les sections précédentes du chapitre, l'estimation Bayésienne consiste soit, à minimiser le risque a posteriori défini par :

$$\int_\lambda \int_\beta \pi(\lambda, \beta | y, t) l(\lambda, \beta, \delta) d\beta d\lambda$$

avec $l(\lambda, \beta, \delta)$ est une fonction de perte donnée. Ou bien à maximiser la loi a posteriori c'est-à-dire déterminer l'estimateur MAP, et dans les deux cas il n'est pas évident d'arriver

TAB. 1.6 – Résultats de la simulation

Paramètres	Moyenne	Écart type
λ_1	0.07113	0.02759
λ_2	0.15098	0.08974
λ_3	0.10447	0.04012
λ_4	0.12321	0.03071
λ_5	0.65680	0.30899
λ_6	0.62212	0.13676
λ_7	0.86522	0.55689
λ_8	0.85465	0,54814
λ_9	1.35524	0.60854
λ_{10}	1.92694	0.40812
β	2.389	0.6986

à déterminer l'estimateur Bayésien, pour cela on fait appel aux méthodes MCMC et plus précisément à l'échantillonnage de Gibbs.

Les distributions a posteriori conditionnelles de λ et β s'obtiennent en éliminant tous les termes qui ne dépendent pas du paramètre dont on veut trouver sa distribution a posteriori conditionnelle. Donc, dans ce cas :

$$\begin{aligned} \pi(\lambda_i|\beta, y, t) &\propto \lambda^{y_i+\alpha-1} e^{(t_i+\beta)\lambda_i} \\ \pi(\beta|\lambda, y, t) &\propto \beta^{10\alpha+\gamma-1} e^{-\beta(\delta+\sum_{i=1}^{10}\lambda_i)} \end{aligned}$$

$\pi(\lambda_i|\beta, y, t)$ est une $\mathbf{gamma}(y_i + \alpha, t_i + \beta)$

$\pi(\beta|\lambda, y, t)$ est une $\mathbf{gamma}(10\alpha + \gamma, \delta + \sum_{i=1}^{10}\lambda_i)$

Les étapes de l'algorithme de Gibbs dans ce cas sont :

1. Premièrement, il faut définir une valeur initiale de β (ici on a pas besoin de donner une valeur initiale à λ parce qu'on commence par générer λ qui dépend seulement de β). $> beta.cur < -1$
2. On génère $\lambda^{(1)}$ selon sa distribution conditionnelle.
   $> lambda.update < -function(alpha, beta, y, t)\{$
   rgamma(length(y),y+alpha,t+beta) }
3. On génère $\beta^{(1)}$ selon sa distribution conditionnelle.
   $> beta.update < -function(alpha, gamma, delta, lambda, y, t)\{$
   rgamma(1,length(y)*alpha + gamma),delta + sum(lambda) }
4. On répète ces étapes en utilisant à chaque fois les valeurs de l'étape précédente comme valeurs initiales de la nouvelle étape.
5. On met tous cela dans une fonction pour calculer les quantité d'intérêt.

Pour un nombre de simulation $n = 10000$, et pour les valeurs initiales suivantes des paramètres : $\beta = 1$, $\alpha = 1.8$, $\gamma = 0.01$ et $\delta = 1$, les résultats de la simulation par l'échantillonnage de Gibbs sont donnés par le tableau(1.6) ci-dessus.

Chapitre 2

La robustesse Bayésienne

2.1 Introduction

Dans la mise en œuvre d'une analyse Bayésienne, le statisticien s'est intéressé comme une première étape à proposer un modèle qui explique le comportement des observations, une loi a priori qui génère le paramètre d'intérêt et une fonction de perte qui est utilisée pour évaluer le risque. Etant donné ces trois éléments, le Bayésien cherche à employer des méthodes qui sont optimales dans un certaine sens.

Cependant dans la pratique, il est rare de pouvoir proposer une détermination explicite du modèle, de la loi a priori et de la fonction de perte même si on dispose de certaines informations. La robustesse Bayésienne consiste à évaluer l'influence de cette indétermination sur les quantités d'intérêt.

Une pléthore de méthodes et d'outils ont été proposés pour faire face à ce problème comme les travaux de Good (1983), Berger et Berliner (1986), Berger et Sellke (1987), Wasserman (1992) et Abraham et Daurés (2000).
La robustesse Bayésienne donc peut être établie par rapport au modèle proposé, à la loi a priori ou parfois par rapport à la fonction de perte quand il s'agit d'un problème de décision. Mais, dans les trois cas elle consiste à construire une classe de modèles/lois a priori/fonctions de perte, et étudier par la suite les changements éffectués sur les quantités a posteriori autours de ces classes.

Ce chapitre est organisé comme suit : la deuxième section aborde quelques notions de base sur la robustesse Bayésienne. Après, nous relieons l'approche de la robustesse Bayésienne par l'approche classique dans la section 3.

2.2 Quelques notions de base

2.2.1 différentes approches

Il existe trois principales approches de la robustesse Bayésienne. La première est **l'approche informelle**, dans laquelle un ensemble de lois a priori est considéré et les moyennes a posteriori correspondantes sont comparées. Cette approche a été (et elle est) trés populaire en raison de sa simplicité. En revanche, il est parfois facile de perdre les lois a priori compatibles avec les connaissances a priori disponibles, ce qui mènerait a des moyennes a posteriori trés différentes.

La deuxième approche est appelée **robustesse globale** (voir Moreno, 2000, pour plus de détails). Cette approche fonctionne mieux que l'approche précédente, elle consiste à considérer une classe de lois a priori compatibles avec les informations a priori disponibles, et évaluer par la suite la différence entre le sup et l'inf des moyennes a posteriori autours de la classe. Cette approche est trés populaire elle même, mais les calculs ne sont pas toujours faciles du fait qu'elle exige l'évaluation du sup et de l'inf des moyennes a posteriori.

La troixième approche est dite **robustesse locale**. Elle est décrite par Gustafson (2000) et Sivaganesan (2000). Elle s'est intéréssée au taux de changements dans l'inférence par rapport aux changements dans la loi a priori utilisant différentes téchniques. Les mesures de sensibilité (robustesse) locale sont généralement plus faciles à calculer que les mesures globales, mais leur interprétation n'est pas toujours claire.

2.2.2 Robustesse par rapport à la loi a priori

Nous allons commencer cette section par un exemple qui montre combien il est important d'introduire la notion de la sensibilité au choix de la loi a priori.

Exemple 7. Supposons qu'on observe une variable aléatoire X qui suit la loi de $Poisson(\theta)$, et supposons aussi qu'il est connu a priori que θ a une distribution continue avec une médiane égale à 2 et un quantile d'ordre 3 égale à 4. i.e. $P^{\pi}(\theta \leq 2) = 0.5$ et $P^{\pi}(\theta \leq 4) = 0.25$. Si ces informations sont les seules connaissances disponibles sur le paramètre θ, les trois distributions suivantes peuvent être considérées comme des lois a priori de θ :

(i) $\pi_1 : \theta \sim exp(a)$ *avec* $a = log(2)$;

(ii) $\pi_2 : log(\theta) \sim \mathcal{N}(log(2), (log(2)/z_{.25})^2)$; et

(iii) $\pi_3 : log(\theta) \sim Cauchy(log(2), log(2))$.

et donc,

(i) sous π_1, $\theta|x \sim Gamma(a+1, x+1)$, et la moyenne a posteriori est $E^{\pi_1}(\theta|x) = (a+1)/(x+1)$

TAB. 2.1 – Les moyennes a posteriori sous π_1, π_2 et π_3

	x									
	0	1	2	3	4	5	10	15	20	50
π										
π_1	.749	1.485	2.228	2.971	3.713	4.456	8.169	11.882	15.595	37.874
π_2	.950	1.480	2.106	2.806	3.559	4.353	8.660	13.241	17.945	47.017
π_3	.761	1.562	2.094	2.633	3.250	3.980	8.867	14.067	19.178	49.402

(ii) sous π_2, si on pose $\gamma = log(\theta)$ et $\tau = log(2)/z_{.25} = log(2)/0.675$ on obtient :

$$\begin{aligned} E^{\pi_2}(\theta|x) &= E^{\pi_2}(exp(\gamma)|x) \\ &= \frac{\int_{-\infty}^{+\infty} exp(-e^{\gamma})exp(\gamma(x+1))exp(-(\gamma - log(2))^2/(2\tau^2))d\gamma}{\int_{-\infty}^{+\infty} exp(-e^{\gamma})exp(\gamma x)exp(-(\gamma - log(2))^2/(2\tau^2))d\gamma} \end{aligned}$$

(iii) sous π_3, et posant aussi $\gamma = log(\theta)$, on obtient :

$$\begin{aligned} E^{\pi_3}(\theta|x) &= E^{\pi_3}(exp(\gamma)|x) \\ &= \frac{\int_{-\infty}^{+\infty} exp(-e^{\gamma})exp(\gamma(x+1)) \left[1 + (\frac{\gamma - log(2)}{log(2)})^2\right]^{-1} d\gamma}{\int_{-\infty}^{+\infty} exp(-e^{\gamma})exp(\gamma x) \left[1 + (\frac{\gamma - log(2)}{log(2)})^2\right]^{-1} d\gamma} \end{aligned}$$

Pour voir l'influence du choix de la loi a priori, on examine les moyennes aposteriori sous les trois différentes lois a priori. Les résultats sont donnés par la table ci-dessus.

On remarque que pour x petit ou modéré ($x \leq 10$), la robustesse est réalisée,i.e. il n'y a pas un grand changement entre les moyennes a posteriori sous les trois lois a priori, et donc le choix d'une loi a priori entre les trois n'a pas d'influence. Par contre pour des grandes valeurs de x, le choix de la loi a priori est trés important et a influencé les moyennes a posteriori, il n'y a pas de robustesse dans ce cas.
Il est clair maintenant qu'il y a des situations où le choix d'une loi a priori parmis d'autres dans une classe peut avoir une influence sur les quantités a posteriori d'intérêt.

Classes de lois a priori

"Comment construire une classe Γ de lois a priori de sorte qu'elle modélise l'incertitude sur la loi a priori ?" est la question fondamentale dans la mise en œuvre d'une robustesse Bayésienne par rapport à la loi a priori. Il existe une litérature vaste qui répond à

cette question, mais quelque soit la méthode, cette construction devrait vérifie les objectifs suivants :

1. La classe doit contenir un nombre maximum de lois a priori raisonnables en évitant les lois a priori déraisonnables qui pourraient conduire à trop manque de robustesse.
2. Pour répondre que Γ ne doit pas exiger l'information a priori qui ne se détermine pas facilement dans la pratique .
3. Le calcul de mesures de robustesse doit être aussi facile que possible.

Suivant la classification de Berger (1990), nous considérons que l'incertitude portante 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.

Nous allons passer en revue dans ce qui suit les types de classes de robustesse les plus couramment utilisés dans la litérature.

Classes de lois conjuguées

Ces classes sont basées sur les lois a priori conjuguées traitées dans le premier chapitre. Elles sont parmi les classes les plus faciles à utiliser dans la pratique, et elles sont typiquement choisies pour des raisons pratiques parce qu'elle fournissent en général des bornes explicites pour les quantités d'intérêt. Par exemple, si $X \sim \mathcal{N}(\mu, \tau^2)$ tels que : $\mu_1 \leq \mu \leq \mu_2$ et $\tau_1^2 \leq \tau^2 \leq \tau_2^2$, on peut considérer la classe :

$$\Gamma_c = \left\{\mathcal{N}(\mu, \tau^2) : \ \mu_1 \leq \mu \leq \mu_2 \ \ et \ \ \tau_1^2 \leq \tau^2 \leq \tau_2^2\right\}$$

pour quelques valeurs spécifiées de $\mu_1, \mu_2, \tau_1^2 \ et \ \tau_2^2$

L'avantage de ces classes est que les quantités a posteriori peuvent être calculées sous forme fermée (pour les lois naturelles conjuguées), ce qui facilite la minimisation et la maximisation des quantités d'intérêt.

Ces classes sont connues aussi par les classes paramétriques et elles sont données en général par :

$$\Gamma_P = \{P : \ p(\theta, \omega), \ \omega \in \Omega\}$$

Si par exemple, notre loi a priori est une $\mathcal{G}(\alpha, \beta)$ on peut considérer comme classe de lois a priori :

- $\Gamma_p = \{\mathcal{G}(\alpha, \beta) : \ \alpha/\beta = \mu\}$
- $\Gamma_p = \{\mathcal{G}(\alpha, \beta) : \ l_1 \leq \alpha \leq \mu_1, \ l_2 \leq \beta \leq \mu_2\}$
- $\Gamma_p = \{\mathcal{G}(\alpha, \beta) : \ l_1 \leq \alpha/\beta \leq \mu_1, \ l_2 \leq \alpha/\beta^2 \leq \mu_2\}$

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, qui sont assez peu compatibles avec l'information a priori.

classes des moments généralisés

Les classes des moments généralisés sont les classes données par :

$$\Gamma_{GM} = \left\{ \pi : \int_{\Theta} H_i(\theta)\pi(\theta)d\theta \leq \alpha_i,\ i = 1, ..., n \right\}$$

Où les H_i sont des fonctions π-intégrables et les α_i sont des nombres réels fixés.

Ces classes ont été considérées premièrement par Betrò et al. (1994) et Goutis (1994). Elles sont trés riches et elles contiennent des classes bien connues comme celle à moments déterminés, i.e. $H_i(\theta) = \theta^i$ qui se considère lorsque l'information a priori disponible ne peut se traduire que par des bornes sur certains moments de la loi a priori. Et la classe des quantiles présentée ci-dessous, i.e. $H_i(\theta) = I_{A_i}(\theta)$.

Les classes à moments déterminés ont été utilisées par Hartigan (1969) et Goldstein (1980) qui a considéré des lois a priori en donnant les deux premiers moments. Bien que la détermination des deux premiers moments : la moyenne et la variance, est souvent trés raisonnable. Il est difficile de préciser les moments d'ordre supèrieur à 2. Ces classes sont assez peu satisfaisantes car elles imposent des conditions fortes sur les queues de la loi a priori et ce d'autant plus qu'elles contiennent des lois peu raisonnables.

La classe des moments généralisés permet d'autres choix des fonctions H_i. Betrò et al. (1994) ont considéré la classe définie par des bornes sur les probabilités marginales pour des ensembles données K_i, dont ils ont pris :

$$H_i(\theta) = \int_{K_i} f(x|\theta)dx$$

et en appliquant le théorème de Fubini, ils ont trouvé que :

$$\int_{\Theta} H_i(\theta)\pi(\theta)d\theta = \int_{K_i} m_\pi(x)dx$$

où $m_\pi(x) = \int_{\theta} f(x|\theta)\pi(\theta)d\theta$ est la densité marginale de x sous π.

Un avantage pour ce choix est qu'à partir de l'information sur θ on peut obtenir des informations sur X.

Classes de voisinages

Ces classes sont introduites par Huber (1964b) pour la détection de point abérrant. Dans ces classes, une loi a priori de référence π_0 est proposée et la sensibilité est étudiée quand un voisinage autour de cette loi a priori est considéré. Bien que certaines classes de lois a

priori, soient les voisinages eux même dans un sens topologique dans l'espace de mesures de probabilités. Qui n'est pas le cas dans le plus populaire entre eux : le ε-contamination voisinage.

Classe d'ε-contamination

Elle est définie comme suit :

$$\Gamma_\varepsilon = \{\pi : \ \pi = (1-\varepsilon)\pi_0 + \varepsilon Q, \ Q \in \mathcal{Q}\}$$

où ε porte l'incertitude sur π_0 et $\mathcal{Q}$ est la classe des contaminations.
Cette classe a été considéré par Huber (1973) dans la robustesse classique. Pour la robustesse Bayésienne, la classe d'ε-contamination est construite à partir des lois a priors qui ressemblent à la loi a priori de référence π_0.

Le problème majeur lié à l'utilisation de telles classes est la détermination difficile de ε et de $\mathcal{Q}$, notamment à partir du degré d'incertitude sur π_0. En fait il convient de mentionner que, pour tout sous-ensemble mesurable A, il s'ensuit que :

$$(1-\varepsilon)\pi_o(A) + \inf_{Q\in\mathcal{Q}} Q(A) \leq \pi(A) \leq (1-\varepsilon)\pi_0(A) + \sup_{Q\in\mathcal{Q}} Q(A), \tag{2.1}$$

La majorité des chercheurs se sont basé sur la détermination des classes de contamination $\mathcal{Q}$ et ils ont proposé une variété de choix dans la litérature. Le choix le plus évident est de prendre $\mathcal{Q}$ la classe de toutes les distributions. Cependant, ce choix donne une classe trés large et par la suite, la robustesse est à peine atteinte.

Une classe importante qui améliore la classe des contaminations implique l'addition des contraintes de forme, comme l'unimodalité et la symétrie. De teles contraintes peuvent souvent être facilement obtenues, et peuvent de manière trés significative réduire le rang des quantités a posteriori d'intérêt.

Une autre classe d'amélioration, qui est souvent considérée, consiste à ajouter des contraintes de quantiles, parce que la spécification des probabilités des ensembles est généralement plus facile que d'autres considérations.

Parfois, il y a plus d'incertitude dans les queues des lois a priori contaminées. Plus d'incertitude devrait mener à une grande variation dans les queues de l'ensemble A dans la formule (2.1), i.e. ε devrait être plus grand pour θ dans les queues de la distribution. La classe d'ε-contamination avec $\varepsilon = \varepsilon(\theta)$ a été considérée par Moreno et al. (1996).

Moreno et Cano (1992) ont adressé le problème de la robustesse dans un espace des paramètres multidimensionnel, vu la classe des contaminations avec des marginales partiellement connues, i.e. avec quelques quantiles des marginales unidimensionnelles données, ou avec une marginale complètement donnée et les autres connues juste par l'intermédiare

de quelques quantiles. Ce problème a été aussi bien adressé par Lavine et al. (1991).

Comme une classe relative, Gelfand et Dey (1991) ont proposé une classe d'ε-contamination géométrique, définie comme suit

$$\Gamma_g = \left\{\pi : \pi = c_g(\varepsilon)\pi_0^{1-\varepsilon} q^{\varepsilon}, Q \in \mathcal{Q}\right\},$$

avec q est la densité de la mesure Q dans la classe des contaminations $\mathcal{Q}$.

Autres classes de voisinages

Un inconvénient des classes d'ε-contamination est qu'elles ne sont pas des vrais voisinages dans un sens topologique, une variété d'autres classes qui ont une interprétation formelle dans ce sens ont été considérées.

Un premier exemple trés important est le voisinage de variation d'une mesure de probabilité a priori Π_0 qui contient toutes les mesures de probabilité Π qui satisfont

$$sup|\Pi(A) - \Pi_0(A)| \leq \varepsilon$$

pour ε fixé dans l'intervalle [0,1].

Un autre exemple est donné par le voisinage basé sur la fonction de concentration, un outils défini par Cifarelli et Regazzini (1987) comme une généralisation de la courbe de Lorenz. Tandis que, la courbe de Lorenz compare une distribution discrète avec la distribution uniforme, la fonction de concentration tient compte de la comparaison entre deux mesures de probabilité Π et Π_0 considérant la distance enjambée (the range spanned) par la probabilité sous Π, de tous les sous-ensembles avec une probabilité donnée sous Π_0. Fortini et Ruggeri (1994b) ont proposé une méthode basée sur la fonction de concentration pour définir le voisinage d'une mesure de probabilité, et ils l'ont appliqué dans la robustesse Bayésienne (voir Fortini et Ruggeri, 1994a), et nous renvoyons le lecteur intéressé à examiner le papier de Fortini et Ruggeri (2000) et les références qui y sont. Les voisinages de la fonction de concentration sont trés riches puisqu'ils incluent des classes bien connues comme la classe d'ε-contamination et le voisinage de variation. Cependant, la détermination d'un voisinage de fonction de concentration n'est pas trés simple de fait qu'il exige la spécification des fonctions monotones, continues et convexes.

Tandis que les classes de voisinages précdentes ont été définies en utilisant des probabilités ou des densités données en ce qui concerne des mesures dominantes, Basu et DasGupta (1990, 1995) et Basu (1995) ont considéré des classes des bornes de la distribution, définies comme suit

$$\Gamma_{BDG} = \{F : F \textit{ est une cdf et } F_L(\theta) \leq F(\theta) \leq F_U(\theta), \forall\theta\},$$

avec F_L et F_U sont des fonctions de densités cumulatives données (cdf), et $F_l(\theta) \leq F_U(\theta)$. Cette classe est importante d'un point de vue mathématique puisqu'elle inclut, comme cas

I_i	$(-\infty,-2]$	(-2,-1]	(-1,0]	(0,1]	(1,2]	$(2,\infty)$
p_i	.08	.16	.26	.26	.16	.08

TAB. 2.2 – Les intervalles et les probabilités a priori

spéciaux, les voisinages bien connus de Klmogorov et Lèvy. L'elicitation de cette classe est relativement simple et d'autres contraintes, comme la symétrie et l'unimodalité des distributions peuvent être ajoutées comme en Basu (1992).

Des dérivés fonctionnelles, connues par les dérivés de Fréchet, sont parfois considérées dans la robustesse Bayésienne et elles sont définies sur un espace linéair, normé. Dans ce cas, une loi a priori de référence π_0 est donnée, et la classe $\{\pi; \pi = \pi_0 + \delta\}$ est considérée, où δ est une mesure avec $\delta(\Theta) = 0$. Notons que, l'ensemble Δ des δ est un espace linéair, normé par la norme $\|\delta\| = d(\delta, 0)$ où $d(P,Q) = \sup_{A \in B(\theta)} |P(A) - Q(A)|$ est la métrique de variation totale (the total variation metric). Choisissant $\delta = (1-\varepsilon)(Q - \pi_0)$, avec Q est une mesure de probabilité, $\pi_0 + \delta$ devient une loi a priori d'ε-contamination.

Classes des quantiles

Dans ce cas, il n'y aura pas une loi a priori de référence et la classe des lois a priori sera tout simplement celle qui satisfait les contraintes déjà décrites. L'exemple le plus commun d'une condition du moment généralisé est la spécification d'un quantile.

Un exemple d'une classe des quantiles est celui considéré dans Berger et O'Hagan (1988), O'Hagan et Berger (1988) et Moreno et Cano (1989). La mise en place de cet exemple se procède par la détermination des probabilités a priori pour chacun des intervalles $I_i, i = 1, 6$, dans la table (2.2) ci-dessus. Les probabilités indiquées sont les valeurs obtenues, et la classe des quantiles se compose de toutes les distributions qui sont compatibles avec cette évaluation. Notons que $\mathcal{N}(0, 2)$ est une de ces lois a priori.

Le calcul des rangs des quantités a posteriori est plutôt simple pour une classe des quantiles, puisque les bornes supèrieures et inférieures sont réalisées pour les distributions discrète donnant la masse à un point dans chaque intervalle I_i. Le lecteur intéressé par ce point peut consulter Ruggeri (1990).

Autres classes

Nous concluons cette illustration globale des classes de lois a priori par quelques autres classes qui ont joué un rôle approprié dans le développement de la robustesse Bayésienne au début des années 90.
DeRobertis et Hartigan (1981) ont considéré la classe de rapport de densités Γ_{DR} définie comme suit

$$\Gamma_{DR} = \{\pi : L(\theta) \leq c\pi(\theta) \leq U(\theta) \textit{ pour quelques } c > 0\},$$

où L et U sont des fonctions non négatives donnés.
Le choix de ces fonctions est délicat et a des conséquances importantes, car, si elles sont similaires, toutes les lois dans Γ_{DR} auront les mêmes queues. Voir DeRobertis et Hartigan (1981) et Abraham et Daurés (2000) pour des classes similaires. Ruggeri et Wasserman (1995) ont considéré un voisinage du rapport de densités autour de Π_0 (avec probabilité π_0), en prenant $L(\theta) = \pi_0(\theta)$ et $U(\theta) = k\pi_0(\theta)$ pour $k > 0$ et presque tout θ.

Lavine (1991) a considéré la classe des densités bornées Γ_B définie comme suit

$$\Gamma_B = \{\Pi : L(A) \leq \Pi(A) \leq U(A) \forall A \in \mathcal{F}\},$$

avec $\mathcal{F}$ est une σ-algèbre de Θ et L et U sont des mesures finies. Ruggeri et Wasserman (1991) ont considéré un voisinage de densités bornés autour de Π_0, en prenant $L = (1/k)\Pi_0$ et $U = k\Pi_0$ pour $k > 0$.

Une autre classe intéressante, mais tout à fait négligée dans la litérature est la classe des fonctions de densité $\pi(\theta)$ monotoniquement croissantes. Elle a été proposée par Madanski (1990).

Finalement, les classes mixtes de la forme

$$\Gamma_M = \left\{\pi : \pi(\theta) = \int \pi(\theta|\omega) dG(\omega)\right\},$$

ont été considérées par Bose (1994), et elles proviennent de l'idée que beaucoup de classes de lois a priori sont des mélanges des lois a priori.

Une fois une classe Γ de lois a priori est construite, des mesures de sensibilité sont nécessaires pour examiner la robustesse des procédures d'inférence sous cette classe. Pendant ces dernières années, deux types de ces mesures ont été étudiés. Les mesures globales comme, les bornes (the range) des quantités a posteriori ; et les mesures locales comme, les dérivées de ces quantités que nous aborderons ultèrierement.

2.2.3 Les mesures globales de la sensibilité

Nous allons présenter dans cette section quelques mesures globales qui ont été introduites pour étudier les changements éffectués sur les quantités a posteriori d'intérêt quand, la loi a priori varie dans la classe. Ces mesures donnent, en général, un nombre qui, en principe, devrait être interpréter de la façon suivante :

- Si ce nombre est petit, alors la robustesse est réalisée et n'importe quel a priori dans la classe Γ peut être choisi.
- Si ce nombre est grand, et de nouvelles données peuvent être considérées et/ou une autre classe Γ_1 rétrécit la classe Γ, alors recalculons les mesures de la robustesse en s'arrêtant lorsque le nombre deviendra petit.

. Si ce nombre est grand et la classe ne peut pas être modifiée, alors on peut choisir un a priori dans la classe mais on doit être circonspects de l'influence de ce choix sur les quantités d'intérêt.

En donnant une classe Γ de lois a priori, l'analyse globale de sensibilité s'intéressera au rang de la variation d'une quantité a posteriori d'intérêt $T(h,\pi)$ quand la loi a priori π varie dans la classe Γ. C'est-à-dire elle s'intéresse à évaluer la quantité

$$\sup_{\pi\in\Gamma} T(h,\pi) - \inf_{\pi\in\Gamma} T(h,\pi)$$

Comme il est expliqué dans Berger (1990), cette quantité a posteriori d'intérêt peut être en général, dans un sens Bayésien, une des trois catégories suivantes :

(i) Des fonctions linéaires de la loi a priori : $T(h,\pi) = \int_\Theta h(\theta)\pi(d\theta)$, où h est une fonction donnée.
Si h est la fonction de vraisemblance $f(x|\theta)$, on obtient une fonction linéaire importante, la densité marginale des données, i.e., $m_\pi(x) = \int_\Theta f(x|\theta)\ \pi(d\theta)$.

(ii) Les rapports des fonctions linéaires de la loi a priori :

$$T(h,\pi) = E^\pi(h(\theta)|x) = \frac{1}{m_\pi(x)}\int_\Theta h(\theta)\ f(x|\theta)\ \pi(d\theta)$$

pour quelques fonctions h données. Si on prend $h(\theta)=\theta$, $T(h,\pi)$ est la moyenne a posteriori, et pour $h(\theta) = I_{\mathcal{C}}(\theta)$, la fonction indicatrice de l'ensemble $\mathcal{C}$, on obtient la probabilité a posteriori de $\mathcal{C}$.

(iii) Les rapports des fonctions non linéaires : $T(h,\pi) = \frac{1}{m_\pi(x)}\int_\Theta h(\theta,\phi(\theta))\ l(\theta)\ \pi(d\theta)$ pour quelques fonctions h. Pour $h(\theta,\phi(\theta)) = (\theta-\mu(\pi))^2$ avec $\mu(\pi)$ est la moyenne a posteriori, on obtient $T(h,\pi) =$ la variance a posteriori.

Notons que les valeurs extrêmes des fonctions linéaires de la loi a priori quand cette dernière varie dans une classe Γ sont faciles à calculer si les points extrêmes de Γ peuvent être identifiés.

Théorème 2.2.1 ([54]). *Soit Γ_{su} la classe de toutes les distributions a priori unimodales et symétriques, du mode θ_0, alors on a :*

$$\sup_{\pi\in\Gamma_{su}} E^{\pi}(h(\theta)|x) = \sup_{r>0} \frac{\frac{1}{2r}\int_{\theta_0-r}^{\theta_0+r} h(\theta)\ f(x|\theta)\ d\theta}{\frac{1}{2r}\int_{\theta_0-r}^{\theta_0+r} f(x|\theta)\ d\theta}$$

$$\inf_{\pi\in\Gamma_{su}} E^{\pi}(h(\theta)|x) = \inf_{r>0} \frac{\frac{1}{2r}\int_{\theta_0-r}^{\theta_0+r} h(\theta)\ f(x|\theta)\ d\theta}{\frac{1}{2r}\int_{\theta_0-r}^{\theta_0+r} f(x|\theta)\ d\theta}$$

Notons que $E^{\pi}(h(\theta)|x) = \frac{\int_{\Theta} h(\theta)\ f(x|\theta)\ \pi(d\theta)}{\int_{\Theta} f(x|\theta)\ \pi(d\theta)}$

Exemple 8. Supposons que $x|\theta \sim \mathcal{N}(\theta, \sigma^2)$ et qu'on s'intéresse à tester $H_0 : \theta \leq \theta_0$ contre $H_1 : \theta > \theta_0$. Et soit Γ_{su} la classe de toutes les distributions a priori unimodales et symétriques dont on s'intéresse à examiner sa robustesse.
On a

$$\begin{aligned} P^{\pi}(H_0|x) &= P^{\pi}(\theta \leq \theta_0|x) \\ &= \frac{\int_{-\infty}^{\infty} I_{(-\infty,\theta_0]}(\theta) f(x|\theta)\ d\pi(\theta)}{\int_{-\infty}^{\infty} f(x|\theta)\ d\pi(\theta)} \end{aligned}$$

ainsi, en appliquant le théorème précédent on obtient :

$$\begin{aligned} \sup_{\pi\in\Gamma_{su}} P^{\pi}(H_0|x) &= \sup_{r>0} \frac{\frac{1}{2r}\int_{\theta_0-r}^{\theta_0} \frac{1}{2r}\phi(\frac{x-\theta}{\sigma})\ d\theta}{\frac{1}{2r}\int_{\theta_0-r}^{\theta_0+r} \frac{1}{2r}\phi(\frac{x-\theta}{\sigma})\ d\theta} \\ &= \sup_{r>0} \frac{\Phi(\frac{\theta_0-x}{\sigma}) - \Phi(\frac{\theta_0-r-x}{\sigma})}{\Phi(\frac{\theta_0+r-x}{\sigma}) - \Phi(\frac{\theta_0-r-x}{\sigma})} \end{aligned}$$

et de même :

$$\inf_{\pi\in\Gamma_{su}} P^{\pi}(H_0|x) = \inf_{r>0} \frac{\Phi(\frac{\theta_0-x}{\sigma}) - \Phi(\frac{\theta_0-r-x}{\sigma})}{\Phi(\frac{\theta_0+r-x}{\sigma}) - \Phi(\frac{\theta_0-r-x}{\sigma})}$$

Ces bornes sont atteintes respectivement pour 0.5 et α, avec $\alpha = \Phi(\frac{x-\theta_0}{\sigma})$, la p-value.

La détermination de ces mesures reste toujours un problème non défini puisqu'il n'est pas clair lorsque nous pouvons dire qu'il y a plus de robustesse dans un problème que dans un autre juste en comparant deux nombres, il n'est même pas possible de trouver un critère objectif qui indique quand la robustesse est atteinte. Ces inconvénients ont poussé les chercheurs à l'ídée de mesurer la sensibilité par une quantité a posteriori appropriée afin de se ramener à une comparaison ou évaluation plus facile.

Gustafson (1994) et sivaganesan (1991) ont suggéré de diviser la mesure de sensibilité d'intérêt par la moyenne a posteriori, tandis que Ruggeri et Wasserman (1995), et Gustafson (1994) ont suggéré de diviser les mesures de sensibilité (locale, dans leur cas) par l'écart type de la loi a posteriori correspondante à une loi a priori de référence π_0. Plus tard et en reposant sur une interprétation de la théorie de la décision, Ruggeri et Sivaganesan (2000) ont suggéré de diviser par l'écart type a posteriori correspondant à la loi a priori alternative π, et ils l'ont appellé la sensibilité relative. Pour une quantité d'intérêt $h(\theta)$, ils l'ont défini par

$R_\pi = \frac{(T(h,\pi)-T(h,\pi_0))^2}{V^\pi}$

Où chaqu'un des termes $T(h,\pi)$ et $T(h,\pi_0)$ égale à $E(h(\theta)|x)$ sous π et π_0 respectivement. Et V^π est la variance a posteriori de $h(\theta)$ relativement à la loi a priori π.
L'idée dans cette considération est que la variance a posteriori mesure l'exactitude dans l'estimation de $h(\theta)$, par conséquent, si la distance au carrés entre $T(h,\pi)$ et $T(h,\pi_0)$ relativement à V^π n'est pas trop grande, la robustesse peut être prévue.

Exemple 9. Soit X une variable aléatoire de distribution de probabilité $\mathcal{N}(\theta,1)$. Et soit $\pi_0 = \mathcal{N}(0,2)$ une loi a priori de référence pour θ. on veut évaluer la sensibilité des inférences a posteriori de $h(\theta)=\theta$ sous la classe Γ de toutes les distributions a priori $\mathcal{N}(0,\tau^2)$ avec $0 \leq \tau^2 \leq 10$.
La distribution a posteriori de θ donnant x sous la loi a priori $\mathcal{N}(0,\tau^2)$ est une loi normale de moyenne $\tau^2 x/(\tau^2+1)$ et de variance $\tau^2/(\tau^2+1)$.
D'où on obtient

$$T(h,\pi) - T(h,\pi_0) = \left(\frac{\tau^2}{\tau^2+1} - \frac{2}{3}\right)x \quad et \quad R_\pi(x) = \frac{(\tau^2-2)^2x^2}{9\tau^2(\tau^2+1)}$$

Il est claire que le rang de $T(h,\pi)-T(h,\pi_0)$ est 8x/33 et $\sup R_\pi(x) = 6.4x^2/99$. Ainsi, la robustesse peut être atteinte pour $0 \leq x \leq 4$ et biensûr pas pour x=10.

2.2.4 Robustesse par rapport au modèle

Le modèle est le composant le plus important dans la statistique inférentielle, et par conséquent les imprécisions liées à la spécification du modèle qui peuvent mener à des inférences imprécises doivent être considérées avec une grande importance.

Dans la statistique classique, beaucoup de travaux ont été effectués à cet égard, notons que la majorité d'eux s'intéressaient au problème de l'influence d'un outlier dans un modèle donné.

L'approche Bayésienne pour la robustesse par rapport au modèle est comme par rapport à la loi a priori, consiste à considérer une classe de modèles compatibles avec les informations disponibles, et évaluer par la suite les changements dans les quantités d'intérêt.
Nous allons donner bréièvement les classes des modèles les plus utilisées dans la littérature.

Classes finies

Ces classes peuvent être constituées en considérant un nombre fini de modèles.
Shyamalkumar (2000) a donné un exemple d'une classe de deux modèles :

$$\mathcal{M} = \{\mathcal{N}(\theta, 1),\ Cauchy(\theta, 0.675)\}$$

En d'autres termes soit $X \sim \mathcal{N}(\theta, 1)$ ou bien $X \sim Cauchy(\theta, 0.675)$. Il a aussi considéré deux classes de lois a priori pour le paramètre commun θ données par :

$$\begin{aligned}\Gamma_{0.1}^{A} &= \{\pi : \pi = 0.9\pi_0 + 0.1q,\ q\ est\ arbitraire\ \}\\ \Gamma_{0.1}^{SU} &= \{\pi : \pi = 0.9\pi_0 + 0.1q,\ q\ est\ une\ distribution\ unimodale\ symetrique\ autour\ de\ zero\ \}\end{aligned}$$

et il s'est interessé à évaluer la quantité $\sup_{\pi in \Gamma} E(\theta|x) - \inf_{\pi in \Gamma} E(\theta|x)$. Les résultats sont donnés dans la table (2.3) ci-dessous.

Classes paramétriques

Une des plus riches classes paramétriques est celle proposée par Box et Tiao. (1962).

$$\Lambda_{BT} = \left\{ f(y|\theta, \sigma, \beta) = \frac{exp\{\frac{-1}{2}|\frac{y-\theta}{\sigma}|^{\frac{2}{1+\beta}}\}}{\sigma 2^{(1.5+0.5\beta)}\Gamma(1.5+0.5\beta)}\ \forall\theta,\ \sigma > 0,\ \beta \in (-1, 1] \right\}$$

Une application de cette classe peut être trouvée dans Shyamalkumar (2000).

TAB. 2.3 – Les rangs des moyennes a posteriori (Shyamalkumar (2000))

Data	Likelihood	$\Gamma^{A}_{0.1}$	$\Gamma^{A}_{0.1}$	$\Gamma^{SU}_{0.1}$	$\Gamma^{SU}_{0.1}$
		$\inf E(\theta\|x)$	$\sup E(\theta\|x)$	$\inf E(\theta\|x)$	$\sup E(\theta\|x)$
$x=2$	Normal	0.93	1.45	0.97	1.12
	Cauchy	0.86	1.38	0.86	1.02
$x=4$	Normal	1.85	4.48	1.69	3.34
	Cauchy	0.52	3.30	0.57	1.62
$x=6$	Normal	2.61	2.48	2.87	5.87
	Cauchy	0.20	5.54	0.33	2.88

Classes des voisinages

On distingue trois types de classes de voisinages :

1. Classes d'ϵ-contaminations

Les classes de modèles de type ϵ-contaminations ont été considérées dans Sivaganesan (1993). Elles sont données par :

$$\Gamma_\epsilon = \{f : f(x|\theta) = (1-\epsilon)f_0(x|\theta) + (1-\epsilon)g(x|\theta),\ g \in \mathcal{G}\}$$

avec f_0 est la densité de référence et $\mathcal{G}$ est la classe des contaminations.

2. Classes des rapports des densités

Ces classes ont été considées par Basu (1995). Elles sont données par

$$\Gamma_{DR} = \{f : L(x-\theta_0) \leq \alpha f(x|\theta_0) \leq U(x-\theta_0),\ \forall x\}$$

avec L et U sont des fonctions non négatives données, et $L(.) \leq U(.)$.

3. Classes des voisinages de la vraisemblance

Sur la base des résultats obtenues par Guevas et Sanz (1988), et d'une manière similaire à celle utilisée par Ruggeri et Wasserman (1993)pour calculer les dérivées de Fréchet, Ruggeri (1991, manuscript) a proposé un voisinage topologique pour la fonction de vraisemblance $l(\theta)$. De plus, il a proposé une classe de voisinages de la vraisemblance donnée par :

$$\Gamma_L = \{l \in \mathcal{L}_p;\ L(\theta) \leq l(\theta) \leq U(\theta)\}$$

où L et U sont des fonctions non négatives mesurables données, et n'appartiennent pas nécessairement à $\mathcal{L}_p$, l'espace de toutes les fonctions intégrables.

2.3 Robustesse par rapport à la fonction de perte

Etant donné le même problème de décision, il est possible que différents décideurs aient des différentes évaluations des conséquences de leurs actions. Ainsi, ils peuvent avoir des différentes fonctions de perte. Dans une telle situation, il peut être nécessaire d'évaluer la sensibilité des procédures Bayésiennes au choix de la fonction de perte.

Exemple 10. Supposons que $X \sim Poisson(\theta)$, et que la loi a priori pour θ est exponentielle de moyenne 1. Supposons aussi que $x = 0$ est observée. Alors la distribution a posteriori de θ est exponentielle de moyenne $1/2$. Ainsi, l'estimateur Bayésien de θ est égale à $1/2$ sous la perte quadratique, qui est la moyenne a posteriori. Et est égale à 0.3465, sous la perte absolute, qui est la médiane a posteriori. Il est claire que ces deux estimateurs sont différents, et cette différence peut avoir un certain impact selon l'utilisation de ces estimateurs.

Une étude de la robustesse Bayésienne par rapport à la fonction de perte peut être établie exactement de la même manière que par rapport à la loi a priori et au modèle. Si une classe de fonctions de perte est disponible, les changement effectués sur les quantités a posteriori peuvent être examinés.

Quelques classes de fonctions de perte

Nous allons présenter dans cette section les classes de fonctions de pertes les plus utilisées dans les études de la robustesse Bayésienne.

Les classes d'ϵ-contamination

Comme nous avons déjà vu, les classes d'ϵ-contamination sont très populaires pour définir les classes de lois a priori. On peut aussi définir un voisinage d'une fonction de perte L_0 comme suit :

$$\mathcal{L}^{\epsilon} = \{L : L(c) = (1-\epsilon)L_0(c) + \epsilon M(c) : M \in \mathcal{W}\}$$

où ϵ représente l'imprécision sur L_0, c est une conséquence de l'ensemble des conséquences C et $\mathcal{W}$ est la classe des fonctions de perte qui contient aussi L_0.

Classes partiellement connues

Sur la base des classes des quantiles des lois a priori, Martín et al. (1998) ont considéré une partition finie C_1, ..., C_n de l'ensemble des conséquences C et ont donné les bornes supèrieurs et inférieurs des fonctions de pertes pour chaque élément de la partition.

$$\mathcal{L}_k = \{L : v_{i-1} \leq L(c) \leq v_i,\ \forall c \in C,\ i = 1, ..., n\}$$

où quelques ensembles C_i pourraient être vides.

Classes paramétriques

la classe paramétrique des fonctions de perte la plus populaire est définie dans Varian (1974) par

$$\mathcal{L}_\gamma = \{L_\gamma : L_\gamma(c) = e^{\gamma c} - \gamma c \ - 1,\ \gamma_L \leq \gamma \leq \gamma_u,\ c = a - \theta\}$$

Un autre exemple est

$$\mathcal{L}_k = \left\{L_k : L_k(c) = -e^{-k(c)},\ k > 0,\ c = a - \theta\right\}$$

D'autres exemples peuvent être trouvés dans Bell(1995). En général, les classes paramétriques sont définies comme suit :

$$\mathcal{L}_w = \{L = L_w,\ w \in \Omega\}\,.$$

2.4 La robustesse Bayésienne et l'approche fréquentiste

Nous pouvons constater une certaine relation entre la robustesse Bayésienne et la robustesse classique. Certaines notions développées dans la robustesse classiques ont été utilisées dans la robustesse Bayésienne. Comme les classes d'ϵ-contamination qui ont été introduites par Hubber (1973) pour modéliser les problèmes des outliers.

L'approche de la sensibilité locale, introduite par Gustafson (2000) et Sivaganisan (2000), a aussi utilisé la notion de la fonction d'influence de la robustesse classique (Hampel et al., 1986). Gustafson a remarqué que la dérivée de Gâteaux peut être écrite de la façon suivante :

$$G_{\rho_0}(Q - \Pi_0) = \int I(z) d[Q - \Pi_0](z),$$

où $I(z)$ est la fonction d'influence. Les graphes de la fonction d'influence peuvent être utiles pour évaluer visuellement la sensibilité.

Peña et Zamar (1996) ont présenté une approche asymptotique de la robustesse Bayésienne qui utilise les notions de la robustesse classique comme les fonctions d'influence. Ceci permet de présenter les usages de la liaison entre les deux approches.

Les deux approches reposent sur des perspectives différentes. Chaqu'une possède des avantages et des inconvénients. Ainsi, on ne peut pas dire que cette approche est meilleure que l'autre, mais par contre, on peut retirer les avantages de chaque approche et les combiner afin d'obtenir une meilleure inférence.

Chapitre 3

Inférence Bayésienne des modèles AR(1)

3.1 Introduction

Le chapitre précédent a présenté une inférence Bayésienne sur un paramètre θ, où les observations sont considérées indépendantes. L'analyse Bayésienne ne se restreint pas à ce cadre et nous allons voir dans ce présent chapitre comment l'analyse Bayésienne peut faire face à des structures de dépendance en étudiant des modèles standards des séries temporelles. Nous n'allons pas proposer une théorie unifiée mais plutôt, nous allons décrire via des exemples les éléments essentiels dans l'étude Bayésienne des séries temporelles en se basant sur les modèles autoregressifs.

Une série temporelle est une suite formée d'observations au cours du temps $\{X_t, t \in T\}$, nous pouvons songer par exemple à l'évolution du nombre de voyageurs utilisant le train, à l'accroissement relatif mensuel de l'indice des prix ou encore à l'occurence d'un phénomène naturel.

Une fois que la série temporelle est observée, divers problèmes liés à cette série doivent être résolus. Deux problèmes intéressants forment souvent l'objectif principal d'une série temporelle. le premier concerne la prévision de la série qui consiste à prévoir les valeurs futures de la série à partir de ses valeurs passées, c'est-à-dire à construire des distributions de probabilités de valeurs non encore observées. Et le deuxième problème concerne la détection des outliers résultants. La résolution de ces problèmes se ramène alors à modéliser la série,à estimer les paramètres inconnus du modèle et à l'ajuster.

Dans un sens classique, la série temporelle s'écrit selon le modèle d'ajustement comme suit :

$$X_t = m_t + S_t + \varepsilon_t \tag{3.1}$$

telles que les observations X_t sont modélisées comme superposition additive d'une tendance déterministe m_t, d'une saisonalité S_t, et une perturbation aléatoire ε_t qui représente les erreurs, de moyenne nulle mais qui possède une strecture de corrélation non nulle.
Ce modèle classique connaît comme généralisation

$$X_t = f(t, \varepsilon_t) \tag{3.2}$$

et on distingue deux types d'ajustement :
- Un ajustement additif de type : $f(t, \varepsilon_t) = g(t) + \varepsilon_t$, pour lequel, le modèle (3.1) est un exemple avec $g(t) = m_t + S_t$
- Et un ajustement multiplicatif, c'est-à-dire : $f(t, \varepsilon_t) = g(t)\varepsilon_t$

Une autre manière de modéliser une série temporelle est d'utiliser les modèles autoprojectifs, dont la variable aléatoire X_t s'explique par une relation fonctionnelle $f(.)$ avec ses valeurs passées et une perturbation aléatoire :

$$X_t = f(X_{t-1}, X_{t-2}, ..., \varepsilon_t) \tag{3.3}$$

Une classe importante de modèles autoprojectifs sont les modèles autoregressifs-moyenne mobile (ARMA), dont une étude détaillée est faite par Box et Jenkis (1976).

Les perturbations ε_t dans le modèle (3.1) peuvent être de corrélation forte et suivre un processus ARMA.

La mojorité de méthodes d'analyse des séries temporelles reposent sur l'hypothèse de stationnarité de cette dernière, et si cette hypothèse n'est pas satisfaite alors, la série doît être stationnarisée avant de passer à une étape ultérieure d'analyse. Une des causes de non stationnarité d'une série temporelle est la présence de la tendance m_t.

Après avoir donné des outils préliminaires sur la stationnarité des processus et surtout des processus autoregressifs dans la deuxième section de ce chapitre, nous aborderons dans la troixième section les problèmes de l'inférence Bayésienne dans les séries temporelles en se basant sur le modèle autoregressif (AR). Par la suite, la dernière section du chapitre contient une conclusion.

3.2 Outils préliminaires sur les séries temporelles

3.2.1 Stationnarité d'une série temporelle

Dans ce paragraphe, nous donnons quelques notions de stationnarité d'une série temporelle avec quelques méthodes de stationnarisation de la série dans le cas où cette dernière ne satisfait pas la stationnarité.

Comme nous l'avons déjà vu, toute série temporelle admet la décomposition classique :

$$X_t = m_t + S_t + \varepsilon_t$$

avec
m_t est la tendance qui est une fonction linéaire continue du temps décrite par un nombre fini de paramètres.
S_t est la saisonalité qui est une fonction périodique : Si d est la période de S_t, on a

$$\begin{cases} S_t = S_{t+d} \\ \sum_{j=1}^{d} S_{t+j} = 0 \end{cases}$$

et ε_t est la composante aléatoire.

Nous donnons dans ce qui suit sous formes de notes, les divers sens de stationnarité d'une série temporelle

1. La série temporelle $\{X_t, t \in T\}$ est dite **stationnaire en moyenne** lorsque la moyenne de chacune des variables de la suite est identique.i.e

$$E(X_t) = E(X_0), \forall t \in T$$

2. De même, on dit que cette série est **stationnaire en variance** lorsque :

$$Var(X_t) = Var(X_0), \forall t \in T$$

3. La série $(X_t)_t$ est dite **stationnaire au sens strict** ou **strictement (fortement) stationnaire** si le vecteur $X_{t_1}, X_{t_2}, ..., X_{t_n}$ a la même loi que le vecteur $X_{t_1+k}, X_{t_2+k}, ..., X_{t_n+k}$ pour tout $t_1, t_2, ..., t_n \in Z$ et $n \in N$
4. Et enfin, elle est dite **stationnaire de second ordre ou au sens faible** si les trois conditions suivantes sont vérifiées :

   (i) $E(X_t) = m, \forall t \in Z$

   (ii) $E((X_t)^2) < \infty, \forall t \in Z$

   (iii) $cov(X_t, X_{t+h}) = cov(X_{t-1}, X_{t-1+h}) = ... = cov(X_0, X_h) = \gamma(h)$

La stationnarité au sens strict est peu réalisable en pratique sauf un peu dans le cas gaussien, et est plus exigeant que le concept de la stationnarité faible comme l'indique le lemme suivant.

Lemme 3.2.1 ([25]). *Si la série (X_t) est strictement stationnaire et $E((X_t)^2) < \infty$, alors (X_t) est faiblement stationnaire. La réciproque est fausse en général.*

Aprés avoir donnée la définition d'un processus bruit blanc, nous introduisons un théorème fondamental de l'analyse des séries temporelles stationnaires.

3.2.2 Le processus bruit blanc (white noise)

Soit le processus $\{X_t, t \in \mathbb{Z}\}$. Si pour tout n-uple du temps $t_1 < t_2 < ... < t_n$, les variables aléatoires réelles $X_{t_2} - X_{t_1}$, ..., $X_{t_n} - X_{t_n-1}$ sont indépendantes, il s'agit d'un processus à accroissements indépendants.
Un bruit blanc est un processus stochastique à accroissements non corrélés. Il est dit bruit blanc "fort" si les accroissements sont indépendants. On l'appelle aussi processus purement aléatoire.

Un bruit blanc est donc tel que :

. $E[X_t] = m, \forall t \in \mathbb{Z}$
. $V[X_t] = \sigma^2, \forall t \in \mathbb{Z}$
. $cov[X_t, X_{t+k}] = \gamma_x(k) = 0, \forall t \in \mathbb{Z}, \forall k \in \mathbb{Z}$

Si $E[X_t] = 0$, le bruit blanc est centré.

Dans le reste de ce travail, en écrivant bruit blanc, nous sous-entendons bruit blanc stationnaire et on le note souvent ε_t.

Théorème 3.2.1 ([25]). *Théorème de Wold*

Tout processus stationnaire du second ordre peut être représenté sous la forme :

$$X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} + \mu_t$$

où

. ε_t *est le bruit blanc*
. μ_t *est une composante linéaire déterministe avec* $cov(\mu_t, \varepsilon_{t-j}) = 0$
. ψ_j *des paramètres satisfaisant :* $\psi_0 = 1$, $\psi_j \in \mathbb{R}$, $\sum_{j=o}^{\infty} (\psi_j)^2 < \infty$

La condition sur les ψ_j assure l'existence des moments d'ordre deux du processus.

3.2.3 Processus autorégressifs

On dit qu'un processus $\{X_t, t \in \mathbb{Z}\}$ satisfait une représentation autoregressive d'ordre p notée $AR(p)$ s'il vérifie l'équation :

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} + \varepsilon_t \tag{3.4}$$

où les coéfficients ϕ_i sont fixés avec $\phi_p \neq 0$ et où ε_t est un bruit blanc centré.

On note $\Phi(B)$ le polynôme en B

$$\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p$$

avec B est l'opérateur de retard défini par : $B^i X_t = X_{t-i}$
d'où l'équation (3.4) peut s'écrire :

$$\Phi(B) X_t = \varepsilon_t, \forall t \tag{3.5}$$

Dire que le processus $AR(p)$ est complètement déterminé par l'équation (3.4) implique que toute l'explication (linéaire) du présent par le passé est contenue dans seulement p variables explicatives, celles du passé le plus proche. Ce qui donne une structure markovienne aux processus autoregressifs.

Les processus satisfaisant une représentation $AR(p)$ sont toujours inversibles. Et ils sont stationnaire lorsque les racines du polynôme $\Phi(B)$ notés $\lambda_i \in \mathbb{C}, \forall i < p$ sont en module strictement supèrieur à l'unité.

Une solution de l'équation (3.4) est donnée par :

$$X_t = \frac{1}{\Phi(B)} \varepsilon_t \tag{3.6}$$

et en décomposant $\Phi^{-1}(B)$ en éléments simples on obtient :

$$X_t = \Phi^{-1}(B)\varepsilon_t = \left[\sum_{i=1}^{p} \frac{k_i}{1-\mu_i B} \right] \varepsilon_t \tag{3.7}$$

les k_i sont des constantes, et les μ_i sont les racines du polynôme caractéristique de l'équation (3.4) qui est donné par $g(z) = z^p + \phi_1 z^{p-1} + ... + \phi_p$.

Si $|\mu_i| < 1$, pour tout i, chaque terme de (3.7) peut être développé en une série convergente de $\varepsilon_t, \varepsilon_{t-1}, ...$ et X_t sera une somme de série convergente qui représente la solution stationnaire de l'équation (3.4).

Si $|\mu_i| > 1$, pour tout i, c'est-à-dire elles sont toutes à l'extèrieur du cercle unité, la solution (3.7) ne serait pas convergente . Mais il est toujours possible d'obtenir une solution stationnaire de (3.4) qui sera en fonction de ε_t, ε_{t+1},... , et dans ce cas Hannan (1970) a montré qu'on peut attribuer une représentation autoregressive à un tel processus tel que son polynôme caractéristique ait ces racines à l'intèrieur du cercle unité.

En revanche, le cas le plus délicat à traiter est celui où le polynôme caractéristique de l'équation (3.4) a une racine ou plus sur le cercle unité i.e. il existe i tel que $|\mu_i| = 1$ (la racine unité). Dans ce cas l'équation (3.4) n'admet pas une solution stationnaire.

Donc on peut conclure que la condition suffisante et necessaire pour qu'un processus autoregressif soit stationnaire est que Les racines de son polynôme caractéristique qui sont l'inverse des racines du polynôme $\Phi(B)$ soient à l'intérieur du cercle unité c'est-à-dire $|\mu_i| < 1, \forall i$.

Processus autoregressif d'ordre 1

Un processus autoregressif d'ordre 1 est un cas particulier du processus AR(p) où $p = 1$, on le note $AR(1)$ et il satisfait l'équation

$$X_t = \phi X_{t-1} + \varepsilon_t \tag{3.8}$$

où ϕ est le paramètre autoregressif, et ε_t est un bruit blanc centré. Le polynôme caractéristique de l'équation (3.8) est donné par, $g(z) = z - a$, et dans ce cas la racine du polynôme caractéristique est le paramètre autoregressif lui même.
Donc la condition de stationnarité est : $|\phi| < 1$ et une solution stationnaire de (3.8) est

$$X_t = \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j} \tag{3.9}$$

Si $|\phi| > 1$, la série (3.9) ne converge pas, mais il est toujours possible d'obtenir une solution stationnaire en réecrivant le processus $AR(1)$ sous la forme

$$X_t = \frac{1}{\phi} X_{t+1} - \frac{1}{\phi} \varepsilon_{t+1} \tag{3.10}$$

et la solution sera donnée par

$$X_t = -\sum_{j=1}^{\infty} \phi^{-j} \varepsilon_{t+j} \tag{3.11}$$

Cependant cette solution relie X_t avec les valeurs futures des innovations $\varepsilon_{t+j\,j>1}$. Elle n'est donc pas naturelle car le processus X_t est corrélé avec les réalisations non encore observées du processus ε_t.
Si $|\phi| = 1$, l'équation (3.8) n'admet pas de solution stationnaire.

3.2.4 Processus moyenne mobile

Un processus moyenne mobile d'ordre q est un processus $\{X_t, t \in \mathbb{Z}\}$ dont les variables vérifient

$$X_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + ... + \theta_q \varepsilon_{t-q} \tag{3.12}$$

où θ_1, θ_2, θ_q sont des paramètres réels, et ε_t est un bruit blanc centré.

En utilisant l'opérateur retard B, l'équation (3.12) est équivalente à

$$X_t = \Theta(B)\varepsilon_t \tag{3.13}$$

avec

$$\Theta(B) = 1 + \theta_1 B + ... + \theta_q B^q = \sum_{j=o}^{q} \theta_j B^j$$

Un processus qui satisfait une représentation $MA(q)$ est toujours stationnaire sans tenir compte des valeurs de θ_1, θ_2,..., θ_q.
En effet

(i) $E(X_t) = E(\varepsilon_t) \sum_{j=0}^{q} \theta_j = 0$

(ii) $V(X_t) = V(\varepsilon_t) \sum_{j=0}^{q} \theta_j^2 = \sigma_\varepsilon^2 \sum_{j=0}^{q} \theta_j^2$

(iii) $cov(X_t, X_{t+h}) = E(X_t X_{t+h}) = \sigma_\varepsilon^2 \sum_{j=0}^{q} \theta_j \theta_{j+1}$

Les conditions de la stationnarité de second ordre sont toujours vérifiées, d'où la stationnarité. Mais contrairement au processus autoregressif, le processus moyenne mobile n'est pas toujours inversible et pour qu'il le soit, il faut que son polynôme caractéristique ait toutes ces racines à l'intérieur du cercle unité.

3.2.5 Processus autoregressif-moyenne mobile

On dit que $\{X_t, t \in \mathbb{Z}\}$ est un processus autoregressif-moyenne mobile d'ordre (p,q) noté $ARMA(p, q)$ s'il satisfait l'équation

$$X_t + \phi_1 X_{t-1} + ... + \phi_p X_{t-p} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + ... + \theta_q \varepsilon_q \tag{3.14}$$

Et en utilisant l'opérateur de retard, on peut écrire

$$\Phi(B) X_t = \Theta(B) \varepsilon_t \tag{3.15}$$

Comme le processus $MA(q)$ est toujours stationnaire, la stationnarité d'un processsus $ARMA(p, q)$ repose sur la stationnarité du processus $AR(p)$ qui est vérifiée comme nous l'avons vu si les racines de son polynôme caractéristique sont à l'intérieur du cercle unité. En revanche, l'inversibilité de ce processus repose sur l'inversibilité de la partie $MA(q)$.

Et la solution stationnare est

$$X_t = \Phi^{-1}(B)\Theta(B)\varepsilon_t$$

Comme nous avons déjà mentionné auparavant, la majorité des méthodes d'analyse d'une série temporelle repose sur l'hypothèse de la stationnarité qui n'est pas toujours atteinte, et dans ce cas il est préférable de stationnariser la série avant de passer à une étude quelconque.

Une des raisons de la non stationnarité d'une série temporelle est la présence d'une tendance déterministe ou même d'une tendance stochastique qui se caractérise par la présence

d'une racine unité dans la partie autoregressive de la série. Et dans ce cas pour stationnariser la série il suffit d'éliminer (d'extraire) cette tendance.
Deux méthodes d'extraction d'une série sont les plus utilisées :

(i) L'estimation de la tendance : qui peut se faire en deux manière. Estimation paramétrique dont la tendance m_t est une fonction linéaire, exponentielle, quadratique,..., et utilisant la méthode des moindres carrés on peut estimer facilement les paramètres de m_t. Sinon, si la tendance est une fonction quelconque : ni linéaire ni quadratique, on fait une estimation non paramétrique en effectuant un lissage par moyenne mobile.

(ii) La différenciation qui consiste à appliquer à la série l'opérateur de différence $\triangle$ défini par $\triangle Y_t = Y_t - Y_{t-1}$. La série sera dite alors différenciée ou intégrée.

3.3 inférence Bayésienne des séries temporelles

Cette section décrit l'utilisation des méthodes Bayésiennes dans l'analyse statistique des séries temporelles, et l'importance des méthodes MCMC qui ont rendu même les modèles les plus compliqués des séries temporelles favorables à l'analyse Bayésienne. Comme l'échantillonneur de Gibbs et la technique de Metropolis-Hastings. Le modèle qui va être discuté en détails est le modèle AR(1).

Tandis qu'avant les années 1990, l'inférence Bayésienne a été au mieux une entreprise difficile dans la pratique ; réservée à un nombre restreint de chercheurs spécialisés et limitée à un nombre plutôt restreint des modèles. Elle est devenue maitenant un procédé trés accessible et performant qui peut assez facilement être appliqué à presque tout type de modèles ; grâce aux algorithmes de calcul qui sont globalement numériques et en particulier les méthodes de Monte Carlo par chaînes de Markov (les MCMC).

La comparaison des modèles peut être aussi faite dans un cadre Bayésien utilisant l'odds a posteriori, qui est le produit de l'odds a priori et le facteur de Bayes comme il est indiqué dans le premier chapitre du mémoire. Le facteur de Bayes est trés flexible dans la comparaison des modèles et il résume comment les données d'un modèle sont préférées à un autre. En outre, le paradigme Bayésien semble naturel pour la prévision, en tenant compte de tous les paramètres du modèle ou encore l'incertirude sur le modèle. La distribution prédictive est la distribution d'échantillonnage où les paramètres sont intégrés par rapport à la distribution a posteriori, c'est exactement ce dont nous avons besoin pour prévoir, et prévoir est un objectif clé dans l'analyse des séries temporelles.

Un autre passage éssentiel dans l'analyse d'un modèle des séries temporelles est l'estimation des paramètres, qui peut s'établir classiquement comme elle peut être traitée par les outils Bayésiens, une fois que la loi a priori est choisie, suivant les indications fournis dans le chapitre précédent.

L'approche Bayésienne fait face aux problèmes rencontrés dans l'analyse fréquentiste des séries temporelles telle qu'elle combine les informations apportées par les données avec celles de la loi a priori. Beaucoup d'économistes ont trouvé les méthodes Bayésiennes attrayantes non seulement pour des raisons philosophiques, mais surtout pour leur efficacité dans l'inférence à distances finies.

Un inevitable passage dans l'utilisation du paradigme Bayésien dans l'analyse statistique des séries temporelles est la spécification des distributions a priori pour toutes les quantitées dans le modèle qui sont traitées comme inconnues. Le plus souvent, les lois a priori utilisées sont des lois non informatives.

Quand des informations sur la nature des paramètres d'intérêt sont disponibles, les chercheurs essayent de les transformer à une distribution dite a priori informative comme une façon de traduire leurs croyances. Par exemple, Doan et al.(1984) et Litterman (1986) ont observé que plusieurs séries temporelles en macroéconomie forment des processus aléatoires et ont développé une loi a priori informative connue par "Minnesota prior" . Pastor (2000) et Pastor et Stambaugh (2000) ont aussi utilisé la théorie de la finance pour mettre des lois a priori informatives.

Dans plusieurs applications des modèles autoregressifs et en raison du nombre important de paramètres impliqués, les chercheurs trouvent souvent qu'il est souhaitable d'utiliser des lois a priori non informatives malgré que l'examination de l'effet de cette utilisation sur les distributions a posteriori est relativement rare, voir Kadiyala et Karlsson (1997), Ni et Sun (2003). Généralement ils travaillent avec une loi a priori constante pour les coefficients de régression et une loi a priori de Jeffreys pour la matrice des covariances des erreurs, parce que la combinaison de ces lois a priori conduit à des simulations simples de la distribution a posteriori, et elles sont largement utilisées dans les études de macroéconomie.

Comme nous avons vu dans le premier chapitre du mémoire, les estimateurs Bayésiens sont dérivés de la minimisation du coût a posteriori dans l'espace des paramètres, et nous avons vu aussi que la difficulté de cette minimisation se détermine selon le choix de la loi a priori et de la fonction de perte. Toutefois, la fonction de perte détermine la forme de l'estimateur Bayésien. Pour plusieurs applications des procédures Bayésienne, la moyenne a posteriori des coefficient du modèle AR et de la matrice des covariances des erreurs sont généralement déclarés comme des estimateurs Bayésiens, mais il faut prendre acte, que la moyenne a posteriori comme estimateur Bayésien est optimale que pour certaines fonctions de perte.

Un autre point important dans la mise en œuvre de toute étude d'une série temporelle est la condition de la stationnarité. Même si l'inférence Bayésienne d'un processus non stationnaire peut être conduite, il est préférable d'imposer cette condition avant de passer à d'autres étapes, pour des raisons allant de l'asymptotique à la causalité, en passant par l'identifiabilité et la pratique. De telles contraintes se traduisent dans la distribution

a priori par une restriction sur les valeurs du paramètre. Cette restriction entraine des complications philosophiques et méthodologiques : d'une part, comme le raisonnement Bayésien est éffectué conditionnellement au modèle et à l'échantillon observé, le fait que le processus sous-jacent soit stationnaire ne devrait pas être imposé d'entrer sur le modèle mais exhibé par les données elles-même. D'autre part, parfois la forme de la contrainte sur le paramètre rend délicate la génération d'un paramètre suivant la loi a posteriori, même conditionnellement aux autres paramètres du modèle. Mais cette difficulté ne doit pas être considérée comme un inconvenient car il existe une solution efficace pour ce problème qui passe par une reparamétrisation du modèle.

3.3.1 le modèle AR

Comme indiqué précédement, le modèle AR(p) est donné par

$$x_t = \mu + \sum_{i=1}^{p} \phi_i (x_{t-i} - \mu) + \varepsilon_t \tag{3.16}$$

avec $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ et où le paramètre de position μ est introduit pour plus de généralité. Il est défini par la distribution de x_t conditionnellement au passé $(x_{t-1}, ..., x_{t-p})$. i.e,

$$x_t \sim \mathcal{N}(\mu + \sum_{i=1}^{p} \phi_i (x_{t-i} - \mu), \sigma^2). \tag{3.17}$$

Ce modèle présente la particularité, parmi les autres modèles de séries temporelles de fournir une vraisemblance explicite :

$$L(\mu, \phi_1, ..., \phi_p, \sigma) = \sigma^{-n} \prod_{i=1}^{n} exp \left\{ - \frac{(x_t - \mu + \sum_{i=1}^{p} \phi_i (x_{t-i} - \mu))^2}{2\sigma^2} \right\} \tag{3.18}$$

Si, par convention, nous posons $x_0 = ... = x_{-p} = 0$, il est possible de trouver des lois a priori conjuguées naturelles, normale pour μ, inverse gamma pour σ^2 et normale pour le vecteur $(\phi_1, ..., \phi_p)$. Nous pouvons également utiliser une loi a priori non informative comme celle de Jeffreys qui est controversée dans ce cadre, ou une loi a priori non informative plus courante comme $\pi(\mu, \sigma, \phi) = 1/\sigma$.

Si nous imposons la contrainte de stationnarité du modèle (3.16), cette contrainte se traduit par une restriction du support de ϕ à l'ensemble des ϕ_i tels que le polynôme

$$\Phi(x) = 1 + \phi_1 x + ... + \phi_p x^p.$$

ait toutes ses racines à l'extérieur du cercle unité. Pour des valeurs de p plus grandes que 3, l'espace des paramètres est trop complex pour proposer comme loi a priori la loi conjuguée

normale restreinte à cet espace. Une solution à ce problème est de reparamétriser le modèle. Cette solution est connue par la récurrence de Durbin-Levinson et elle consiste à écrire les paramètres ϕ_1, ..., ϕ_p en fonction des autocorrélations partielles ψ_i suivant la technique de Monahan (1984), qui satisfont sous la contrainte de stationnarité,

$$\psi_i \in]-1, 1[\ , \ i = 1, ..., p.$$

Ce qui autorise l'emploi d'une loi a priori uniforme sur les ψ_i. L'algorithme suivant donne une méthode constructive de calcul des ϕ_i en fonction des ψ_i.

Lemme 3.3.1. *(Barnett et al. (1996))*

Les coefficients ϕ_i se déduisent des ψ_i par l'algorithme suivant :

1. *Définir $\varphi^{ii} = \psi_i$ et $\varphi^{ij} = \varphi^{(i-1)j} - \psi_i \varphi^{(i-1)(j-1)}$, pour $i > 1$ et $j = 1, ..., i-1$.*
2. *Prendre $\phi_i = \varphi^{pi}$ pour $i = 1, ..., p$.*

Bien que les lois a priori et a posteriori de $(\phi_1, ..., \phi_p)$ résultante ne soient pas explicites, nous pouvons simuler les ϕ_i un à un dans une méthode d'échantillonnage de Gibbs comme nous allons le voir dans le prochain exemple. Les autres paramètres du modèle, μ et σ^2 se simulent aisément comme variables normale et inverse gamma, respectivement (voir Barnett et al.(1996)).

L'exemple suivant présente une estimation des paramètres d'un processus AR(1) basée sur les méthodes de simulation MCMC.

Exemple 11. Estimation des paramètres d'un AR(1) par les méthodes de Monte Carlo par chaînes de Markov

Soit X_t un processus autoregressive d'ordre 1 stationnaire,

$$X_t = \mu + \phi_1(x_{t-1} - \mu) + \varepsilon_t \tag{3.19}$$

avec $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ et X_t est normalement distribuée conditionnellement au passé.i.e,

$$X_t \sim \mathcal{N}(\mu + \phi_1(x_{t-1} - \mu), \sigma^2).$$

Les paramètres inconnus dans le modèle sont μ, ϕ_1 et σ^2. On vise à estimer ces paramètres dans un sens Bayésien utilisant les techniques MCMC qui consistent à générer à partir des distributions a posteriori un échantillon afin de mettre en place une chaîne de Markov dont la loi ergodique converge vers cette distribution a posteriori. On emploiera l'échantillonneur de Gibbs.

Comme nous avons indiqué dans le premier chapitre, pour générer un échantillon utilisant la technique de Gibbs il est nécessaire de déterminer une distribution conditionnelle a posteriori pour chaque paramètre conditionnellement aux autres paramètres. Pour cela et comme une première étape, nous devons trouver l'expression de la vraisemblance du modèle, $P(x|\mu, \phi_1, \sigma^2)$.

La vraisemblance du modèle (3.19) est donnée par

$$P(x|\mu, \phi_1, \sigma^2) = P(x_1|\mu, \phi_1, \sigma^2)\prod_{t=2}^{n} P(x_t|x_{t-1}, \mu, \phi_1, \sigma^2)$$

chaque terme dans le produit s'écrit comme suit

$$P(x_t|x_{t-1}, \mu, \phi_1, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} exp\left\{\frac{-[(x_t - \mu) - \phi_1(x_{t-1} - \mu)]^2}{2\sigma^2}\right\}$$

et le premier terme est donné par

$$P(x_1|\mu, \phi_1, \sigma^2) = \frac{\sqrt{(1-\phi_1^2)}}{\sigma\sqrt{2\pi}} exp\left\{\frac{-(x_1 - \mu)^2(1-\phi_1^2)}{2\sigma^2}\right\}$$

Comme chaque étude Bayésienne doit passer par la proposition de la loi a priori, soient $\pi(\mu)$, $\pi(\phi_1)$ et $\pi(\sigma^2)$ les lois a priori des paramètres μ, ϕ_1 et σ^2 respectivement.

En supposant que μ, ϕ_1 et σ^2 sont indépendants a priori et utilisant le théorème de Bayes, la distribution conditionnelle de ϕ_1 est approximative à

$$P(\phi_1|x, \mu, \sigma^2) \propto P(x|\mu, \phi_1, \sigma^2)\pi(\mu)\pi(\phi_1)\pi(\sigma^2).$$

Il est claire que ce n'est pas facile d'extraire les éléments qui dépendent seulement de ϕ_1 dans la dernière expression. Il existe différentes méthodes pour surmonter cette difficulté, l'approche approximative est l'une de ces méthodes.

Cette approche approximative consiste à conditionner à une variable x_0 non observée et la traiter comme un paramètre de plus. Dans ce cas, nous avons besoin de proposer une loi a priori pour x_0, $\pi(x_0)$. On obtient

$$P(\phi_1|x, \mu, \sigma^2, x_0) \propto P(x|\mu, \phi_1, \sigma^2, x_0)P(\mu|\phi_1, \sigma^2, x_0)P(\phi_1|\sigma^2, x_0)P(\sigma^2|x_0)\pi(x_0)$$

où la vraisemblance est donnée par

$$P(x|\mu,\phi_1,\sigma^2,x_0)=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n exp\left\{\frac{-\sum_{t=1}^{n}[(x_t-\mu)-\phi_1(x_{t-1}-\mu)]^2}{2\sigma^2}\right\}$$

Supposons que chacun de μ et x_0 possède la loi a priori $\mathcal{N}(0,\tau^2)$ avec $\tau\longrightarrow\infty$, que $\pi(\sigma^2)\propto 1/\sigma^2$ et que $\pi(\phi_1)$ est la densité a priori uniforme sur (-1,1).
Dans ce cas et comme $\pi(\mu)$, $\pi(\sigma^2)$ et $\pi(x_0)$ ne dépendent pas de ϕ_1, on obtient

$$P(\phi_1|x,\mu,\sigma^2,x_0)\propto P(x|\mu,\phi_1,\sigma^2,x_0)I(-1,1).$$

où I représente la fonction indicatrice. De plus on peut écrire

$$\sum_{t=1}^{n}[(x_t-\mu)-\phi_1(x_{t-1}-\mu)]^2=\sum_{t=1}^{n}(x_{t-1}-\mu)^2\left[\phi_1-\frac{\sum_{t=1}^{n}(x_{t-1}-\mu)(x_t-\mu)}{\sum_{t=1}^{n}(x_{t-1}-\mu)^2}\right]^2.$$

il en résulte que

$$\phi_1|x,\mu,\sigma^2,x_0\sim\mathcal{N}_{(-1,1)}\left(\frac{\sum_{t=1}^{n}(x_{t-1}-\mu)(x_t-\mu)}{\sum_{t=1}^{n}(x_{t-1}-\mu)^2},\frac{\sigma^2}{\sum_{t=1}^{n}(x_{t-1}-\mu)^2}\right),$$

où $\mathcal{N}_{(-1,1)}$ est la distribution Normale tronquée à $(-1,1)$.

De la même manière et suivant les mêmes calculs, on trouve

$$\mu|x,\phi_1,\sigma^2,x_0\sim\mathcal{N}\left(\frac{1}{n(1-\phi_1)}\sum_{t=1}^{n}(x_t-\phi_1x_{t-1}),\frac{\sigma^2}{n(1-\phi_1)^2}\right).$$

encore

$$P(\sigma^2|x,\mu,\phi_1,x_0)\propto P(x|\mu,\phi_1,\sigma^2,x_0)\pi(\sigma^2),$$

et comme $\pi(\sigma^2)=1/\sigma^2$, il en rèsulte que

$$\sigma^2|x,\mu,\phi_1,x_0\sim IG\left(\frac{n}{2}+1,\frac{1}{2}\sum_{t=1}^{n}[(x_t-\mu)-\phi_1(x_{t-1}-\mu)]^2\right),$$

où IG est la loi inverse gamma.

Concernant x_0, il est traité dans cet exemple différemment aux autres paramètres. Puisque la série est stationnaire le rapport dans l'ordre du temps renversé (the relationship in the reverse time order) reste toujours un AR(1). Ainsi, on peut écrire

$$x_0 = \mu + \phi_1(x_1 - \mu) + \eta,$$

avec $\eta \sim \mathcal{N}(0, \sigma^2)$, donc

$$x_0|x, \mu, \phi_1, \sigma^2 \sim \mathcal{N}(\mu + \phi_1(x_1 - \mu), \sigma^2)$$

Dans ce cas, on peut utiliser la technique de Gibbs pour obtenir des chaînes de Markov où les lois ergodiques de ces chaînes sont les distributions a posteriori $p(\mu|x)$, $p(\phi_1|x)$, $p(\sigma^2|x)$ et $p(x_0|x)$ en générant à partir des distributions conditionnelles $p(\mu|x, \phi_1, \sigma^2, x_0)$, $p(\phi_1|x, \mu, \sigma^2, x_0)$,$p(\sigma^2|x, \mu, \phi_1, x_0)$ et $p(x_0|x, \mu, \phi_1, \sigma^2)$ respectivement, et appliquant le théorème ergodique on peut avoir l'estimateur Bayésien de chaque paramètre.

3.3.2 Conclusion

La démarche Bayésienne pour inférer un paramètre θ est la même pour tout type de modèle : Puisque nous avons davantage informations sur ce paramètre, pourquoi ne pas les utiliser pour obtenir des résultats plus exactes ?.

L'inférence Bayésienne d'un modèle autoregressif se fonde sur le même principe : Construire des lois a priori sur les quantités inconnues dans le modèle, transférer cette loi a priori à une distribution dite loi a posteriori en combinnat l'information apportée dans la loi a priori avec celle apportée par les donnée, par le théorème de Bayes. Et enfin établir l'inférence qu'on souhaite faire sur les paramètres d'intérêt. Les différences entre les modèles dépendent tout d'abord de la problématique à étudier, et ensuite du choix de la loi a priori.

Une étude de la robustesse Bayésienne pour les modèles AR(1) et pour tout type de modèles des séries temporelles peut être établie en considérant une classes de lois a priori pour les paramètres du modèle et évaluer les changements sur les quantités a posteriori. Ou de même, si on est devant un problème de modélisation, on peut considérer une classe de modèles compatibles avec les informations disponibles, et évaluer les changements sur les quantités a posteriori quand le modèle change dans la classe.

Nous allons maintenant présenter une application sur l'estimation Bayésienne des paramètres d'un modèle autoregressif d'ordre 1 en utilisant les méthodes MCMC présentées dans le premier chapitre du mémoire.

3.4 Application

L'objectif de cette section est d'appliquer les méthodes de calcul Bayésien (les MCMC) abordées dans le premier chapitre de ce mémoire pour estimer les paramètres d'un modèle autoregressif d'ordre 1, $X_t = \rho X_{t-1} + Y_t$ avec $0 \leq \rho \leq 1$ et Y_t sont des variables aléatoires indépendantes distribuées selon la loi exponentielle de paramètre θ. Cette étude est une autre variante des méthodes présentées par Ibazizen et Fellag (2003) qui ont de leur part une généralisation des résultats de Turkman (1990). Nous avons appliqué les méthodes MCMC (l'algorithme de Gibbs et celui de Metropolis Hastings) pour estimer les paramètres de ce modèle en utilisant la même loi a prioi proposée par Ibazizen et Fellag (2003).

Considérons le modèle autoregressif d'ordre 1 donné par

$$X_t = \rho X_{t-1} + Y_t, \quad t = ... - 1, 0, 1, ... \tag{3.20}$$

où $0 \leq \rho < 1$ et les Y_t sont i.i.d $\sim Ex(\theta)$ de densité

$$f(y) = \theta \exp(-\theta y) I_{[0,\infty[(y)}, \quad \theta > 0$$

X_1 est supposée distribuée selon la loi $Ex((1-\rho)\theta)$ sachant que le processus (3.20) est stationnaire.
La vraisemblance pour $x = \{x_1, x_2, ..., x_n\}$ est donnée par

$$f(x|\theta, \rho) = (1-\rho)\theta^n e^{-\theta(n\overline{x} - \rho S)} I_A(x)$$

où $A = \{x : x_1 > o, x_t - \rho x_{t-1} \geq 0, t = 2, ..., n\}$ et

$$n\overline{x} = \sum_{i=1}^{n} x_t, \quad S = n\overline{x} - (x_n - x_1)$$

Les estimateurs du maximum de vraisemblance ρ_0 et θ_0 pour ρ et θ respectivement sont introduits par Andẽl (1988) comme suit

$$\rho_0 = \min(1, \frac{x_2}{x_1}, ..., \frac{x_n}{x_1}), \quad \theta_0 = \frac{n}{n\overline{x} - \rho_0 S}$$

Turkmann (1990) a considéré une analyse Bayésienne pour le modèle (3.20) basée sur une loi a priori non informative définie comme suit

$$\pi(\theta, \rho) \propto \frac{1}{\theta(1-\rho)} I_{[0,\infty[\times[0,1)}(\theta, \rho) \tag{3.21}$$

Il a derivé les estimateurs de Bayes pour ρ et θ, respectivement

$$\rho^B = \frac{\rho_0}{n-2}\left(\frac{n-1}{1-r^{n-1}} - \frac{1}{1-r}\right)$$

et

$$\theta^B = \frac{n-1}{n\overline{x}r}\frac{1-r^n}{1-r^{n-1}}$$

où $0 < r = 1 - \rho_0(\frac{S}{n\overline{x}}) < 1$.

En 2003, Ibazizen et Fellag ont proposé une analyse Bayésienne pour le même modèle mais utilisant une loi a priori pour ρ et θ plus générale que celle proposée par Turkman (1990)

$$\pi(\theta, \rho, \beta) \propto \frac{1}{\theta}\frac{\rho^{\beta-1}}{1-\rho} I_{[0,\infty[\times[0,1)}(\theta, \rho), \quad \beta > 0 \tag{3.22}$$

qui contient comme cas special ($\beta = 1$) la loi a priori proposée par Turkmann (1990).
Ils ont établi une estimation Bayésienne des paramètres ρ et θ en utilisant les fonctions hypergéométriques, ainsi qu'une prévision.

En s'intéressant aux n premières observations $x = (x_1, x_2, ..., x_n)$, la loi a priori de θ, ρ dans (3.22) se transforme par le théorème de Bayes à

$$\pi(\theta.\rho|x) = C\rho^{\beta-1}\theta^{n-1}e^{-\theta(n\overline{x}-\rho S)} I_{(0,\infty)\times[0,\rho_0]}(\theta, \rho) \tag{3.23}$$

où la constante C est donnée par

$$C^{-1} = \int_0^{\rho_0}\int_0^{\infty} \rho^{\beta-1}\theta^{n-1}e^{-\theta(n\overline{x}-\rho s)}d\theta d\rho = \Gamma(n)\int_0^{\rho_0} \rho^{\beta-1}(n\overline{x} - \rho S)^{-n}d\rho$$

Sous la fonction de perte quadratique, Ibazizen et Fellag ont trouvé les estimateurs Bayésiens suivants :

$$\begin{aligned}
\hat{\rho}_B(\beta) &= \frac{{}_2F_1(\beta+1, n, \beta+2; 1-r)\rho_0\beta}{{}_2F_1(\beta, n, \beta+1; 1-r)(\beta+1)}. \\
\hat{\theta}_B(\beta) &= \frac{{}_2F_1(\beta, n+1, \beta+1; 1-r)}{\overline{x}\,{}_2F_1(\beta, n, \beta+1; 1-r)}
\end{aligned}$$

où

$$\begin{aligned}
{}_1F_1(a, b; x) &= \sum_{m=0}^{\infty}\frac{(a,m)x^m}{(b,m)m!} \qquad pour|x| < 1 \\
&= \int_0^1 \frac{e^{xu}u^{a-1}(1-u)^{b-a-1}}{B(a, b-a)}du
\end{aligned}$$

$$
\begin{aligned}
{}_2F_1(a,b,c;x) &= \sum_{m=0}^{\infty} \frac{(a,m)(b,m)x^m}{(c,m)m!} \qquad pour|x|<1 \\
&= \int_0^1 \frac{u^{a-1}(1-u)^{c-a-1}(1-xu)^{-b}}{B(a,c-a)} du
\end{aligned}
$$

sont les fonctions hypergéométriques pour deux et trois paramètres respectivement, avec $u = \rho/\rho_0$.

Notons que $\hat{\rho}_B(1)$ et $\hat{\theta}_B(1)$ sont exactement les estimateurs Bayésiens proposés par Turkmann (1990).

Ibazizen et Fellag ont établi une étude de la simulation basée sur 10 000 réplications et pour une valeur initiale de $\theta = 0.25$, et différentes valeurs de n, ρ et β. Les résultats de la simulation sont présentés dans les tables (3.1) et (3.2).

L'idée de notre travail est d'utiliser les méthodes MCMC pour estimer les paramètres ρ et θ au lieu d'utiliser les fonctions hypergéométriques.

TAB. 3.1 – Les valeurs simulées de $\hat{\rho}_B$ et des variances a posteriori

n	ρ	ρ_0	$\beta = 0.5$	$\beta = 1.0$	$\beta = 1.5$	$\beta = 2.0$
10	0.1	0.1921 (0.0845)	0.1115 (0.0742)	0.1326 (0.0752)	0.1434 (0.0757)	0.1503 (0.0761)
	0.3	0.3730 (0.0680)	0.2816 (0.0730)	0.3011 (0.0709)	0.3112 (0.0690)	0.3179 (0.0680)
	0.6	0.6435 (0.0425)	0.5921 (0.0471)	0.5960 (0.0458)	0.5988 (0.0451)	0.6011 (0.0446)
20	0.1	0.1455 (0.0436)	0.0952 (0.0430)	0.1075 (0.0417)	0.1138 (0.0412)	0.1179 (0.0410)
	0.3	0.3355 (0.0340)	0.2944 (0.0366)	0.2985 (0.0357)	0.3013 (0.0352)	0.3036 (0.0348)
	0.6	0.6204 (0.0199)	0.5984 (0.0208)	0.5989 (0.0208)	0.5993 (0.0207)	0.5998 (0.0206)
30	0.1	0.1302 (0.0292)	0.0943 (0.0303)	0.1020 (0.0289)	0.1061 (0.0284)	0.1088 (0.0282)
	0.3	0.3231 (0.0225)	0.2975 (0.0234)	0.2988 (0.0357)	0.3000 (0.0231)	0.3010 (0.2999)
	0.6	0.6136 (0.0132)	0.5995 (0.0136)	0.5997 (0.0136)	0.5998 (0.0136)	0.6000 (0.0136)

TAB. 3.2 – Les valeurs simulées de $\hat{\theta}_B$ et des variances a posteriori

n	ρ	θ_0	$\beta = 0.5$	$\beta = 1.0$	$\beta = 1.5$	$\beta = 2.0$
10	0.1	0.3124 (0.1179)	0.2848 (0.1059)	0.2913 (0.1083)	0.2947 (0.1096)	0.2970 (0.1105)
	0.3	0.3123 (0.1166)	0.2771 (0.1036)	0.2830 (0.1053)	0.2864 (0.1063)	0.2887 (0.1071)
	0.6	0.3151 (0.1209)	0.2820 (0.1084)	0.2836 (0.1088)	0.2849 (0.1093)	0.2861 (0.1096)
20	0.1	0.2777 (0.0673)	0.2628 (0.0635)	0.2662 (0.0643)	0.2680 (0.0647)	0.2692 (0.0649)
	0.3	0.2778 (0.0672)	0.2627 (0.0636)	0.2640 (0.0639)	0.2649 (0.0641)	0.2657 (0.0642)
	0.6	0.3124 (0.1179)	0.2848 (0.1059)	0.2913 (0.1083)	0.2947 (0.1096)	0.2970 (0.1105)
30	0.1	0.2674 (0.0514)	0.2572 (0.0494)	0.2592 (0.0497)	0.2603 (0.0499)	0.2611 (0.0500)
	0.3	0.2677 (0.0513)	0.2583 (0.0495)	0.2588 (0.0496)	0.2592 (0.0497)	0.2595 (0.0497)
	0.6	0.2677 (0.0517)	0.2587 (0.0500)	0.2588 (0.0500)	0.2589 (0.0500)	0.2590 (0.0501)

3.4.1 Estimation Bayésienne des paramètres par les méthodes MCMC

En gardant la même loi a priori proposée par Ibazizen et Fellag (2003), la loi a posteriori des paramètres θ et ρ est comme dans la formule (3.23) donnée par :

$$\pi(\theta.\rho|x) = \frac{\rho^{\beta-1}\theta^{n-1}e^{-\theta(n\overline{x}-\rho S)}I_{(0,\infty)\times[0,\rho_0]}(\theta,\rho)}{\Gamma(n)\int_0^{\rho_0}\rho^{\beta-1}(n\overline{x}-\rho S)^{-n}d\rho} \tag{3.24}$$

Il est claire qu'il n'est pas facile de calculer cette loi a posteriori pour cela, nous avons utilisé les méthodes MCMC et plus précisemment la stratégie de Gibbs pour la simuler. Comme nous savons,

$$\pi(\theta.\rho|x) \propto \rho^{\beta-1}\theta^{n-1}e^{-\theta(n\overline{x}-\rho S)}I_{(0,\infty)\times[0,\rho_0]}(\theta,\rho) \tag{3.25}$$

L'échantillonnage de Gibbs considère que la loi cible est le produit des densités complètes. et pour l'appliquer, il faut tout d'abord déterminer la distribution conditionnelle de chaque paramètre. Dans notre cas, les distributions conditionnelles complètes de θ et ρ sont faciles à extraire : On obtient la distribution complète de θ en enlèvant tous les termes dans (3.25) qui ne dépendent pas de θ, et on fait la même chose pour ρ, d'où

$$P(\theta|x) \propto \theta^{n-1}e^{-\theta(n\overline{x}-\rho S)} \tag{3.26}$$

$$P(\rho|x) \propto \rho^{\beta-1}e^{\theta\rho S} \tag{3.27}$$

Remarquons que la loi dans (3.26) est une $gamma(n, n\overline{x}-\rho S)$. Donc l'échantillonnage de Gibbs dans ce cas est mieux placé pour simuler θ. Par contre, pour (3.27), il serait utile d'introduire l'algorithme de Metropolis-Hastings pour simuler ρ.

Comme il est déjà indiqué, l'algorithme de Metropolis-Hastings repose sur la mise en place de la distribution de proposition. Pour cela, nous avons choisi dans notre cas la distribution $gamma(\beta, n.s^2)$ comme distribution de proposition pour générer un candidat α.

Nous avons fait des simulations à partir du modèle (3.20) pour n.sims=10000, $\theta = 0.25$ et pour différentes valeurs de ρ, n et β. Nous avons calculé pour chaque cas les valeurs des estimateurs Bayésiens $\hat{\theta}_B$ et $\hat{\rho}_B$ de θ et ρ et les variances a posteriori. L'objectif est d'étudier l'impact du changement de β et n sur les quantités d'intérêt et de comparer les résultats obtenus avec ceux obtenus par Ibazizen et Fellag (2003).

Les résultats de la simulation sont donnés dans les tables (3.3) pour $\hat{\theta}_B$ et (3.4) pour $\hat{\rho}_B$ ci-dessous. Les variances a posteriori sont entre parnthèses. Nous avons aussi donné pour chaque cas, les valeurs de ρ_o et θ_0, les MLE de ρ et de θ respectivement.

TAB. 3.3 – Les valeurs simulées de $\hat{\rho}_B$ et des variances a posteriori

n	ρ	ρ_0	$\beta = 0.5$	$\beta = 1.0$	$\beta = 1.5$	$\beta = 2.0$
10	0.1	0.1921 (0.0845)	$1.272e-05$ (0.0010)	0.00010 (0.0020)	0.0151 (0.0117)	0.1
	0.3	0.3730 (0.0680)	$3.055e-05$ (0.0030)	0.0017 (0.0224)	0.2339 (0.1243)	0.3
	0.6	0.6435 (0.0425)	0.0013 (0.0281)	0.6	0.6	0.6
20	0.1	0.1455 (0.0436)	$1.040e-05$ (0.0010)	0.00016 (0.0041)	0.0713 (0.0452)	0.1
	0.3	0.3355 (0.0340)	$9.023e-05$ (0.0052)	0.0661 (0.1274)	0.3	0.3
	0.6	0.6204 (0.0199)	0.6	0.6	0.6	0.6
30	0.1	0.1302 (0.0292)	$1.031e-05$ (0.0010)	0.0005 (0.0071)	0.1	0.1
	0.3	0.3231 (0.0225)	0.0010 (0.0175)	0.3	0.3	0.3
	0.6	0.6136 (0.0132)	0.6	0.6	0.6	0.6

TAB. 3.4 – Les valeurs simulées de $\hat{\theta}_B$ et des variances a posteriori

n	ρ	θ_0	$\beta = 0.5$	$\beta = 1.0$	$\beta = 1.5$	$\beta = 2.0$
10	0.1	0.3124 (0.1179)	0.2348 (0.0749)	0.2559 (0.0800)	0.2137 (0.0685)	0.2436 (0.0773)
	0.3	0.3123 (0.1166)	0.2126 (0.0663)	0.2689 (0.0876)	0.2152 (0.0760)	0.2607 (0.0820)
	0.6	0.3151 (0.1209)	0.1102 (0.0356)	0.2765 (0.0876)	0.2681 (0.0854)	0.2519 (0.0809)
20	0.1	0.2777 (0.0673)	0.2407 (0.0544)	0.2188 (0.0472)	0.2787 (0.0636)	0.2544 (0.0568)
	0.3	0.2778 (0.0672)	0.2881 (0.0510)	0.1738 (0.0468)	0.2407 (0.0530)	0.2686 (0.0558)
	0.6	0.3124 (0.1179)	0.2356 (0.0528)	0.2455 (0.0557)	0.2547 (0.0570)	0.2513 (0.0558)
30	0.1	0.2674 (0.0514)	0.2445 (0.0446)	0.2269 (0.0413)	0.2528 (0.0461)	0.2618 (0.0477)
	0.3	0.2677 (0.0513)	0.1564 (0.0228)	0.2518 (0.0459)	0.2478 (0.0451)	0.2671 (0.0486)
	0.6	0.2677 (0.0517)	0.2462 (0.0444)	0.2482 (0.0454)	0.2425 (0.0409)	0.2557 (0.0470)

3.4.2 Interprétation des résultats

On remarque d'après la table (3.3) que pour chaque cas, $\hat{\rho}(\beta)$ est plus proche de la vraie valeur de ρ que ρ_0. De plus pour des valeurs de β plus grandes que 1, et pour n plus grand que 20, $\hat{\rho}$ est exactement égal à la vraie valeur de ρ. Pour θ, et d'après la table (3.4), $\hat{\theta}(\beta)$ est aussi toujours meilleur que θ_0.
Si on compare ces résultats avec ceux obtenus par Ibazizen et Fellag (2003) donnés dans les tables (3.1) et (3.2), on constate que nos résultats sont proches des valeurs obtenues dans Ibazizen et Fellag (2003), et sont plus exactes pour des valeurs de β plus grande de 1 et pour n plus grand que 20. Et ça confirme l'utilité et l'importance des méthodes MCMC dans l'analyse Bayésienne des séries temporelles.

Conclusion et perspectives

Notre mémoire est une préparation nécessaire qui va nous servir dans notre futur travail et que nous avons commencé à réaliser. Nous avons essayé de rassembler les notions de base nécessaires pour établir une inférence Bayésienne d'un paramètre θ : estimation, tests et prévision.

Comme les méthodes MCMC ont rendu la statistique Bayésienne appliquable à presque tout type de modèles, nous nous sommes intéressé dans notre travail à établir une estimation Bayésienne des paramètres d'un modèle autoregressif d'ordre 1 comme un type d'une série temporelle dont l'importance des méthodes MCMC est apparue clairement.

En perspective, il serait intéressant de généraliser ce travail en choisissant judicieusement une loi a priori qui permet de meilleures performances d'estimation Bayésienne.

Aussi, il serait très utile d'aborder la question de la robustesse Bayésienne de ces techniques.

Bibliographie

[1] Abraham, C., et Daurés, J. (2000). Global robustness with respect to the loss function and the prior. Theory and Decision, 48(4), 359-381.

[2] Albert, J.and Chib, S. (1993). Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. Journal of Business and Economic Staistics, 11, 1-15.

[3] Albert, J., Delampady, M. and Polasek, W. (1991). A class of distributions for robustness studies. J. Statist. Inference, 28, 291-304.

[4] Amaral Turkmann, M. A. (1990). Bayesian analysis of an autoregressive process with exponential white noise. Statistics, 4, 601-608.

[5] Anděl, J. (1988). On AR(1) process with exponential white noise. Comm. In Stat., Theory and Methods, 17, 1481-1495.

[6] Barnett, G. et al. (1995). Markov Chain Monte Carlo estimation of autoregressive models with application to Metal concentration in Sludge, 22, 7-13.

[7] Barnett, G., Kohn, R., et Sheather, S. (1996). Bayesian estimation of an autoregressive model using Markov chain Monte Carlo. J. Econometrics, 74, 237-254.

[8] Basu, S. (1992). Variations of posterior expectations for symmetric unimodal priors in a distribution band. Technical Report 214, Departement of statistics and applied probability, University of California, Santa Barbara, USA.

[9] Basu, S. (1995). Ranges of posterior probability over a distribution band. J. Staist. Plan. Inference, 44, 149-166.

[10] Basu, S., et DasGupta, A. (1990). Bayesian analysis under distribution bands. Technical Report. 90-48, Departement of statistics, Purdue University, USA.

[11] Basu, S., et DasGupta, A. (1995). Robust Bayesian analysis with distribution bands. Statist. Decisions, 13, 333-349.

[12] Berger, J.O. (1990). Robust Bayesian Analysis : Sensitivity to the prior. J. Statist. Plan. Inference, 25, 303-328.

[13] Berger, J.O, et Berliner, L.M. (1986). Robust Bayes and empirical Bayes analysis with ϵ-contaminated priors. Ann. Statist., 14, 461-486.

[14] Berger, J. et Bernardo, J. (1989). Estimating a product of means : Bayesian analysis with reference priors. J. American statist. Assoc., 84, 200-207.

[15] 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.

[16] Berger, J. et Bernardo, J. (1992b). Ordered group reference priors with application to the multinomial problem. Biometrika, 79, 25-37.

[17] Berger, J.O, et O'Hagan, A. (1988). Ranges of posterior probabilities for unimodal priors with specified quantiles. In J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds., Bayesian statistics 3, Oxford University Press, Oxford, U.K.

[18] Berger, J.O, et Sellke, T. (1987). Testing a point null hypothesis : The irreconcilability of p-values and evidence. J. American Staist. Assoc., 82, 112-122.

[19] Berger, J.O., Rios Insua, D. and Ruggeri, F. (2000). Bayesian Robustness. In Robust Bayesian analysis (D. Rios Insua and F. Ruggeri, eds.), Springer-Verlag, New York.

[20] Bernardo, J. (1979). Reference posterior distributions for Bayesian inference (with discussion). J. Royal Statist. Soc. Series B, 41, 113-147.

[21] Bernardo, J. et Smith, A. (1994). Bayesian Theory. John Wiley. New York

[22] Betrò, B., Meczaeski, M., et Ruggeri, F. (1994). Robust Bayesian analysis under generalized moments conditions. J. Statists. Plan. Inference, 41, 257-266.

[23] Box, G. et Jenkins, G. (1976). Time series Analusis : Forecasting and control. Holden-Bay, San Fransisco.

[24] Box, G.E.P. et Tiao, G.C. (1962). A further look at robustness via Bayes theorem. Biometrika, 49, 419-432 et 546.

[25] Brockwell. P. J. et Davis. R. (2000). An introduction to time series forcasting. Second edition. Springer-Verlag, New York, USA.

[26] Chib, S. et Greenberg, E. (1994). Bayes inference in regression models with ARMA(p,q) errors. J. Econometrics, 64, 183-206.

[27] Chib, S., and Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. Journal of the American Statistical Association, 49, 327-335.

[28] Chib, S., and Greenberg. (1995). Markov Chain Monte Carlo simulation methods in Econometrics. Washington University, St. Louis MO.USA.

[29] Christian, P. Robert. (2006). Le choix Bayésien : Principes et pratique. Springer-Verlag France, Paris.

[30] Congdon, P. (2001). Bayesian statistical modelling. John Wiley, New York.

[31] DeRobertis, L., et Hartigan, J.A. (1981). Bayesian inference using intervals of measure, Ann. Statist., 1, 235-244.

Doan, T., Litterman, R.B. et Sims, C.A. (1984). Forcasting and conditional projection using realistic prior distributions. Econometric Reviews, 3, 1-100.

[32] Fortini, S., et Ruggeri, F. (1994b). On defining neighbourhoods of measures through the concentration function. Sanklya Ser. A, 56, 444-457.

[33] Fortini, S., et Ruggeri, F. (2000). On the use of concentration function in Bayesian robustness. In D. Rios Insua and F. Ruggeri, eds, Robust Bayesian Analysis, Springer-Verlag, New York, USA.

[34] Gelfand, A., et Dey, D.K. (1991). On Bayesian robustness of contaminated classes of priors. Statist. Decisions, 9, 63-80.

[35] Gelfand, A.E., et Dey, D.K. (1994). Bayesian model choice : Asymptotics and exact calculations. Journal if the Royal Statistical Society, 56, 501-514.

[36] Gelfand, A. et Smith, A. (1990). Sampling based approaches to calculating marginal densities. J. American statist. Assoc., 85, 398-409.

[37] Gelman, A., Carlin, J.B., Stern, H.S., et Rubin, D.B. (1995). Bayesian data analysis. Chapman and Hall, London.

[38] Geman, S. et Geman, D. (1984). Stochstic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell., 6, 721-741.

[39] George, E.P.Box. 1980. Sampling and Bayes inference in scientific modelling and robustness. Journal of the Royal Statistical Society, 143, 383-430.

[40] Glen Barnett, Robert Kohn and Simon Sheather. 1996. Bayesian estimation of an autoregressive model using Markov Chain Monte Carlo. Journal of Econometrics.

[41] Goldstein, M. (1980). The linear Bayes regression estimator under weak assumptions. Biometrika, 67, 621-628.

[42] Good, I.J. (1965). The estimation of probabilities : An essay on modern Bayesian methods. M.I.T. Press, Cambridge, Massachusetts.

[43] Good, I. (1983). Good thinking : The foundations of probability and its applications. University of Minnesota Press, Minneapolis.

[44] Goutis, C. (1994). Ranges of posterior measures for some classes of priors with specified moments. In statist. Review, 62, 245-257.

[45] Gustafson, P. (1994). Local sensitivity of posterior expectations. Ph. D. Dissertation, Departement of Statistics, Carnegie Mellon University, USA.

[46] Gustafson, P. (2000). Local robustness in Bayesian analysis. In D. Rios Insua and F. Ruggeri, eds., Robust Bayesian analysis, Springer-Verlag, New York, USA.

[47] Hampel, F.R., Ronchetti, E.M., Rousseuw, P.J., et Stahel, W.A. (1986). Robust Statistics : The approach Based on Influence Functions. Wiley, New York, USA.

[48] Hartigan, J.A. (1969). Linear Bayesian methods. J. R. Stat. Soc. Ser. B Stat. Methodol., 31, 446-454.

[49] Hastings, W. (1970). Monte carlo sampling methods using Markov chains and their application. Biometrika, 57, 97-109.

[50] Huber, P.J. (1964). Robust estimation of a location parameter. Ann. Math. Statist. 35, 73-101.

[51] Huber, P.J. (1973). The use of Choquet capacities in statistics. Bulletin of the international Statistics Institue, 45, 181-191.

[52] Huerta, G. et West, M. (1999). Priors component structures in autoregressive time series models. Journal of the Royal Statistical Society, 61, 881-899.

[53] Ibazizen, M. et Fellag, H. (2003). Bayesian estimation of an AR(1) process with exponential white noise. Statistics : A journal of Theoritical and Applied Statistics, 37(5), 365-372.

[54] Jayanta, K. Ghosh, Mohan Delampady et Tapas Samanta. (2006). An introduction to Bayesian analysis : Theory and methods. Springer Science and Business Media, LLC, USA.

[55] Jaynes, E. (1980). Marginalization and prior probabilities. In Zellner, A., éditeur, Bayesian Analysis Econometrics and statistics. North-Holland, Amesterdam.

[56] Jaynes, E. (1983). Papers on probability. Statistics and statistical physics. R.D. Rosencrantz. Reidel. Dordrecht.

[57] Jean-Jacques Boreux, Éric Parent et Jacques Bernier. Pratique du calcul Bayésien. Springer-Verlag France, Paris, 2010.

[58] Jerôme Dupuis (LSP-UPS). Technical report (2007). Statistique Bayésienne et Algorithmes MCMC.

Kadiyala, K.R. et Karlsson, S. (1977). Numerical methods for estimation and inference in Bayesian VAR-models. Journal of Applied Econometrics, 12, 99-132.

[59] Kass, R. et Raftery, A. (1995). Bayes factor and model uncertainty. J. American Statist. Assoc., 90, 773-795.

[60] Kass, R. et Wasserman, L. (1996). Formal ruls of selecting prior distributions : A review and annotated bibliography. J. American statist. Assoc., 91, 343-1370.

[61] Lavine, M., Wasserman, L., et Wolpert, R.L. (1991). Bayesian inference with specified marginals. J. Amer. Statist. Assoc., 86, 400-403.

[62] Litterman, R.B. (1986). Forecasting with Bayesian vector autoregression - five years of experience. Journal of Business and Economic Statistics, 4, 25-38.

[63] Madansky, A. (1990). Bayesian analysis with incompletely specified prior distributions. In S. Geisser et al, eds, Bayesian and Liklihood Methods.

[64] Mark Steel. (2008). Bayesian time series analysis. The new palgrave dictionary of economics, second edition.

[65] Martin, J., Rios Insua, D. et Ruggeri, F. (1998). Issues in Bayesian loss robustness. Sankhya, A, 60, 405-417.

[66] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. et Teller, E. (1953). Equations of stats calculations by fast computing machines. J. Chem. Phys., 21, 1087-1092.

[67] Meyns, S. et Tweedie, R. (1993). Markov chaines and stochastic stability. Springer-Verlag, New York.

[68] Monahan, J. (1984). A note on enforcing stationarity in autoregressive-moving average models. Biometrika, 71, 403-404.

[69] Moreno, E. (2000). Global Bayesian robustness for some classes of prior distributions. In Robust Bayesian Analysis, (D.Rios Insua and F.Ruggeri,eds). New York : Springer-Verlag.

[70] Moreno, E., et Cano, J.A. (1992). Classes of bidimensional priors specified on a collection of sets : Bayesian robustness. J. Statist. Plan. Inference, 46, 325-334.

[71] Moreno, E., Martinez, C., et Cano, J.A. (1996). Local robustness and influence for contamination classes of prior distributions (with discussion). In Berger, J.O., Betrò, B., Moreno, E., Pericchi, L.R., Ruggeri, F., Salinetti, G., et Wasserman, L., eds, Bayesian Robustness, Institute of Mathematical Statistics, Hayward, California, USA.

Ni, S. et Sun, D. (2003). Noninformative priors and frequentist risks of Bayesian estimators of vector autoregressive models. Journal of Econometrics, 115, 159-197.

[72] Nummelin, E. (1984). General irreducible Markov chains and non-negative operators. Cambridge : Cambridge university Press.

[73] O'Hagan, A., et Berger, J.O. (1988). Ranges of posterior probabilities for quasi-unimodal priors with specified quantiles. J. Amer. Statist. Assoc., 83, 503-508.

[74] Parent, E. et Bernier, J. (2007).Le raisonnement Bayésien : Modélisation et Inférence. Springer-Verlag France, Paris.

[75] Pasanisi. A. (2004). Aide à la décision dans la gestion des parcs de compteurs d'eau potable. Thèse de Doctorat de l'École Nationale du Génie Rural, des Eaux et des Forêts (ENGREF).

[76] Pastor, L. (2000). Portfolio selection and asset pricing models. Journal of Finance, 55, 179-223.

[77] Pastor, L. et Stanbaugh, R.F. (2000). Comparing asset pricing models : an investment perspective Journal of Financial Economics, 56, 335-81.

[78] Peña, D., et Zamar, R.H. (1996). On Bayesian robustness : An asymptotic approach. In H. Rieder, ed., Robust Statistics, Data Analusis and Computer Intensive Methods. Springer-Verlag, New York, USA.

[79] Peter Congdon. (2003). Applied Bayesian Modelling. John Wiley et Sons Ltd, England.

[80] Peter C.B. Phillips et Werner Poberger. (1996). An asymptotic theory of Bayesian inference for time series. Journal of Econometrica, 64, 381-412.

[81] Rios Insua, D., Ruggeri, F., et Martin, J. (2000). Bayesian Sensitivity Analysis : A review. To appear in Handbook on sensitivity analysis, (A. Saltelli et al. eds.). New York : Wiley.

[82] Robert, C. (1996a). Inference in mixture models. In Gilks, W., Richrdson, S. et Spiegelhalter, D., éditeurs, Markov chain Monte Carlo in practice, pages 441-464. Chapman and Hall, New York.

[83] Robert, C. (1996b). Intrinsic loss functions. Theory and Decision, 40(2), 191-214.

[84] Robert, C et Cassela, G. (1999). Monte Carlo statistical methods. Springer-Verlag, New York.

[85] Robert E.McCulloch and Ruey S.Tsay. (1993). Bayesian inference and prediction for mean and variance shifts in autoregressive time series. Journal of the American Statistical Association, 88, 968-978.

[86] Ruggeri, F. (1990). Posterior ranges of functions of parameters under priors with specified quantiles. Comm. Statist. Theory methods, 19, 127-144.

[87] Ruggeri, F., Rios Insua, D. et Martin, J. (2000). Robust Bayesian Analysis. In Robust Bayesian analysis (D. Rios Insua and F. Ruggeri, eds). Springer-Verlag, New York.

[88] Ruggeri, F., et Sivaganesan, S. (2000). On a global sensitivity measure for Bayesian inference. Sanklya Ser. A, 62, 110-127.

[89] Ruggeri, F., et Wasserman, L. (1993). Infinitesimal sensitivity of posterior distributions. Canad. J. Statist, 21, 195-203.

[90] Ruggeri, F. et Wasserman, L. (1995). Density based classes of priors : infinitesimal properties and approximations. J. Statist. Plan. Inference, 46, 311-324.

[91] Shyamalkumar, N.D. (2000). Likelihood Robustness. In Robust Bayesian Analysis, (D.Rios Insua and F.Ruggeri,eds). New York : Springer-Verlag.

[92] Sivaganesan, S. (1988). Range of posterior measures for priors with arbitrary contaminations. Communication in Statistics A : Theory and Methods, 17, 1591-1612.

[93] Sivaganesan, S. (1991). Sensitivity of some standard Bayesian estimates to prior uncertainty - a comparison. J. Statist. Plan. Inference, 27, 85-103.

[94] Sivaganesan, S. (1993). Robust Bayesian diagnostics. J. Statist. Plan. Inference, 35, 171-188.

[95] Sivaganesan, S. (2000). Global and Local robustness approaches : Uses and limitations. In D. Rios Insua and F. Ruggeri, eds., Robust Bayesian Analysis. Springer-Verlag, New York, USA.

[96] Smith, A.F.M., et Robert, G.O. (1993). Bayesian computation via the Gibbs sampler and related Markov Chain Monte Carlo methods. Journal of the Royal Statistical Society, 55, 3-23.

[97] Tanner, M et Wong, W. (1987). The calculation of posterior distributions by data augmentation. J. American Statist. Assoc., 82, 528-550.

[98] Tierney, L. (1994). Markov chains for exploring posterior distributions. Annals of statistics, forthcomming.

[99] Varian, H.R. (1974). A Bayesian approache to real assessment. In S. Fienberg, and A. Zellner, eds. Studies in Bayesian Econometrics and Statistics in Honor of L. J. Savage. North-Holland, Amesterdam, The Netherlands.

[100] 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.