REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEURE ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE MOULOUD MAMMERI DE TIZI OUZOU
FACULTE DES SCIENCES
DEPARTEMENT DES MATHEMATIQUES

MEMOIRE

Présenté pour l'obtention du diplôme de :

MASTER EN MATHEMATIQUES

OPTION

Probabilités-Statistiques

THEME

Méthodes Monte-Carlo, application Bayésienne

Présenté par :

TOUAM ABDERRAZAK

Devant le jury :

ATIL Lynda	Maître de conférence A	UMMTO	Présidente
BELKACEM Cherifa	Maître de conférence B	UMMTO	Rapporteur
MERABET Dalila	Maître de conférence B	UMMTO	Examinatrice

Soutenu le : 27 septembre 2018.

DEDICACES

Avant toute chose, j'ai un dévouement de reconnaissance à notre **Dieu** le tout puissant, pour m'avoir donner la force dans les moments difficiles d'éditer ce mémoire $\preccurlyeq$ **Dieu Merci** $\succcurlyeq$.

Je dédie ce travail à ma mère, qui à oeuvré pour ma réussite, de son amour, de son soutien, et de tous les sacrifices counsentis durant toute ma vie.

À mon brave père, qui peut être fière et trouvera ici le résultat de longues années de sacrifices pour m'avoir aider à avancer dans la vie, merci pour les valeurs nobles, l'éducation et le soutient permanent venu de toi, et à tout le reste de ma famille .

À tous mes professeurs de l'unversité Mouloud Mammeri de Tizi-Ouzou, qui doivent voir dans mon travail la fierté d'un savoir bien acquis en commençant par:

♣ Monsieur: FELLAG Hocine

♣ Monsieur: BERKOUN Youcef

♣ Monsieur: HAMAZ Abdelghani

♣M^{me} : ATIL Lynda

À mon professeur encadreur:

♣M^{me}: BELKACEM Cherifa

À tous les membres de jury en commençant par:

♠M^{me} ATIL Lynda

♠M^{me} MERABET Dalila

À tous mes collègues de ma promotion en commençant par mademoiselle:

◇*HARROUCHE Lyasmine*

◇*HADDADOU Kamilia*

◇*AIDEN Zahia*

Et enfin, je remercie tous mes amis et tous ceux qui m'ont encouragé et soutenu, je ne saurai citer chacun par son nom, que tous trouvent ici l'éxpression de ma franche et profonde reconnaissance.

Sans oublier de remercier tous **les kabyles** spécialement.

Merci à tous.

Table des matières

Introduction générale

Les méthodes de Monte-Carlo sont des méthodes d'intégration numérique qui utilisent l'aléatoire. Elles permettent de résoudre de nombreux problèmes autrement insolubles comme par exemple l'évaluation d'intégrales sur des domaines complexes et/ou de grande dimension, le calcul de fonctionnelles de processus stochastiques ou l'exploration de densités de probabilité complexes. Les grandes avancées méthodologiques dans le domaine des méthodes de Monte Carlo sont souvent dues aux physiciens (Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E). Néanmoins, ces propositions ne couvrent pas tous les domaines d'application : la recherche méthodologique autour des méthodes de simulation est donc un axe de recherche d'actualité en prise avec de nombreux domaines scientifiques (physique statistique, ingéniérie, statistiques computationnelles, finance,...)

Les algorithmes de Monte-Carlo dépendent de paramètres d'implémentation dont le choix est crucial pour la convergence de l'algorithme et pour son efficacité. Par exemple, dans les algorithmes de type échantillonnage d'importance, les tirages se font sous une loi instrumentale et ces points sont ensuite corrigés par un mécanisme de pondération, pour approcher une loi cible. Dans ce cas, la loi instrumentale est un paramètre d'implémentation. Dans les algorithmes de type Monte Carlo par chaînes de Markov (MCMC), il s'agit de construire une chaîne de Markov stationnaire, de mesure invariante égale à la loi cible, le choix du noyau de transition dépend des propriétés d'ergodicité de la chaîne et donc l'efficacité de l'échantillonneur.

Choisir des paramètres appropriés est un problème difficile, qui demande une certaine expertise de l'utilisateur. En dépit de la facilité d'implémentation des algorithmes de simulation, cet aspect de la mise en oeuvre est un frein à la large diffusion des méthodes de Monte Carlo. En conséquence, les nouveaux algorithmes de simulation s'orientent vers des procédures adaptatives, c'est-à-dire des procédures itératives au cours desquelles l'algorithme corrige la valeur des paramètres d'implémentation en fonction de son comportement passé.
La procédure d'adaptation est guidée par la recherche des paramètres d'implémentation qui optimisent un critère d'efficacité de l'échantillonneur.

Proposer une méthodologie, si possible assez générale pour avoir un domaine d'application très large, est une chose. Vérifier que l'algorithme fait ce pour quoi il a été développé en est une autre. L'étude de la convergence des méthodes de simulation adaptatives s'avère plus complexe que celle des échantillonneurs non-adaptatifs. Par exemple, les tirages issus d'une procédure MCMC adaptative ne sont plus la réalisation d'une chaîne de Markov homogène, et l'étude de ces algorithmes ne relève plus seulement de résultats sur l'ergodicité d'un noyau de transition. La recherche sur les méthodes de simulation doit donc aussi permettre d'aider à la compréhension fine du fonctionnement des échantillonneurs pour identifier les paramètres d'implémentation optimaux ; et permettre d'aider à l'analyse du comportement des algorithmes proposés.

Le document comporte trois chapitres :

Le premier chapitre comprend quelques notions de base de l'analyse Bayésienne, qui apparait comme une approche attrayante devant les difficultés théoriques des approches paramétriques et non paramétriques classiques.

Au chapitre deux, nous introduisons une classe de méthodes de Monte-Carlo utilisant les chaînes de Markov pour contourner les problèmes inhérents aux algorithmes. Les méthodes de Monte-Carlo par chaînes de Markov, notée, MCMC selon l'acronyme anglais, consiste à utiliser une chaîne de Markov ayant pour distribution stationnaire la fonction à simuler pour générer les points nécessaires à l'approximation.

Dans le troisième chapitre, nous verrons l'application des méthodes MCMC à la statistique Bayésienne.
La mise en oeuvre des principes Bayésiens, en dehors de cas d'école, s'est longtemps heurtée aux difficultés pratiques de calcul. Les moyens informatiques, dont on disposait avant les années 1990, n'étaient pas suffisamment puissants. Les problèmes réels, avec leurs dimensions et leurs complexités importantes, faisaient alors la part belle aux méthodes statistiques classiques. La situation, depuis lors, a subi une véritable révolution . Maintenant on peut affirmer qu'il n'existe, au moins a priori, aucun contre-argument justifie a l'emploi des méthodes Bayésiennes quelle que soit la complexité du cas envisagé. On doit ce nouveau paysage scientifique au développement de nouveaux outils de calcul: les méthodes MCMC. Nous verrons l'application des méthodes MCMC à la détection des points de rupture, la détection des points de changement est un sujet d'intérêt pour de nombreuses statistiques appliquées et théoriques, depuis les années soixante-dix.

Chapitre 1

1 L'analyse statistique Bayésienne

1.1 Introduction :

Dans de nombreuses situations d'expériences aléatoires, il semble raisonnable d'imaginer que le praticien a une certaine idée du phénomène aléatoire qu'il est en train d'observer. Or, la démarche statistique classique repose essentiellement sur un principe de vraisemblance qui consiste à considérer que ce qui a été observé rend compte de manière exhaustive du phénomène. Mais l'observation ne fournit qu'une image et celle-ci peut être mauvaise. Certes cet inconvénient est en général gommé par les considérations asymptotiques et un certain nombre de théorèmes permettent d'évaluer la bonne qualité des estimateurs si le nombre d'observations est suffisant. L'analyse bayésienne des problèmes statistiques propose d'introduire dans la démarche d'inférence, l'information dont dispose a priori le praticien. Dans le cadre de la statistique paramétrique, ceci se traduira par le choix d'une loi sur le paramètre d'intérêt.

Dans l'approche classique, le modèle statistique est le triplet $(\mathfrak{X},\mathcal{A},P_\theta,\theta \in \Theta)$.
Le modèle statistique est définie par :

$\mathfrak{X}$: l'espace des observations.
Θ : l'espace des états de la nature (l'espace des paramètres dans le cas d'un problème statistique).
$\mathcal{A}$: l'espace des actions ou décisions.
P_θ : L'ensemble des probabilités.

Ayant un a priori sur le paramètre, modélisé par une densité de probabilité que nous noterons $\pi(\theta)$, on "réactualise" cet a priori au vu de l'observation en calculant la densité a posteriori $\pi(\theta|x)$, et c'est à partir de cette loi que l'on mène l'inférence.

On peut alors, par exemple, de manière intuitive pour le moment, retenir l'espérance mathématique ou encore le mode de cette densité a posteriori comme estimateur de θ.

Le paramètre θ devient donc en quelque sorte une variable aléatoire, à laquelle on associe une loi de probabilité dite **loi a priori**.

On sent bien d'emblée que les estimateurs bayésiens sont très dépendants du choix de l'a priori.

Différentes méthodes existent pour déterminer ces lois a priori. On peut se référer à des techniques bayésiennes empiriques, où l'on construit la loi a priori sur la base d'une expérience passée, usant de méthodes fréquentistes, pour obtenir forme et valeurs de paramètres pour cette loi. Nous verrons que l'on peut aussi modéliser l'absence d'information sur le paramètre au moyen des lois dites **lois non informatives**.

L'approche bayésienne se différencie donc de l'approche classique dans le sens où le paramètre θ n'est plus considéré comme étant totalement inconnu ; il est devenu une v.a. dont le comportement est supposé connu.

1.2 Notions de base

On appelle **modèle statistique bayésien**, la donnée d'un modèle statistique paramétré $(\mathfrak{X},\mathcal{A},P_\theta,\theta \in \Theta)$ avec $f(x|\theta)$ densité de P_θ et d'une loi $\pi(\theta)$ sur le paramètre.
La démarche de l'analyse bayésienne conduit au calcul d'une **loi a posteriori** $\pi(\theta|x)$; actualisation de la loi a priori $\pi(\theta)$ au vu de l'observation.

L'ensemble des observations est noté $\mathfrak{X}$. Dans ce chapitre $x = (x_1,...x_i,...,x_n)$; autrement dit, on dispose d'un échantillon de taille **n**. Le cadre statistique dans ce chapitre étant celui de la statistique inférentielle, les observations x_i sont donc considérées comme des réalisations de variables aléatoires, notées X_i.

Définitions

- On entend par information a priori sur le paramétre θ toute information disponible sur θ en dehors de celle apportée par les observations.

- L'information a priori sur θ est entachée d'incertitude (si ce n'était pas le cas, le paramétre θ serait connu avec certitude et on n'aurait pas à l'estimer!). Il est naturel de modéliser cette information a priori au travers d'une loi de probabilité, appelée loi a priori. Sa densité est notée $\pi(\theta)$.

- Le modèle statistique paramétrique bayésien consiste en la donnée d'une loi a priori et de la loi des observations. On appelle loi des observations, la loi conditionnelle de X sachant θ. Sa densité est notée $f(x|\theta))$, que la variable aléatoire X soit discrète ou continue. Si X est discrète, $f(x|\theta)$ représente $\Pr(X = x|\theta)$. Dans notre travail, on fera l'hypothèse que, sachant θ, les v.a. X_i sont indépendantes et identiquement distribuées. Autrement dit on aura:

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

Indiquons maintenant les autres lois de probabilité qui interviennent en statistique bayésienne.

La loi a posteriori. C'est la loi conditionnelle de θ sachant x. Sa densité est notée $\pi(\theta|x)$. En vertu de la formule de Bayes, on a:

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

La loi du couple (θ,X). Sa densité est notée $h(\theta,x)$. On a donc: $h(\theta,x) = f(x|\theta)\pi(\theta)$.

La loi marginale de X. Sa densité est notée $m(x)$. On a donc: $m(x) = \int_\Theta f(x|\theta)\pi(\theta)d\theta$.

1.3 Théorème de Bayes

Soient A et B deux événements aléatoires tels que $P[B] \neq 0$. La probabilité de A, conditionnellement à la réalisation de B, est par définition exprimée par la relation suivante :

$$P[A|B] = \frac{P[A,B]}{P[B]}.$$

où $P[A,B]$ est la probabilité que les deux événements A et B aient lieu simultanément.

Puisque $P[B,A] = P[A,B]$, alors les deux probabilités conditionnelles $P[A|B]$ et $P[B|A]$ sont reliées par :

$$P[A|B] = \frac{P[A].P[B|A]}{P[B]}$$

Cette équation est une conséquence triviale de la définition de la probabilité conditionnelle, est appelée Formule de Bayes (ou Théorème de Bayes) en l'honneur du Révérend Thomas Bayes (1702 - 1761).
Une version continue de ces résultats, nous donne la distribution conditionnelle de y sachant x définie par

$$g(y \mid x) = \frac{f(x \mid y)g(y)}{\int f(x \mid y)g(y)dy}$$

Avec $f(x|\theta)$ la probabilité des observations conditionnellement à la valeur θ du paramètre du modèle statistique qu'on utilise pour leur description. Il s'agit de la vraisemblance des données, sous le modèle paramètré par θ.
Comme le dénominateur est indépendant de θ, c'est uniquement une constante de normalisation, la formule de Bayes peut s'écrire comme suit :

$$\pi(\theta \mid x) \propto f(x \mid \theta)\pi(\theta)$$

Le passage de la distribution a priori à la distribution a posteriori des paramètres du modèle statistique, exprimé par la formule de Bayes, peut être alors interprété comme une mise à jour de la connaissance, sur la base des observations.

Exemple 1.1. *Considérons un n-échantillon i.i.d.* $\underline{X} = (X_1,...,X_n)$ *de loi exponentielle de paramètre* $\theta > 0$. *(i.i.d. signifie que les v.a.* X_i *sont indépendantes et identiquement distribuées) La vraisemblance a pour expression :*

$$f(\underline{x}|\theta) = (1/\theta)^n \exp\{-\sum_{i=1}^{n} x_i/\theta\}.$$

On prend une loi a priori de type gamma-inverse sur θ. *La densité de cette loi a priori est donnée par :*

$$\pi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}(1/\theta)^{\alpha+1} \exp\{-\beta/\theta\}, \qquad \theta \in \mathbb{R}^+, \qquad \alpha,\beta > 0$$

La loi jointe est donc :

$$f(\underline{x},\theta) = f(\underline{x} \mid \theta)\pi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}(1/\theta)^{n+\alpha+1} \exp\{-(\beta + \sum_{i=1}^{n} x_i)/\theta\}.$$

et la prédictive s'écrit

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^{+\infty} (1/\theta)^{n+\alpha+1} \exp\{-(\beta + \sum_{i=1}^n x_i)/\theta\} = \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(n+\alpha)}{(\sum_{i=1}^n x_i + \beta)^{n+\alpha}}.$$

Ainsi, la loi a posteriori a pour expression:

$$\pi(\theta \mid \underline{x}) = \frac{(\sum_{i=1}^n x_i + \beta)^{n+\alpha}}{\Gamma(n+\alpha)} (1/\theta)^{n+\alpha+1} e^{-(\beta+\sum_{i=1}^n x_i)/\theta}.$$

Il s'agit d'une loi gamma inverse de paramètres $(n+\alpha, \sum_{i=1}^n x_i + \beta)$.

1.4 Les lois a priori

Le choix des lois a priori est une étape fondamentale dans l'analyse bayésienne. Ce choix peut avoir différentes motivations. Les stratégies sont diverses. Elle peuvent se baser sur des expériences du passé ou sur une intuition, une idée que le praticien a du phénomène aléatoire qu'il est en train de suivre. Elles peuvent être également motivées par des aspects de calculabilité. Enfin, ces stratégies peuvent également tenir compte du fait qu'on ne sait rien par des lois non informatives.

1.4.1 Lois a priori conjuguées

L'approche a priori conjuguée, introduite par Raiffa et Schlaifer en (1961), peut être considérée comme un point de départ pour l'élaboration de distributions a priori fondées sur des informations a priori limitées.

Définition 1.1. Famille conjuguée

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

Un exemple trivial d'une famille conjuguée est l'ensemble $\mathcal{F}_0$ de toutes les lois de probabilité sur Θ. L'avantage des familles conjuguées est avant tout de simplifier les calculs. Avant l'essor du calcul numérique, ces familles étaient pratiquement les seules qui permettaient de faire aboutir des calculs.

Les lois a priori conjuguées sont généralement associées à un type particulier de lois d'échantillonnage qui permet toujours leur obtention ; Ces lois constituent ce qu'on appelle des familles exponentielles.

Exemples de lois conjuguées :

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Normale $\mathcal{N}(\theta,\sigma^2)$	Normale $\mathcal{N}(\mu,\tau^2)$	$\mathcal{N}(\varrho(\sigma^2\mu+\tau^2x),\varrho\sigma^2\tau^2)$
Gamma $\mathcal{G}(n,\theta)$	Gamma $\mathcal{G}(\alpha,\beta)$	$\mathcal{G}(\alpha+n,\beta+x)$
Binomiale $\mathcal{B}(n,\theta)$	Bêta $Be(\alpha,\beta)$	$Be(\alpha+x,\beta+n-x)$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha,\beta)$	$\mathcal{G}(\alpha+x,\beta+1)$
Binomiale Négative $Neg(m,\theta)$	Bêta $Be(\alpha,\beta)$	$Be(\alpha+m,\beta+x)$
Normale $\mathcal{N}(\mu,1/\theta)$	Gamma $\mathcal{G}(\alpha,\beta)$	$\mathcal{G}(\alpha+0.5,\beta+(\mu-x)^2/2)$
Multinomiale $\mathcal{M}_n(\theta_1,\theta_2,...,\theta_n)$	Dirichle $\mathcal{D}(\alpha_1,\alpha_2,...,\alpha_n)$	$\mathcal{D}(\alpha_1+x_1,\alpha_2+x_2,...,\alpha_n+x_n)$

Une loi conjuguée peut être déterminée en considérant la forme de la vraisemblance $f(x|\theta)$ et en prenant une loi a priori de la même forme que cette dernière vue comme une fonction du paramètre. Les lois a priori conjuguées obtenues par ce procédé sont dites **naturelles.**

Exemples :

• Considérons une loi Pareto de paramètres (θ,a).

$$f(x|\theta,a)=\frac{\theta a^\theta}{x^{\theta+1}}\mathbb{1}_{[a,+\infty[}(x).$$

Supposons a connu, $f(x|\theta)\propto\theta e^\theta\log(a/x)$, on pourrait donc prendre une loi a priori de type gamma.

• Dans le cas d'une loi binomiale négative de paramètre (n,p),

$$P(X=x\mid p)=C^x_{n+x-1}p^x(1-p)^n,0<p<1,\ x\in\mathbb{N}.$$

On voit clairement qu'une loi naturelle conjuguée sera une loi bêta puisque :

$$P(X=x\mid p)\propto p^x(1-p)^n.$$

Etudions maintenant en détail le cas d'une famille de lois très importante : la famille des lois exponentielles.

1.4.2 Cas du modèle exponentiel

On rappelle tout d'abord la définition du modèle exponentiel.

Définition 1.2. Famille exponentielle

On appelle famille exponentielle à s-paramètres, toute famille de loi de distribution P_θ dont la densité a la forme suivante :

$$f(x|\theta)=\exp\left[\sum_{i=1}^{s}\eta_i(\theta)T_i(x)-B(\theta)\right]h(x). \tag{1}$$

où $\eta_i(.)$ et $B(.)$ sont des fonctions du paramètre θ et $T_i(.)$ sont des statistiques.

Exemples :

• Loi exponentielle :

$$f(x \mid \theta) = \frac{1}{\theta} e^{-x/\theta} \mathbb{1}_{(x)}$$
$$= \exp\left\{-\frac{1}{\theta}x - \log\theta\right\} \mathbb{1}_{(x)}.$$

Ici, s vaut 1, $\eta_1(\theta) = 1/\theta$,$T_1(x) = x$,$B(\theta) = \log\theta$ et $h(x) = \mathbb{1}_{(x)}$.

• Loi binomiale

$$P(X = x \mid \theta) = C_n^x \theta^x (1-\theta)^{n-x}$$
$$= C_n^x \exp\{x\log\theta + (n-x)\log(1-\theta)\}$$
$$= C_n^x \exp\{x\log[\theta/(1-\theta)] + n\log(1-\theta)\}$$

On a $s = 1$,$\eta_1(\theta) = \log(\theta/(1-\theta))$,$T_1(x) = x$,$B(\theta) = n\log(1-\theta)$ et $h(x) = C_n^x$.

• Loi Gamma $f(x|\alpha,\beta) = \dfrac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$.

En remarquant que : $x^{\alpha-1}e^{-\beta x} = \exp\{-\beta x + (\alpha-1)\log(x)$, on peut écrire :

$\eta_1(\alpha,\beta) = -\beta$, $\eta_2(\alpha,\beta) = \alpha - 1$, $T_1(x) = x$, $T_2(x) = \log x$, $B(\alpha,\beta) = \log(\Gamma(\alpha)/\beta^\alpha)$.

Il est classique d'écrire le modèle exponentiel sous la forme dite **canonique** en reparamétrisant : $\eta_i(\theta) \equiv \theta_i$.

$$f(x \mid \theta) = \exp\left[\sum_{i=1}^{s} \theta_i T_i(x) - A(\theta)\right] h(x).$$

On a le résultat suivant qui donne la forme des lois naturelles conjuguées dans le cas du modèle exponentiel.

Proposition 1.1. *(Christian P. Robert 2006)*

- Soit $f(x \mid \theta)$ appartenant à une famille exponentielle. Alors une famille de lois a priori conjuguée pour $f(x \mid \theta)$ est donnée par :

$$\pi(\theta \mid \mu,\lambda) = K(\mu,\lambda)\exp(\theta\mu - \lambda A(\theta)) \tag{2}$$

où $K(\mu,\lambda)$ est une constante de normalisation.

Et la loi a posteriori est de la forme :

$$\pi(\theta \mid x) \propto \exp((\mu + x)\theta - (\lambda + 1)A(\theta)).$$

Exemple 1.2. *Considérons le modèle* $P(X=x)=\dfrac{e^{(\theta-\beta)x}}{1+e^{\theta-\beta}} \qquad x\in\{0,1\}$

Il s'agit d'une loi logistique. Elle appartient bien à la famille exponentielle.

On a :

$$P(X=x)=\exp[(\theta-\beta)x-\log(1+e^{\theta-\beta})]$$

et la représentation : $h(x)=1$, $\theta=[\theta\ ,\ \beta]'$, $T(x)=[x\ ,\ -x]'$

et $A(\theta,\beta)=\log(1+e^{\theta-\beta})$.

En appliquant la proposition (1.1), on obtient une loi a priori de la forme :

$$\pi(\theta,\beta\mid\mu_1,\mu_2,\lambda)\propto\exp\{[\theta\ \beta][\mu_1\ \mu_2]'-\lambda A(\theta,\beta)]=\frac{e^{\mu_1\theta+\mu_2\beta}}{(1+e^{\theta-\beta})^{\lambda}}.$$

On remarquera que cette loi est impropre. La loi a posteriori aura la forme suivante :

$$\pi(\theta,\beta\mid x)\propto\frac{e^{(\mu_1+x)\theta+(\mu_2+x)\beta}}{(1+e^{\theta-\beta})^{\lambda+1}}.$$

1.4.3 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 le cadre paramètrique, 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. Ces distributions sont dites subjectives parce qu'elles sont propre à l'expert. Elles doivent être interprétées comme un pari de l'expert.

1.4.4 Lois a priori impropres

La loi a priori peut être impropre $\displaystyle\int_\Theta\pi(\theta)d\theta=+\infty$. Ce choix de type de loi n'a donc plus d'intérêt que calculatoire et s'interprète difficilement. Nous verrons par la suite que la construction de lois non informative peut conduire à des lois a priori de ce type.

Exemple 1.3. *Considérons la loi* $\mathbb{1}_{\mathbb{R}^+}(x)$*, loi uniforme sur* $\mathbb{R}^+$*. Supposons* $f(x|\lambda)=\lambda\exp\{-\lambda x\}$.

La loi a posteriori est : $\pi(\lambda|x)=\lambda\exp\{-\lambda x\}/\int_0^{+\infty}\lambda e^{-\lambda x}dx$.

Le dénominateur est égal à $\Gamma(2)x^2$ *d'où* $\pi(\lambda|x)=[x^2/\Gamma(2)]\lambda\exp\{-\lambda x\}$*, une loi gamma de paramètre* $(2,x)$.

1.4.5 Lois a priori non informatives

Une loi non informative est une loi qui porte une information sur le paramètre à estimer dont le poids dans l'inférence est réduit. Certains auteurs la définissent également comme une

loi a priori qui ne contient aucune information sur θ ou encore comme une loi qui ne donne pas davantage de poids à telle ou telle valeur du paramètre. Par exemple, supposons Θ un ensemble fini de taille q, une loi a priori non informative pourra être une loi de la forme :

$$P(\theta_i) = \frac{1}{q} \qquad i = \overline{1,q}$$

On a l'équiprobabilité, les valeurs possibles de θ se voit attribuer le même poids.

1.4.6 Lois a priori de Jeffreys

La règle de Jeffreys

– Une méthode proposée par Jeffreys (1961) permet de fabriquer des lois a priori non informative. Cette méthode utilise l'information de Fischer : $I(\theta)$. L'argument pourrait être le suivant. $I(\theta)$) représente une mesure de la quantité d'information sur θ contenue dans l'observation. Plus $I(\theta)$ est grande, plus l'observation apporte de l'information. Il semble alors naturel de favoriser (au sens rendre plus probable suivant $\pi(\theta)$), les valeurs de θ pour lesquels $I(\theta)$ est grande; ce qui minimise l'influence de la loi a priori au profit de l'observation.
Le choix de ce type de loi conduit ainsi souvent à des estimateurs classiques du type maximum de vraisemblance.

La règle de Jeffreys consiste donc à considérer des lois a priori de la forme :

$$\pi(\theta) = C\sqrt{I(\theta)} \text{ où } I(\theta) = E\left[-\frac{\partial^2}{\partial\theta^2}\log f(x \mid \theta)\right]. \tag{3}$$

dans le cas unidimensionnel.

Rappels sur l'information de Fisher

– Soit un n-échantillon $(X_1,...,X_n)$ de loi $f(x \mid \theta)$. On a le résultat suivant. Sous certaines conditions de régularité, l'estimateur du maximum de vraisemblance $\hat{\theta}_n$ de θ est tel que, pour n assez grand, $\sqrt{n}(\hat{\theta} - \theta)$ suit une loi normale centrée, de variance $\dfrac{1}{I(\theta)}$.

Autre résultat important : l'inégalité de Cràmer-Rao.
Soit t_n, un estimateur de $g(\theta)$, alors :

$$E(t_n - g(\theta))^2 \geq \frac{g'(\theta) + b'_n(\theta))^2}{nI(\theta)},$$

où $b_n(\theta) = E(t_n) - g(\theta)$, est le biais.

Si $g(\theta) = \theta$ et si l'estimateur est sans biais, on a :

$$Var(t_n) \geq \frac{1}{nI(\theta)}.$$

Ces résultats illustrent l'idée que l'information de Fisher se compare à une variance.

1.4.7 Lois a priori d'entropie maximum

L'entropie est une grandeur bien connue des physiciens comme étant une mesure du désordre. Dans le cadre de la statistique, elle mesure la quantité d'incertitude inhérente à la loi de probabilité.

Définition 1.3. – Soit Θ un espace de paramètres discret, soit π une probabilité sur Θ. L'entropie de π notée $\mathcal{E}_q(\pi)$ est définie par :

$$\mathcal{E}_q(\pi) = -\sum_{\Theta} \pi(\theta_i) \log \pi(\theta_i). \tag{4}$$

Convention : si $\pi(\theta_i) = 0$ alors $\pi(\theta_i) \log \pi(\theta_i) = 0$

Exemple 1.4. *On considère :* $\Theta = \{\theta_1,...,\theta_q\}$.

On pose : $\pi(\theta) = 1$ *si* $\theta = \theta_k$ *et* $\pi(\theta) = 0$ *pour tout theta* $= \theta_i, i \neq k$.

La loi a priori donne exactement la valeur de θ*; l'incertitude est donc ici nulle et* $\mathcal{E}_q(\pi) = 0$. *Si maintenant* $\pi(\theta_i) = \dfrac{1}{q}$ *pour tout i alors*

$$\mathcal{E}_q(\pi) = -\sum_{i=1}^{q} \frac{1}{q} \log(\frac{1}{q}) = \log q.$$

Ce qui correspond a l'incertitude la plus grande. En effet, on peut montrer que : $\mathcal{E}_q(\pi) \leq \log q$ *pour tout* π *non degenere.* π *est appelee la loi d'entropie maximum.*

Supposons que l'on dispose d'informations a priori concernant θ telles que l'on puisse écrire :

$$E^{\pi}[g_k(\theta)] = \sum_{i=1}^{q} g_k(\theta_i)\pi(\theta_i) = \mu_k,\ k = 1,...,m. \tag{5}$$

Une loi a priori d'entropie maximum est une solution d'un problème d'optimisation sous la contrainte (5) :

$$\bar{\pi} = \underset{\pi}{Argmax} \mathcal{E}_q(\pi)$$

Si π est propre ($\sum \pi(\theta_i) = 1$), on montre que la solution est de la forme :

$$\bar{\pi}(\theta_i) = \frac{\exp\{\sum_{k=1}^{m} \lambda_k g_k(\theta_i)\}}{\sum_{\Theta} \exp\{\sum_{k=1}^{m} \lambda_k g_k(\theta_i)\}}$$

.

Exemple 1.5. $\Theta = \mathbb{N}$. *Supposons* $E^{\pi}(\theta) = 5, m = 1$. *On a :* $g_1(\theta) = \theta$ *et* $\mu_1 = 5$.
Supposons $\lambda_1 < 0$,

$$\bar{\pi}(\theta) = \frac{e^{\lambda_1 \theta}}{\sum_{\theta=0}^{+\infty} e^{\lambda_1 \theta}} = (1 - e^{\lambda_1})(e^{\lambda_1}).$$

On reconnait une loi géométrique. On déterminé alors λ_1 *par :*

$$E^{\bar{\pi}}(\theta) = \frac{e^{\lambda_1}}{1-e^{\lambda_1}} = 5 \Longleftrightarrow \lambda_1 = \log\left(\frac{5}{6}\right)$$

$\bar{\pi}$ *est donc une loi géometrique de parametre* $\dfrac{5}{6}$.

Définition 1.4. Si Θ est continu, on peut proposer la définition suivante de l'entropie :

$$\mathcal{E}(\pi) = -E^{\pi}\left[\log\frac{\pi(\theta)}{\pi_0(\theta)}\right] = -\int_{\Theta}\pi(\theta)\log\left(\frac{\pi(\theta)}{\pi_0(\theta)}\right)d\theta$$

où $\pi_0(\theta)$ est la loi a priori non informative naturelle pour le probléme.

Comme précédemment, si on dispose d'information a priori du type :

$$E^{\pi}[g_k(\theta)] = \int_{\Theta} g_k(\theta)\pi(\theta)d\theta = \mu_k,\ k = 1,...,m.$$

la loi a priori du maximum d'entropie est alors donnée par :

$$\bar{\pi}(\theta_i) = \frac{\pi_0(\theta)\exp\{\sum_{k=1}^{m}\lambda_k g_k(\theta)\}}{\int_{\Theta}\exp\{\sum_{k=1}^{m}\lambda_k g_k(\theta)\}d\theta},$$

où λ_k sont des constantes obtenues par la contrainte.

Exemple 1.6. $\Theta = \mathbb{R}$. *Supposons que* θ *est un paramètre de position. La loi a priori naturel non informative est alors* $\pi_0(\theta) = 1$.
Si on fixe les valeurs de la moyenne et de la variance de la loi a priori à (μ,σ^2), *on a* $E^{\pi}(\theta) = \mu$ *et* $g_1(\theta) = \theta, \mu_1 = \mu$, *et* $E^{\pi}[(\theta-\mu)^2] = \sigma^2$ *et* $g_2(\theta) = (\theta-\mu)^2, \mu_2 = \sigma^2$.

La loi du maximum d'entropie est alors :

$$\bar{\pi}(\theta) = \frac{\exp\{\lambda_1\theta + \lambda_2(\theta-\mu)^2\}}{\int_{\Theta}\exp\{\lambda_1\theta + \lambda_2(\theta-\mu)^2\}d\theta}.$$

Calculons λ_1,λ_2. *On montre sans difficultés que* $\lambda_1\theta + \lambda_2(\theta-\mu)^2 \propto \lambda_2[\theta - (\mu_1\lambda_1/2\lambda_2)]^2$.

$\bar{\pi}$ *est donc une loi normale de paramètres* $[(\mu - \lambda_1/2\lambda_2); -1/2\lambda_2]$.

On cherche alors λ_1 *et* λ_2 *tels que :* $\mu - \lambda_1/2\lambda_2 = \mu$ *et* $\sigma^2 = -1/2\lambda_2$.

Il vient donc : $\lambda_1 = 0$ *et* $\lambda_2 = -1/2\sigma^2$ *d'où* $\bar{\pi}(\theta)$ *est une loi normale de paramètres* (μ,σ^2).

1.5 Les bases de la théorie de la décision

1.5.1 Coût et Décision

Le problème général auquel on s'intéresse ici est celui d'un individu plongé dans un environnement donné (**nature**) et qui, sur la base **des observations**, est conduit à mener des **actions** et à prendre des décisions qui auront un **coût**. Les espaces intervenant dans un **modèle de décision** sont :

$\mathfrak{X}$: l'espace des observations,

Θ : l'espace des états de la nature (l'espace des paramètres dans le cas d'un problème statistique),

$\mathcal{A}$: l'espace des actions ou décisions, dont les éléments sont des images de l'observation par une application δ appelée règle de décisions (une statistique (c'est-à-dire fonction des observations) dans le cas d'un problème statistique),

$\mathcal{D}$: l'ensemble des règles de décisions δ, applications de $\mathfrak{X}$ dans $\mathcal{A}$ (les estimateurs possibles). On note à une action. On a : $a = \delta(x)$.

L'inférence consiste à choisir une règle de décision $\delta \in \mathcal{D}$ concernant $\theta \in \Theta$ sur la base d'une observation $x \in \mathfrak{X}$, x et θ étant liés par la loi $f(x|\theta)$.

En statistique, la règle de décision est un estimateur, l'action est une estimation (valeur de l'estimateur au point d'observation x). Pour choisir une décision, on construit une relation de préférence en considérant une mesure du coût ou perte encourue lorsqu'on prend la décision $\delta(x)$ et que l'état de la nature est θ. Pour ce faire on introduit la fonction L, appelée **fonction de coût** (ou de perte) définie de la manière suivante :

1.5.2 Fonction de perte et risque

Définition 1.5. Fonction coût (perte)

– On appelle fonction de coût, toute fonction mesurable L de $\Theta \times \mathcal{A}$ dans $\mathbb{R}$.

$L(\theta,a)$ évalue le coût d'une décision a quand le paramètre vaut θ. Elle permet donc, en quelque sorte, de quantifier la perte encourue par une mauvaise décision, une mauvaise évaluation de θ. Il s'agit d'une fonction de θ.
Un coût négatif correspond à un gain.

1.5.2.1 Risque Fréquentiste

On dira qu'une décision est une bonne décision si elle conduit à un coût nul. Autrement dit, une bonne décision est solution de l'équation :

$$L(\theta,\delta(x)) = 0.$$

θ étant inconnu, on ne peut évidemment pas résoudre cette équation. Classer les décisions par la

seule considération du coût est donc impossible. Celui-ci ne prend pas en compte l'information apportée par le modèle $f(x|\theta)$.
Ces remarques conduisent à considérer la moyenne de la perte, c'est le risque fréquentiste.

Définition 1.6. – On appelle **risque fréquentiste** le coût moyen (l'espérance mathématique) du coût d'une règle de décision :

$$R(\theta,\delta) = E_\theta[L(\theta,\delta(x))] = \int_{\mathfrak{X}} L(\theta,\delta(x))f(x|\theta)dx.$$

On peut alors donner la définition suivante :

– On dira que la décision δ_1 est préférable à δ_2 et on note $\delta_1 \prec \delta_2$ si :

$$R(\theta,\delta_1) \leq R(\theta,\delta_2), \qquad \forall\theta \in \Theta.$$

Cette définition permet d'établir un préordre sur l'ensemble D des décisions. Cependant, ce préordre est partiel puisqu'il ne permet pas de comparer deux règles de décision telles que :

$$R(\theta_1,\delta_1) < R(\theta_1,\delta_2) et\ R(\theta_2,\delta_1) > R(\theta_2,\delta_2), \qquad \text{pour } \theta_1 \text{ et } \theta_2 \in \Theta.$$

Définition 1.7. Risque a posteriori

Une fois donnée la loi a priori sur le paramètre et la fonction de perte, le risque a posteriori est défini par : $\rho(\pi,\delta|x) = E^\pi[L(\theta,\delta)|x]$

$$= \int_\Theta L(\theta,\delta)\pi(\theta|x)d\theta$$

Définition 1.8. Risque intégré

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

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

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

1.5.3 Risque de Bayes

Puisque l'approche Bayésienne met à la disposition du statisticien une loi a priori $\pi(\theta)$, on peut considérer la moyenne du risque fréquentiste. la moyenne du coût moyen suivant la loi a priori : $E^\pi[R(\theta,\delta(X))]$. Il s'agit du **risque bayésien** ou **risque de Bayes** que l'on note $r(\pi,\delta)$).
On a :

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

On définit alors le **coût a posteriori** $\rho(\pi,\delta(x))$ comme étant la moyenne du coût par rapport à la loi a posteriori :

$$\rho(\pi,\delta(x)) = E^{\pi(.|x)}[L(\theta,\delta(x))] = \int_{\Theta} L[\theta,\delta(x)]\pi(\theta|x)d\theta$$

Il s'agit d'une fonction de x.

On a le résultat suivant :

Proposition 1.2. – *Le risque de Bayes $r(\pi,\delta)$ est la moyenne du coût a posteriori $\rho(\pi,\delta(x))$ suivant la loi marginale $f(x)$.*

Preuve. $r(\pi,\delta) = \int_{\Theta}\int_{\mathfrak{X}} L(\theta,\delta(x))f(x|\theta)\pi(\theta)dxd\theta$ *or* $f(x|\theta)\pi(\theta) = \pi(\theta|x)f(x)$

On a donc :

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

1.5.4 Estimateur de Bayes

Comme nous l'avons déjà fait remarquer, la prise d'une décision, ici le choix d'un estimateur, va engendrer un coût que l'on va quantifier à l'aide de la fonction de perte. En pratique, on cherche une décision qui minimise en moyenne la fonction de coût.

Définition 1.9. – On appelle **estimateur de Bayes** associé à un coût L et à une distribution a priori π, toute décision δ^{π} qui minimise le risque de Bayes $r(\pi,\delta)$.
On a :

$$\delta^{\pi}(x) = \arg\min_{\delta\in\mathcal{D}} r(\pi,\delta).$$

Remarquons que d'après la proposition (1.2), un estimateur peut également être défini comme étant une décision qui minimise la moyenne suivant la prédictive $f(x)$ du coût a posteriori.

Théorème 1.1. *Méthode de calcul (Judith Rousseau (2009))*

Si $\exists\delta \in D$, $r(\pi,\delta) < \infty$ et $\forall X \in \mathcal{X}$, $\delta^{\pi}(X) = \arg\min_{\delta\in\mathcal{D}} \rho(\pi,\delta|X)$, alors $\delta^{\pi}(X)$ est un estimateur Bayésien.

1.5.5 Fonctions de coût usuelles

1.5.5.1 coût quadratique

Introduit par Légende en (1805) et Gauss en (1810), ce coût est sans conteste le critère d'évaluation le plus commun. Fondant sa validité sur l'ambiguïté de la notion d'erreur dans un

contexte statistique (soit erreur de mesure, soit variation aléatoire), il a aussi donné lieu à de nombreuses critiques, la plus fréquente étant sans doute le fait que le coût quadratique

$$L(\theta,\delta) = (\theta - \delta)^2 \tag{6}$$

pénalise trop fortement les grandes erreurs.

les estimateurs de Bayes associés au coût quadratique sont les moyennes a posteriori. Cependant, notons que le coût quadratique n'est pas le seul coût à avoir cette caract éristique. Les fonctions de coût conduisant à la moyenne a posteriori comme estimateur de Bayes sont appelées fonctions de coût propres et ont été identifiées par Lindley en (1985) et Hwang et Pemantle (1994).

Proposition 1.3. *(Christian P. Robert 2006)*

L'estimateur de Bayes δ^π associé à la loi a priori π et au coût quadratique est la moyenne a posteriori

$$\delta^\pi(x) = E^\pi[\theta|x] = \frac{\int_\theta \theta f(x|\theta)\pi(\theta)d\theta}{\int_\theta f(x|\theta)\pi(\theta)d\theta} \tag{7}$$

Corollaire 1.1. *(Christian P. Robert 2006)*

Quand $\Theta \in \mathbb{R}^p$, l'estimateur de Bayes δ^π associé à π et au coût quadratique,

$$L(\theta,\delta) = (\theta - \delta)^t Q(\theta - \delta)$$

est la moyenne a posteriori, $\delta^\pi(x) = E^\pi[\theta|x]$, pour toute matrice $Q(p \times p)$ symétrique définie positive.

Le coût quadratique est particulièrement intéressant lorsque l'espace des paramètres est borné et le choix d'un coût plus subjectif est impossible.

Le tableau ci-dessus représente quelques estimateurs de Bayes du paramètre θ sous coût quadratique pour les lois a priori conjuguées des familles exponentielles usuelles.

Loi de x	Loi conjuguée	Moyenne a posteriori
Normale $\mathcal{N}(\theta,\sigma^2)$	Normale $\mathcal{N}(\mu,\tau^2)$	$\frac{\mu\sigma^2+\tau^2x}{\sigma^2+\tau^2}$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha,\beta)$	$\frac{\alpha+x}{\beta+1}$
Gamma $\mathcal{G}(\nu,\theta)$	Gamma $\mathcal{G}(\alpha,\theta)$	$\frac{\alpha+\nu}{\beta+x}$
Binomiale $B(n,\theta)$	Beta $Be(\alpha,\beta)$	$\frac{\alpha+x}{\alpha+\beta+n}$
Binomiale Négative $Nev(m,\theta)$	Bêta $Be(\alpha,\beta)$	$\frac{\alpha+n}{\alpha+\beta+x+n}$
Multinomiale $M_k(\theta_1,...,\theta_k)$	Dirchlet $D(\alpha_1,...,\alpha_k)$	$\frac{\alpha_i+x_i}{(\sum_j\alpha_j)+n}$
Normale $\mathcal{N}(\mu,1/\theta)$	Gamma $\mathcal{G}(\alpha,\beta)$	$\frac{(\alpha+1)^2}{\beta+(\mu-x)}$

Exemple 1.7. *Dans cet exemple nous avions obtenu une loi gamma inverse de paramètres* $(n+\alpha,\sum_{i=1}^n x_i+\beta)$ *comme loi a posteriori. Or l'espérance mathématique d'une loi gamma inverse de paramètres* (a,b) *est égale à* $\frac{b}{a-1}$*, on a donc qu'un estimateur de Bayes* $\delta^\pi(x)$ *du paramètre* θ*, sous l'hypothèse d'un coût quadratique, est égale à :*

$$\delta^\pi(x)=(\sum_{i=1}^n xi+\beta)/(n+\alpha-1).$$

Exemple 1.8. *Si* $X\sim\mathcal{P}(\theta)$ *et si* $\pi(\theta)$*, la loi a priori est une loi Gamma de paramètres* (α,β)*, la loi a posteriori* $\pi(\theta\mid x))$ *est une loi Gamma de paramètres* $(\alpha+x,\beta+1)$ *.*

Sous l'hypothèse d'un coût quadratique, un estimateur de Bayes $\delta^\pi(x)$ *de* θ *sera l'espérance a posteriori de* θ*. Puisque la loi a posteriori est une loi Gamma, l'espérance est le rapport des paramètres et on a :*

$$\delta^\pi(x)=\frac{(\alpha+x)}{(\beta+1)}.$$

1.5.5.2 Coût absolu

Définition 1.10. – La fonction de coût absolue est la fonction définie par :

$$L(\theta,\delta(x))=\begin{cases}\mathrm{k}_2(\theta-\delta(x)) & \text{si } \theta>\delta(x),\ k_1,k_2>0\\ \mathrm{k}_1(\delta(x)-\theta) & \text{sinon .}\end{cases}$$

Proposition 1.4. *(Christian P. Robert 2006)*

– Un estimateur de Bayes associé à π *et au coût absolu, est un fractile d'ordre* $k_2/(k_1+k_2)$ *de* $\pi(\theta|x)$*.*

En particulier, si $k_1=k_2=1$*, dans le cas du coût absolu, l'estimateur de Bayes est la médiane a posteriori, qui est l'estimateur obtenu par Laplace.*

Preuve. $$E^{\pi(.|x)}[L(\theta,\delta(x))] = \int_\Theta L(\xi-\delta(x))\pi(\xi|x)d\xi$$
$$= \int_{\delta(x)}^{+\infty} k_2(\xi-\delta(x))\pi(\xi|x)d\xi + \int_{-\infty}^{\delta(x)} k_1(\delta(x)-\xi)\pi(\xi \mid x)d\xi$$

On remarque que :

$$\pi(\xi \mid x)d\xi = dF(\xi \mid x) = -d(1-F(\xi \mid x)) \text{ et on écrit donc :}$$
$$E^{\pi(.|x)}[L(\theta,\delta(x))] = [k_2(\xi-\delta(x))(1-F(\xi \mid x))]_{\delta(x)}^{+\infty} + \int_{\delta(x)}^{+\infty} k_2 P^{\pi(.|x)}(\theta) > \xi)d\xi$$
$$+[k_1(\delta(x)-\xi)F(\xi \mid x)]_{-\infty}^{\delta(x)} + \int_{-\infty}^{\delta(x)} k_1 P^{\pi(.|x)}(\theta > \xi)d\xi$$
$$= \int_{\delta(x)}^{+\infty} k_2 P^{\pi(.|x)}(\theta < \xi)d\xi$$

On dérive par rapport à $\delta(x)$, il vient :

$$\frac{\partial}{\partial\delta(x)}\rho(\pi,\delta(x)) = -k_2 P^{\pi(.|x)}(\theta > \delta(x)) + k_1 P^{\pi(.|x)}(\theta < \xi) = 0$$
$$\Leftrightarrow -k_2(1 - P^{\pi(.|x)}(\theta > \delta(x))) + k_1 P^{\pi(.|x)}(\theta < \xi) = 0$$
$$\Leftrightarrow (k_1+k_2)P^{\pi(.|x)}(\theta < \delta(x)) - k_2 = 0$$

d'où

$$P^{\pi(.|x)}(\theta < \delta(x)) = \frac{k_2}{k_1+k_2}$$

et le coût est donc maximisé pour $\tilde{\theta} = \delta(x)$ tel que $P^{\pi(.|x)}(\theta < \delta(x)) = \frac{k_2}{k_1+k_2}$.

1.5.5.3 Coût 0 - 1

Définition 1.11. – On appelle coût $0-1$, l'application L définie par :

$$L(\theta,\delta(x)) = \begin{cases} 0 & \text{si la décision est bonne} \\ 1 & \text{sinon} \end{cases}$$

On retrouve en utilisant cette fonction de coût, les résultats de la théorie des tests d'hypothèses. Un problème de test est un problème de choix (de prise de décision) entre $H_0 : \theta \in \Theta_0$ et $H_1 : \theta \in \Theta_1$

On définit donc la décision de la manière suivante :

$\delta(x) = 1$: on accepte H_0;

$\delta(x) = 0$: on rejette H_0.(n.b. : ceci ne dépend pas de θ).

On a un espace d'actions de la forme : $A = \{0,1\}$.

Soit W la région de rejet c'est-à-dire le sous-ensemble de $\mathcal{X}$ qui conduit à rejeter H_0. On peut construire une fonction de coût de la manière suivante : supposons $\theta \in \Theta_0$,

si $X \in W$, on prend la décision de rejeter c'est-à-dire $\delta(X) = 0$, mais la décision n'est pas bonne, on va pénaliser et $L(\theta,\delta(x)) = 1$,

si $X \notin W$, on ne rejette pas, on prend la décision $\delta(X) = 1$, la décision est bonne $L(\theta,\delta(X)) = 0$.

Le coût s'écrit donc :

$$L(\theta,\delta(x)) = \begin{cases} 1 - \delta(x) & \text{si } \theta \in \Theta_0 \\ \delta(x) & \text{sinon.} \end{cases}$$

Ce qu'on peut écrire : $L(\theta,\delta(x)) = \mathbb{1}(x \in W)$ et on calcule la fonction de risque :

$$R(\theta,\delta) = E[L(\theta,\delta(x))] = \int_{\mathfrak{X}} L(\theta,\delta(x))dp_\theta(x) = P_\theta(x \in W), \theta \in \Theta_0.$$

On retrouve le risque de première espèce.

Proposition 1.5. *(Christian P. Robert 2006)*

L'estimateur de Bayes associé à π et au coût $0-1$ est

$$\delta^\pi(x) = \begin{cases} 1 & si\ P(\theta \in \Theta_0|x) > P(\theta \notin \Theta_0|x) \\ 0 & sinon. \end{cases}$$

donc $\delta^\pi(x)$ vaut 1 si et seulement si $P(\theta \in \Theta_0|x) > 1/2$.

1.5.6 Admissibilité et minimaxité

Définition 1.12. Estimateur randomisé

Pour un ensemble de décisions D, on définit D^* comme l'ensemble des probabilités sur D. L'estimateur $\delta^* \in D^*$ est appelé estimateur randomisé.

Cette extension est nécessaire au traitement des notions d'admissibilité et de minimaxité. L'ensemble D^* apparait comme complétion topologique de D.
Cependant cette modification de l'espace de décision ne modifie pas les réponses Bayésiennes, comme le montre le résultat suivant.

Théorème 1.2. *(Christian P. Robert 2006)*

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

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

1.5.6.1 Minimaxité

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

Définition 1.13. On appelle risque minimax associé à la fonction de coût L la valeur :

$$\bar{R} = \inf_{\delta \in D^*} \sup_{\theta} R(\theta,\delta) = \inf_{\delta \in D^*} \sup_{\theta} \mathbb{E}_\theta[L(\theta,\delta(x))] \tag{8}$$

et estimateur minimax tout estimateur δ_0 tel que

$$\bar{R} = \sup_{\theta} R(\theta,\delta_0) \tag{9}$$

L'estimateur minimax correspond au point de vue de faire le mieux dans le pire des cas, c'est-à-dire à s'assurer contre le pire. Il est utile dans des cadres complexes mais trop conservateur dans certains cas où le pire est très peu probable. Il peut être judicieux de voir l'estimation comme un jeu entre le statisticien (choix de δ) et la Nature (choix de θ), l'estimation minimax rejoint alors celle de la Théorie des Jeux.

Régle minimax et stratégie maximin

Une difficulté importante liée à la notion de minimaxité est que les estimateurs minimax n'existent pas nécessairement. En particulier, il existe une stratégie minimax quand Θ est fini et la fonction de coût est continue. Plus généralement, Brown (1976) (voir aussi Le Cam, 1986, et Strasser, 1985) considère l'espace de décision D comme plongé dans un autre espace de manière telle que l'ensemble des fonctions de risque sur D est compact dans ce grand espace. Dans cette perspective et sous des hypothèses supplémentaires, il est alors possible de construire des estimateurs minimax lorsque la fonction de coût est continue.

Théorème 1.3. *(Christian P. Robert 2006)*

Si $D \subset \mathbb{R}^k$ est convexe et compact et si $L(\theta,d)$ est continue et convexe en tant que fonction de d, pour chaque $\theta \in \Theta$, alors, il existe un estimateur minimax non randomisé.

Lemme 1.1. *(Judith Rousseau (2009)*

Le risque de Bayes est toujours plus petit que le risque minimax,

$$\underline{R} = \sup_{\pi} r(\pi) = \sup_{\pi} \inf_{\delta \in D} r(\pi,\delta) \leq \overline{R} = \inf_{\delta \in D^*} \sup_{\theta} R(\theta,\delta)$$

La première valeur est dite risque maximin et une distribution π^* telle que $r(\pi^*) = \underline{R}$ est appelée distribution a priori la moins favorable, quand de telles distributions existent. En général, la borne supérieure $r(\pi^*)$ est atteinte plutôt par une distribution impropre pouvant s'exprimer comme une limite de distributions a priori propres $\prod_n$. Mais ce phénomène n'empêche pas nécessairement la construction d'estimateurs minimax. Quand elles existent, les distributions les moins favorables sont celles qui ont le risque de Bayes le plus grand, donc aussi les moins intéressantes en terme de coût lorsqu'elles ne sont pas suggérées par l'information a priori disponible.

Le résultat ci-dessus est assez logique au sens où l'information a priori ne peut qu'améliorer l'erreur d'estimation, même dans le pire des cas.

Un cas particulier interessant correspond à la définition suivante :

Définition 1.14. Un problème d'estimation est dit admettre une valeur si $\underline{R} = \overline{R}$, c'est-à-dire quand

$$\sup_{\pi} \inf_{\delta \in D} r(\pi,\delta) = \inf_{\delta \in D^*} \sup_{\theta} R(\theta,\delta)$$

Quand le problème admet une valeur, certains estimateurs minimax sont des estimateurs de Bayes correspondant aux lois a priori les moins favorables. Cependant, ils peuvent être randomisés. Par conséquent le principe minimax ne fournit pas toujours des estimateurs acceptables.

D'un point de vue pratique, le lemme suivant fournit des conditions suffisantes de minimaxité.

Lemme 1.2. *(Christian P. Robert 2006)*

Si δ_0 est un estimateur de Bayes pour π_0 et si $R(\theta,\delta_0) \leq r(\pi_0)$ pour tout θ dans le support de π_0, δ_0 est minimax et π_0 est la distribution la moins favorable.

1.5.6.2 Admissibilité

Le critère d'admissibilité induit un ordre partiel sur D^* en comparant les risque fréquentistes des estimateurs $R(\theta,\delta)$.

Définition 1.15. Estimateur admissible

Soit un modèle paramètrique et une fonction de perte L sur $\Theta \times D$ où D est l'ensemble des décisions. On dit que $\delta \in \Theta$ est inadmissible s'il existe un estimateur δ_0 qui domine δ, c'est-a-dire tel que pour tout θ,

$$R(\theta,\delta) \geq R(\theta,\delta_0)$$

et pour au moins une valeur θ_0 du paramètre

$$R(\theta_0,\delta) > R(\theta_0,\delta_0)$$

Dans le cas contraire, δ est admissible.

Proposition 1.6. *(Christian P. Robert 2006)*

S'il existe un unique estimateur minimax, cet estimateur est admissible.

Notons que la réciproque de ce résultat est fausse, car il peut exister plusieurs estimateurs minimax admissibles. Par exemple, dans le cas $N_p(\theta,I_p)$, il existe des estimateurs de Bayes réguliers minimax pour $p \geq 5$. Quand la fonction de coût L est absolument convexe (en d), la caractérisation suivante est aussi possible.

Proposition 1.7. *(Judith Rousseau (2009))*

Si δ_0 est admissible de risque constant, δ_0 est l'unique estimateur minimax.

Théorème 1.4. *Estimateurs Bayésiens admissibles, (Judith Rousseau (2009)*

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 δ^π estimateur bayésien non admissible : $\mathcal{D},\forall\theta \in \Theta,R(\theta,\delta) \geq R(\theta,\delta_0)$ et $\exists\theta_0 \in \Theta,R(\theta_0,\delta) > R(\theta_0,\delta_0)$. En intégrant la première inégalité :

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

donc δ_0 est aussi un estimateur bayésien associé à L et π et $\delta_0 \neq \delta^\pi$ d'après la seconde inégalité. Le théorème se déduit par contraposée. Ce théorème s'applique notamment dans le cas d'un risque fini et d'une fonction de coût convexe. En outre, l'unicité de l'estimateur bayésien implique la finitude du risque :
$r(\pi) = \int_\Theta R(\theta,\delta^\pi)d\pi(\theta) < \infty$ (sinon, tout estimateur minimise le risque).

Proposition 1.8. *(Christian P. Robert 2006)*

Si un estimateur de Bayes, δ^π, associé à une loi a priori (propre ou impropre) π, est tel que le risque de Bayes,

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

soit fini, alors δ^π est admissible.

Définition 1.16. π-admissibilité

Un estimateur δ_0 est π-admissible si et seulement si

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

Propriété 1.1. (Christian P. Robert 2006)

Tout estimateur Bayésien tel que $r(\pi) < \infty$ est π-admissible.

Démonstration:

Soit δ^π un estimateur bayésien à risque fini. Pour δ_0 tel que $\forall\theta R(\theta,\delta) \leq R(\theta,\delta_0)$, on note $A = \{\theta \in \Theta, R(\theta,\delta) < R(\theta,\delta_0)\}$. Nous avons alors :

$$\int_\Theta R(\theta,\delta_0)d\pi(\theta) - \int_\Theta R(\theta,\delta^\pi)d\pi(\theta) = \int_A (R(\theta,\delta) - R(\theta,\delta_0))d\pi(\theta) \leq 0$$

si et seulement si $\pi(A) = 0$. Or, comme δ^πest bayésien et le risque fini, $r(\theta,\delta_0) \geq r(\theta,\delta^\pi)$ donc l'intégrale est nulle (positive et négative !), d'où $\pi(A) = 0$: δ^π est π-admissible.

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

Théorème 1.5. *Continuité et π-admissibilité, (Judith Rousseau (2009)*

Si $\pi > 0$ sur Θ, $r(\pi) < \infty$ pour une fonction de perte L donnée, si δ^π est l'estimateur Bayésien correspondant existe et si $\theta \mapsto R(\theta,\delta)$ est continu, alors δ^π est admissible.

1.5.7 Estimateur du maximum a posteriori MAP

Définition 1.17. On appelle estimateur MAP 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

$$\delta^\pi(x) \in \arg\max_\theta \pi(\theta|x)$$

Cet estimateur naturel peut s'exprimer comme un estimateur du maximum de vraisemblance pénalisée au sens classique, il a le grand avantage de ne pas dépendre d'une fonction de perte et est utile pour les approches théoriques. Ses inconvénients sont les mêmes que l'estimateur du maximum de vraisemblance : non unicité, instabilité (dus aux calculs d'optimisation) et la dépendance vis-à-vis de la mesure de référence (dominant Θ). En outre, il ne vérifie pas la non invariance par reparamétrisation qui peut apparaître importante intuitivement.

1.5.8 Tests et intervalles de crédibilité

D'un point de vue statistique, un test soit au sens Bayésien ou classique, peut être considéré comme une des deux approches :
-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 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, 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 calculé la loi a posteriori.

1.5.8.1 Intervalles de crédibilité

Définition 1.18. Région α-crédible

Pour $0 < \alpha < 1$, une région α-crédible de $100(1-\alpha)\%$ pour θ est un sous-ensemble $C \in \Theta$ tel que $P^{\pi}\{\theta \in C |\ X = x\} = 1 - \alpha$ habituellement C est un intervalle. Il existe une infinité de région α-crédibles, il est logique de s'intéresser donc à celle qui a un volume minimal. Pour cela, nous allons introduire la notion d'une région HPD (Highest Posterior Density).

Définition 1.19. Région HPD

Une région HPD est la région C_{α}^{π} définie par

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

où $h_{\alpha} = \sup\{h; P^{\pi}(\{\theta, \pi(\theta \mid x) \geq h\} \mid x) \geq 1 - \alpha\}$

Les régions HPD peuvent êtres calculées numériquement ou approximativement comme elles peuvent êtres calculées par des méthodes de simulation.

1.5.8.2 Le facteur de Bayes

Définition 1.20. Le facteur de Bayes est le rapport des probabilités a posteriori des hypothèses nulle et alternative sur le rapport des probabilités a priori de ces mêmes hypothèses, soit :

$$B_F = \frac{P(\theta \in \Theta_0 \mid x)}{P(\theta \in \Theta_1 \mid x)} \Big/ \frac{\pi(\theta \in \Theta_0)}{\pi(\theta \in \Theta_1)}$$

Ce rapport est une analogue Bayésienne du rapport de vraisemblances des tests classiques, il évalue la modification de la vraisemblance de l'ensemble Θ_1 par rapport à celle de l'ensemble Θ_1 due à l'observation et peut se comparer naturellement à 1.

En général, le facteur de Bayes dépend de l'information a priori, mais il est souvent proposé comme réponse Bayesienne "objective", car il élimine partiellement l'influence du modèle a priori et souligne le rôle des observations. De ce fait, il peut être perçu comme un rapport de vraisemblance Bayésien, car si π_0 est la loi a priori sous H_0, et π_1 la loi a priori sous H_1, B_F peut s'écrire :

$$B_F = \frac{\int_{\Theta_0} f(x \mid \theta_0)\pi_0(\theta)d\theta}{\int_{\Theta_1} f(x \mid \theta_1)\pi_1(\theta)d\theta}$$

1.6 Conclusion

Nous renvoyons à Robert (2006) pour des exemples convaincants dans le cadre des modèles linéaires généralisés, des modèles de capture-recapture, des modèles de mélange, des séries temporelles... Il existe dans chaque cas une modélisation a priori par défaut et une résolution algorithme qui permettent de fournir une solution bayésienne de référence pour le problème considéré. Bien entendu d'autres lois a priori peuvent être considérées, le modéle de référence servant alors à évaluer l'impact de ce choix a priori. Nous voulions simplement communiquer ici l'idée selon laquelle il est possible de mener une inférence bayésienne sur un probléme réaliste sans disposer d'une expertise particulière pour la construction de lois a priori.

Chapitre 2

2 Méthodes de Monte-Carlo par Chaîne de Markov

2.1 Introduction

Les méthodes de Monte-Carlo par chaînes de Markov, ou méthodes MCMC, sont une classe de méthodes d'échantillonnage à partir de distributions de probabilité. Ces méthodes de Monte-Carlo se basent sur le parcours de chaînes de Markov qui ont pour lois stationnaires les distributions à échantillonner.

Certaines méthodes utilisent des marches aléatoires sur les chaînes de Markov (algorithme de Metropolis-Hastings, échantillonnage de Gibbs), alors que d'autres algorithmes, plus complexes, introduisent des contraintes sur les parcours pour essayer d'accélérer la convergence.

Ces méthodes sont notamment appliquées dans le cadre de l'inférence bayésienne.

2.2 Méthodes de Monte-Carlo

Le terme méthode de Monte-Carlo, désigne une famille de méthodes algorithmiques visant à calculer une valeur numérique approchée en utilisant des procédés aléatoires, c'est-à-dire des techniques probabilistes. Le nom de ces méthodes, qui fait allusion aux jeux de hasard pratiqués à Monte-Carlo, a été inventé en (1947) par Nicholas Metropolis, et publié pour la première fois en (1949) dans un article coécrit avec Stanislaw Ulam.

Les méthodes de Monte-Carlo sont particulièrement utilisées pour calculer des intégrales en dimensions plus grandes que 1 (en particulier, pour calculer des surfaces et des volumes). Elles sont également couramment utilisées en physique des particules, où des simulations probabilistes permettent d'estimer la forme d'un signal ou la sensibilité d'un détecteur. La comparaison des données mesurées à ces simulations peut permettre de mettre en évidence des caractéristiques inattendues, par exemple de nouvelles particules.

2.2.1 Description de la méthode

Supposons que l'on veuille calculer une quantité I. La première étape est de la mettre sous forme d'une espérance $I = \mathbb{E}(X)$ avec X une variable aléatoire. Si on sait simuler des variables X_1,X_2,... indépendantes et identiquement distribuées, alors nous pouvons approcher I par

$$I \approx \frac{X_1 + X_2 + ... + X_N}{N} \tag{10}$$

avec N « grand », sous réserve d'application de la loi des grands nombres. C'est ce type d'approximation que l'on appelle « méthode de Monte-Carlo ».

Exemple 2.1. *Supposons que l'on cherche à calculer*

$$I = \int_{[0;1]} f(u_1,...,u_d)du_1...du_d.$$

Nous posons $X = f(U_1,...,U_d)$ *où les* $U_1,...U_d$, *sont des variables aléatoires indépendantes suivant toutes une loi uniforme sur [0,1]. Alors :*

$$I = \mathbb{E}(f(U_1,...,U_d)) = \mathbb{E}(X).$$

Nous avons donc réalisé la première étape. Tout logiciel de programmation nous permet de simuler des variables uniformes sur [0,1]. De plus, des appels successifs de variables uniformes renvoient des variables indépendantes. Nous pouvons donc facilement simuler $U_1^{(1)},...,U_d^{(1)},U_1^{(2)},...,U_d^{(2)},...$ des variables i.i.d. de loi uniforme sur [0,1] (nous noterons cette loi $U([0,1])$) et donc $X_1 = (U_1^{(1)},...,U_d^{(1)}),X_2 = (U_1^{(2)},...,U_d^{(2)},...)$ sont i.i.d. Le programme 2.1 fournit un exemple de programme approchant I par une méthode de Monte-Carlo (avec $d = 3$, $n = 1000$ itérations, $f(u_1;u_2;u_3) = \sin(u_1)\sin(u_2)\sin(u_3)$).

Programme 2.1 approximation de Monte-Carlo

```
d=3
n=1000
s=0
for (i in 1:n)
{
u=runif(d,0,1) \# simulation de d variables uniformes indépendantes
s=s+sin(u[1])*\sin(u[2])*\sin(u[d])
}
print(s/n)
```

Si nous cherchons à évaluer une intégrale de la forme

$$I = \int_{\mathbb{R}^d} Q(x)f(x)dx$$

avec f une densité de probabilité (c'est à dire f positive et $\int_{\mathbb{R}^d} f(x)dx = 1$), alors nous pouvons écrire $I = \mathbb{E}(Q(X))$ avec X de loi de densité f. Nous sommes encore dans la situation où nous voulons calculer l'espérance d'une variable aléatoire (si X est une variable aléatoire, alors $Q(X)$ est une variable aléatoire, à condition que Q soit mesurable).

2.2.2 Intégration Monte-Carlo

Une application classique des méthodes Monte-Carlo est le calcul de quantités du type

$$I = \mathbb{E}[\varphi(X)] = \int \varphi(x)f(x)dx, \tag{11}$$

où $\varphi : \mathbb{R}^d \to \mathbb{R}$ est une fonction donnée et X un vecteur aléatoire de densité f suivant laquelle on sait simuler. Dans ce contexte, l'estimateur Monte-Carlo de base est défini par

$$\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n} \varphi(X_i), \tag{12}$$

où les X_i sont générées de facon i.i.d. selon f. Outre les propriétés de cet estimateur, ce chapitre explique comment on peut éventuellement améliorer sa précision grâce a des techniques de réduction de variance.

2.2.2.1 Approximation d'une intégrale la par méthode de Monte Carlo

(sur l'intervalle[a,b])

Considérons les fonctions f et Q définies sur [a,b].
L'intégrale suivante

$$I = \int_a^b Q(x)dx$$

peut s'écrire comme ça

$$\begin{aligned} I &= \int_a^b Q(x)dx = \int_a^b Q(x)\frac{(b-a)}{(b-a)}dx \\ &= (b-a)\int_a^b \frac{Q(x)}{b-a}dx \\ I &= (b-a)\int_a^b Q(x)f(x)dx \end{aligned}$$

f est la densité de la loi uniforme sur [a,b],

$$\boxed{\text{I = (b - a)E[Q(X)].}}$$

Exemple 2.2. *soit l'intégrale I donnée par:*

$$\begin{aligned} I &= \int_0^{\frac{\pi}{4}} \frac{1}{\cos x}dx \\ &= \frac{\pi}{4}\int_0^{\frac{\pi}{4}} \frac{1}{\cos x}\ \frac{1}{\frac{\pi}{4}}dx \end{aligned}$$

donc:

$$I = \frac{\pi}{4}\mathbb{E}\left[\frac{1}{\cos X}\right] \approx \frac{\pi}{4}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\cos x_i}\right].$$

2.2.3 Convergence et intervalles de confiance

Il est clair que l'estimateur $\hat{I}_n$ est sans biais, c'est-à-dire que $\mathbb{E}[\hat{I}_n] = I$.
Mieux, la loi forte des grands nombres assure qu'il est convergent.

2.2.3.1 Théorèmes de convergence

1) Convergence

Proposition 2.1. *(Loi forte des grands nombres)*

Soit $(X_n)_{n\geq 0}$ *une suite de variables aléatoires i.i.d., à valeurs dans* $\mathbb{R}^d (d \in \mathbb{N}^*)$.
On suppose que $\mathbb{E}[\varphi(X)] < \infty$, *alors*

$$\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n} \varphi(X_i) \xrightarrow[n\to+\infty]{p.s} I.$$

Rappel : « p.s. » signifie « presque sûrement ». Ce théorème nous dit pourquoi l'approximation de Monte-Carlo Programme(2.1) est valide (et sous quelles hypothèses).

Exemple 2.3. ***estimation de*** π.

Supposons que (X,Y) *suive la loi uniforme sur le carré* $C = [0,1] \times [0,1]$
et que $\varphi(x,y) = \mathbf{1}_{\{x^2+y^2\leq 1\}}$.
En notant $D = \{(x,y) \in \mathbb{R}_+^2,\ x^2 + y^2 \leq 1\}$ *le quart de disque unité, on a donc*

$$I = \int\int_C \mathbf{1}_D(x,y)dxdy = \lambda(D) = \frac{\pi}{4}.$$

Simuler des points (X_i,Y_i) *uniformément dans* C *est très facile et la propriété précédente assure donc que*

$$\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_D(X_i,Y_i) \xrightarrow[n\to+\infty]{p.s} \frac{\pi}{4} \iff 4 \times \hat{I}_n \xrightarrow[n\to+\infty]{p.s} \pi.$$

On dispose donc d'un estimateur Monte-Carlo pour la constante π. Encore faut-il connaître sa precision : c'est tout l'intérêt du théorème central limite.

2) Vitesse de convergence

Proposition 2.2. ***(Théorème central limite)***

Si $\mathbb{E}[\varphi(X)^2] < \infty$, *alors*

$$\sqrt{n}(\hat{I}_n - I) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,\sigma^2),$$

avec $\sigma^2 = Var(\varphi(X)) = \mathbb{E}[\varphi^2(X)] - \mathbb{E}[\varphi(X)]^2 = \int \varphi^2(x)f(x)dx - I^2$.

Ainsi, lorsque n est grand, notre estimateur suit à peu près une loi normale : avec un abus de notation flagrant, on a $\hat{I}_n \approx \mathcal{N}(I,\sigma^2/n)$, c'est-à-dire que $\hat{I}_n$ tend vers I avec une vitesse en $\mathcal{O}(1/\sqrt{n})$. Plus précisément, le TCL doit permettre de construire des intervalles de confiance. Cependant, en général, l'écart-type σ est lui aussi inconnu, il va donc falloir l'estimer. Qu'à cela ne tienne, la méthode Monte-Carlo en fournit justement un estimateur à peu de frais puisque

basé sur le même échantillon $(X_1,...,X_n)$, à savoir .

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} \varphi^2(X_i) - \hat{I}_n^2 \xrightarrow[n\to+\infty]{p.s} \sigma^2 \tag{13}$$

par la loi des grands nombres pour le premier terme et la Proposition (2.1) pour le second. Le lemme de Slutsky implique donc que

$$\sqrt{n}\frac{\hat{I}_n - I}{\hat{\sigma}_n} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Proposition 2.3. ***(Intervalles de confiance)***

Soit $\alpha \in (0,1)$ fixé. Un intervalle de confiance de niveau asymptotique $1-\alpha$ pour I est

$$\left[\hat{I}_n - \Phi^{-1}(1-\alpha/2)\frac{\hat{\sigma}_n}{\sqrt{n}}; \hat{I}_n + \Phi^{-1}(1-\alpha/2)\frac{\hat{\sigma}_n}{\sqrt{n}}\right],$$

où $\Phi^{-1}(1-\alpha/2)$ désigne le quantile d'ordre $1-\alpha/2$ de la loi normale centrée réduite.

Typiquement $\alpha = 0.05$ donne $\Phi^{-1}(1-\alpha/2) = q_{0.975} = 1.96 \approx 2$, qui permet de construire un intervalle de confiance à 95% pour I. Un point de précision en passant : prendre 1.96 plutôt que 2 pour le quantile de la loi normale est tout à fait illusoire si le terme $\hat{\sigma}_n/\sqrt{n}$ n'est pas déjà lui-même de l'ordre de 0.01 (c'est-à-dire, pour un écart-type unité, un échantillon de taille au moins 10000).

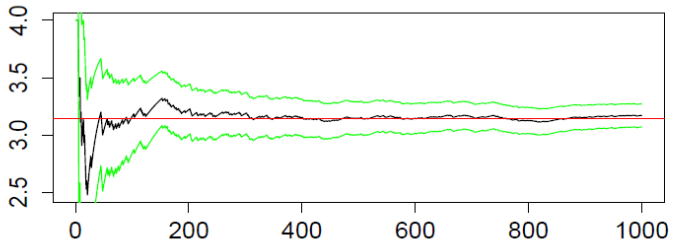


FIG. 1 – *Estimateur Monte-Carlo de π et intervalles de confiance asymptotiques.*

2.2.3.2 Détermination de la valeur de π

Soit un point M de coordonnées (x,y), ou $0 < x < 1$ et $0 < y < 1$. on tire aléatoirement les valeurs de x et y entre 0 et 1 suivant une loi uniforme. Le point M appartient au disque de centre (0,0) de rayon **R=1** si et seulement si $x^2 + y^2 \leq 1$. La probabilité que le point M appartienne au disque est $\frac{\pi}{4}$ puisque le quart de disque est de surface $\sigma = \frac{\pi R^2}{4} = \frac{\pi}{4}$ et le carré qui le contient est de surface $S = R^2 = 1$: la probabilité qu'un point soit tiré dans le quart de disque ou en dehors étant la même, la probabilité de tomber dans le quart de disque vaut $\frac{\sigma}{S} = \frac{\pi}{4}$.

En faisant le rapport du nombre de points dans le disque au nombre de tirages, on obtient une approximation du nombre $\frac{\pi}{4}$ si le nombre de tirages est grand.

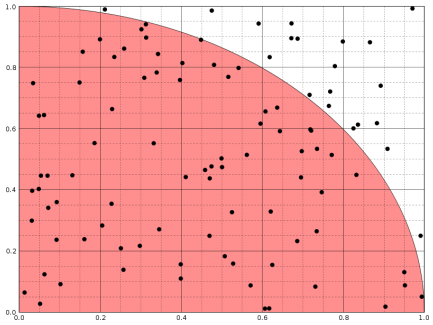


FIG. 2 – *Détermination de la valeur de π*

2.2.3.3 Détermination de la superficie d'un lac

Cet exemple est un classique en vulgarisation de la méthode de Monte-Carlo. Soit une zone rectangulaire ou carrée dont les côtés sont de longueur connue. Au sein de cette aire se trouve un lac dont la superficie est inconnue. Grâce aux mesures des côtés de la zone, on connaît l'aire du rectangle. Pour trouver l'aire du lac, on demande à une armée de tirer X coups de canon de manière aléatoire sur cette zone. On compte ensuite le nombre N de boulets qui sont restés sur le terrain, on peut ainsi déterminer le nombre de boulets qui sont tombés dans le lac :$X - N$. Il suffit ensuite d'établir un rapport entre les valeurs :

$$\frac{superficie_{terrain}}{superficie_{lac}} = \frac{X}{X-N}$$

$$\Longrightarrow \qquad \text{superficie}_{lac} = \frac{X-N}{X} \times superficie_{terrain}$$

Par exemple, si le terrain fait $1000m^2$, que l'armée tire 500 boulets et que 100 projectiles sont tombés dans le lac, alors une estimation de la superficie du plan d'eau est de : $1000 \times 100 \div 500 = 200$.

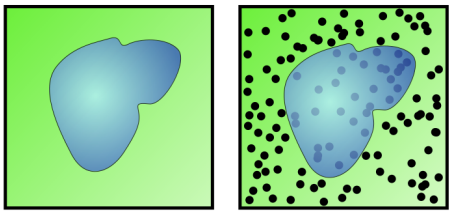


FIG. 3 – *Détermination de la superficie d'un lac*

La qualité de l'estimation s'améliore (lentement) en augmentant le nombre de tirs et en s'assurant que les artilleurs ne visent pas toujours le même endroit mais couvrent bien la zone, de manière uniforme. Cette dernière remarque est à mettre en parallèle avec la qualité du générateur aléatoire qui est primordiale pour avoir de bons résultats dans la méthode de Monte-Carlo. Un générateur biaisé est comme un canon qui tire toujours au même endroit : les informations qu'il apporte sont réduites.

2.2.4 Réduction de variance

Nous restons dans le cadre de la section précédente, à savoir l'estimation de l'intégrale $I = \mathbb{E}[\varphi(x)]$.
Son estimation $\hat{I}_n$ par la méthode Monte-Carlo standard aboutit à une erreur $\sigma/\sqrt{n}$. Dès que la dimension de X est grande ou φ irrégulière, les méthodes déterministes ou quasi-Monte-Carlo ne sont plus concurrentielles et la vitesse en $1/\sqrt{n}$ apparaît donc incompressible. L'idée est alors de gagner sur le facteur σ, ce qui est précisément l'objet des méthodes de réduction de variance.

Une remarque élémentaire au préalable : pour une précision ε voulue sur le résultat, donc en $\sigma/\sqrt{n}$ par Monte-Carlo classique, une méthode de réduction de variance permettant de diviser la variance σ^2 par 2 permet de diviser le nombre n de simulations nécessaires par 2 pour atteindre la même précision. Si la nouvelle méthode n'est pas plus coûteuse en temps de calcul, c'est donc la durée de la simulation qui est ainsi divisée par deux. Si la nouvelle méthode requiert beaucoup plus de calculs, il faut en toute rigueur en tenir compte, par exemple en définissant l'efficacité d'une technique par

$$\text{Efficacité} = \frac{1}{\text{Complexité} \times \text{Variance}}.$$

et en comparant les effcacités des différentes méthodes entre elles.

2.2.4.1 Echantillonnage préférentiel

On souhaite estimer

$$I = \mathbb{E}[\varphi(X)] = \int \varphi(x) f(x) dx.$$

Si la fonction φ prend ses plus grandes valeurs là où la densité f est très faible, c'est à dire là où X a très peu de chances de tomber, et réciproquement, l'estimateur Monte-Carlo classique

$$\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n} \varphi(X_i)$$

est très mauvais puisqu'à moins de prendre n très grand, il va estimer environ 0 même si I vaut 1.

Exemple 2.4. *1. Evénements rares : la variable X suivant une loi normale centrée réduite, on veut estimer la probabilité*

$$\mathbb{P}(X > 6) = \mathbb{E}[\mathbf{1}_{X>6}] = \int_{\mathbb{R}} \mathbf{1}_{x>6} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx.$$

La commande 1-pnorm(6) montre qu'elle est de l'ordre de 10^{-9}. Ceci signifie qu'à moins de prendre n de l'ordre d'au moins un milliard, on a toutes les chances d'obtenir $\hat{I}_n = 0$. 2. Soit m un réel, $X \sim \mathcal{N}(m,1)$ et $\varphi(x) = \exp(-mx + m^2/2)$. Pour tout m, on a donc

$$I = \mathbb{E}[\varphi(X)] = \int_{\mathbb{R}} \varphi(x) f(x) dx = \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 1.$$

Or 95% des X tombent dans l'intervalle $[m-2,m+2]$ tandis que $\varphi(m) = \exp(-m^2/2)$ tend très vite vers 0 quand m augmente. Ainsi, dès lors que m est grand, on obtient proche de 0, ce qui n'est pas une très bonne approximation de la valeur cherchée $I = 1$...

2.2.4.2 Description de la méthode

L'idée de l'échantillonnage préférentiel, ou échantillonnage pondéré, ou importance sampling, est de tirer des points non pas suivant la densité f de X mais selon une densité auxiliaire g réalisant un compromis entre les régions de l'espace où φ est grande et où la densité f est élevée, quitte à rétablir le tir ensuite en tenant compte du fait que la loi de simulation g n'est pas la loi initiale f. Mathématiquement, il s'agit simplement d'une réécriture de I sous la forme

$$I = \mathbb{E}[\varphi(X)] = \int \varphi(x) f(x) dx = \int \frac{f(y)}{g(y)} \varphi(y) g(y) dy = \int \omega(y)\varphi(y) g(y) dy = \mathbb{E}[\omega(Y)\varphi(Y)],$$

où Y a pour densité g et $\omega(y) = f(y)/g(y)$ correspond à la pondération (ou rapport de vraisemblance) dû au changement de loi. Pour que celui-ci soit bien défini, il faut s'assurer qu'on ne divise pas par 0, c'est-à-dire que $g(y) = 0$ implique $f(y)\varphi(y) = 0$. On peut alors voir $\omega(y)$ comme la dérivée de Radon-Nikodym de la loi de X par rapport à celle de Y. Ceci supposé, si l'on sait :

(a) simuler suivant la densité g,

(b) calculer le rapport de vraisemblance $\omega(y) = f(y)/g(y)$ pour tout y,

l'estimateur par échantillonnage préférentiel prend la forme

$$\tilde{I}_n = \frac{1}{n}\sum_{i=1}^{n} \omega(Y_i)\varphi(Y_i),$$

où les Y_i, sont i.i.d. de densité g, les résultats vus pour $\hat{I}_n$ s'appliquent à nouveau ici.

Proposition 2.4. ***(Echantillonnage préférentiel)***

Si $\mathbb{E}|\omega(Y)\varphi(y)| < \infty$, *et* $\mathbb{E}[\tilde{I}_n] = I$

$$\tilde{I}_n = \frac{1}{n}\sum_{i=1}^{n} \omega(Y_i)\varphi(Y_i) \xrightarrow[n\to+\infty]{p.s} \mathbb{E}[\omega(Y)\varphi(Y)] = I$$

Si de plus $\mathbb{E}[\omega^2(Y)\varphi^2(Y)] < \infty$, *alors*

$$\sqrt{n}(\tilde{I}_n - I) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,s^2),$$

avec

$$s^2 = Var(\omega(Y)\varphi(Y)) = \int \omega^2(y)\varphi^2(y)g(y)dy - I^2 = \mathbb{E}[\omega(X)\varphi^2(X)] - I^2.$$

Remarque 2.1. Comme précédemment, la variance σ^2 est naturellement estimée par

$$\tilde{\sigma}_n = \frac{1}{n}\sum_{i=1}^{n} \omega^2(Y_i)\varphi^2(Y_i) - \tilde{I}_n^2$$

pour en déduire des intervalles de confiance asymptotiques.
La variance σ^2 de $\tilde{I}_n$ est à comparer à la variance $\sigma^2 = \mathbb{E}[\varphi^2(X)] - I^2$ de $\hat{I}_n$. Il "suffit" donc de choisir la pondération ω, c'est-à-dire la densité instrumentale g, de sorte que le terme $\mathbb{E}[\omega(X)\varphi^2(X)]$ soit le plus petit possible. La bonne nouvelle, c'est que ce problème a une solution explicite, la mauvaise c'est qu'elle est tout à fait hors d'atteinte...

Lemme 2.1. ***(Loi d'échantillonnage optimale)***

Pour toute densité g *telle que* $\mathbb{E}[\omega^2(Y)\varphi^2(Y)] < \infty$, *on a*

$$s^2 = Var(\omega(Y)\varphi(Y)) \geq \mathbb{E}[|\varphi(X)|]^2 - I^2 = \left(\int |\varphi(x)|f(x)dx\right)^2 - \left(\int \varphi(x)f(x)dx\right)^2,$$

la borne inférieure étant atteinte pour la densité $g^\star$ *définie par*

$$g^\star(y) = \frac{|\varphi(y)|f(y)}{\int |\varphi(y)|f(y)dy}.$$

Si φ est de signe constant, la variance obtenue avec $g^\star$ est nulle, ce qui signifie qu'un seul tirage selon $g^\star$ suffit ! En effet, si $Y \sim g^\star$, alors

$$\tilde{I}_1 = \omega(Y)\varphi(Y) = \frac{f(Y)}{g^\star(Y)}\varphi(Y) = \int \varphi(x)f(x)dx.$$

Même si la proposition précédente présente surtout un intérêt théorique, on retrouve l'idée selon laquelle la densité auxiliaire g doit réaliser un compromis entre la fonction à intégrer φ et la densité f : autant que possible, g doit mettre du poids là où le produit $|\varphi(x)f(x)|$ est le plus élevé.

2.3 Chaîne de Markov

2.3.1 Introduction aux chaînes de Markov

Lorsque deux systèmes identiques à un instant donné peuvent avoir des comportements différents dans le futur, on est amené à introduire une suite de variables aléatoires $(X_t)_t$ pour décrire leurs évolutions : X_t servant à définir l'état du système étudié à l'instant t. Si le système étudié est une population, l'état du système à un instant donné peut être décrit simplement par un nombre lorsqu'on s'intéresse uniquement à la taille de cette population, ou par un ensemble de nombres tels que l'ensemble des positions de chaque individu lorsqu'on s'intéresse à la répartition spatiale de la population. On va ici se limiter à des systèmes dont l'état peut être décrit par une variable aléatoire ou un vecteur aléatoire discret. En général l'évolution futur d'un système dépendant au moins de son état présent, les variables aléatoires décrivant l'état du système à chaque instant ne pourront pas être considérées comme indépendantes. On va s'intéresser aux situations où l'évolution future d'un système ne dépend du passé qu'au travers de son état présent et pour simplifier on n'étudiera pas l'évolution du système en temps continu, mais son évolution en une suite infinie d'instants $0 = t_0 < t_1 < ... < t_n < ....$ On travaillera donc avec une suite de variables aléatoires discrètes $(X_n)_{n\in\mathbb{N}}$, chaque variable aléatoire étant à valeurs dans un ensemble fini ou infini dénombrable noté X que l'on identifiera à $\{1,...,N\}$ si $\mathcal{X}$ est composé de N éléments et à $\mathbb{N}^*$ si $\mathcal{X}$ est infini.

2.3.2 Généralités

Soit $X_0,X_1,...,X_n,...$ une suite de variables aléatoires définies sur un même espace de probabilité $(\Omega,\mathcal{A},\mathcal{P})$ et à valeurs dans $\mathcal{X}$.

2.3.2.1 Définitions et exemples

Définition 2.1. Formellement, soit $\mathcal{X}$ un espace fini ou dénombrable. Ce sera **l'espace d'états**. Soit $X = \{X_n; n \geq 0\}$ une suite de variables aléatoires à valeurs dans $\mathcal{X}$. On dit que X est une chaîne de Markov si pour tout $x_1,...,x_{n+1} \in \mathcal{X}$, on a

$$\mathrm{P}(\underbrace{X_{n+1} = x_{n+1}}_{\text{futur}} \mid \underbrace{X_1 = x_1, X_2 = x_2, ..., X_n = x_n}_{\text{le passé (et le présent)}}) = \mathbb{P}(\underbrace{X_{n+1} = x_{n+1}}_{\text{le futur}} \mid \underbrace{X_n = x_n}_{\text{le présent}})$$

Cette propriété des chaînes de Markov est aussi connue comme **propriété de Markov.**

Définition 2.2. Soit $(X_n)_{n\in\mathbb{N}}$ une chaîne de Markov à valeur dans $\mathcal{X}$.
La chaîne de Markov est dite **homogène** si pour tout $(i,j) \in \mathcal{X}^2$ et tout $n \in \mathbb{N}$, la probabilité de passer à l'état j à l'instant $n+1$ sachant qu'on l'état i à l'instant n ne dépendent pas de n :

$$P(X_{n+1} = j | X_n = i) = Q(i,j), \qquad \text{pour tout } n \in \mathbb{N}.$$

Les nombres $Q(i,j),(i,j) \in \mathcal{X}^2$ s'appellent les **probabilité de transition** de la chaîine de Markov.

Définition 2.3. Lorsque $\mathcal{X} = \{1,...,N\}$, la matrice Q représentée par un tableau à N lignes et N colonnes est appelée **matrice de transition** de la chaîne de Markov $(X_n)_n$:

$$Q = \begin{pmatrix} Q(1,1) & . & . & . & Q(1,N) \\ . & & & & . \\ . & & & & . \\ . & & & & . \\ Q(N,1) & . & . & . & Q(N,N) \end{pmatrix}$$

Que $\mathcal{X}$ soit un ensemble fini ou dénombrable, on appllera $Q = (Q(i,j))_{1 \leq (i,j) \leq N}$ la **matrice de transition** de la chaîne de Markove $(X_n)_n$.

Définition 2.4. La loi de X_0 est appelée la **loi initiale de la chaîne de Markov**. On l'écrira

$$\pi_0 = (P(X_0 = 1), P(X_0 = 2), ..., P(X_0 = N-1), P(X_0 = N)) \text{ si } \mathcal{X} = \{1,...,N\}.$$

Comme la i-ième ligne de la matrice de transition d'une chaîne de Markov homogène $(X_n)_n$ définit la loi conditionnelle de X_n sachant que $\{X_{n-1} = i\}$, on a la propriété suivante :

Définition 2.5. On dit qu'une matrice Q de transition(éventuellement infinie) est **stochastique** ssi tous ses coefficients sont > 0 et si la somme de chaque ligne fait 1 :

$$\sum_{j \in \mathcal{X}} Q(i,j) = 1.$$

On dit aussi est une **matrice markovienne.**

Propriété 2.1. Soit $(X_n)_{n \in \mathbb{N}}$ une chaîne de Markov homogène à valeurs dans $\mathcal{X}$ de matrice de transition Q. La matrice Q est une matrice stochastique, c'est-à-dire : pour tout $i,j \in \mathcal{X}, Q(i,j) \geq 0$ et pour tout $i \in \mathcal{X}, \sum_{j \in \mathcal{X}} Q(i,j) = 1$.

Démonstration :

Soit $i \in \mathcal{X}$. Alors

$$\begin{aligned} \sum_{j \in \mathcal{X}} Q(i,j) &= \sum_j \mathbb{P}(X_1 = j | X_0 = i) \\ &= \mathbb{P}(\text{"}X_n \text{ aille quelque part après être allé en } i \text{ ..."}) \\ &= 1 \end{aligned}$$

Plus formellement,

$$\begin{aligned} \sum_j \mathbb{P}(X_1 = j | X_0 = i) &= \sum_j \mathbb{E}(\mathbb{1}_{X_1 = j} | X_0 = i) \\ &= \mathbb{E}(\sum_j \mathbb{1}_{X_1 = j} | X_0 = i) \end{aligned}$$

Où

$$\sum_j \mathbb{1}_{X_1 = j} = 1$$

tout le temps car X_1 est égal à un et un seul des j.

Donc $\sum_j \mathbb{P}(X_1 = j \mid X_0 = i) = \mathbb{E}(1 \mid X_0 = i) = 1$

On va voir que réciproquement, pour toute matrice stochastique Q sur un produit $E \times E$, il existe une chaîne de Markov sur $\mathcal{X}$ qui admet Q comme matrice de transition. Mais elle n'est pas unique, car il faut encore préciser la loi du premier état X_0.

2.3.2.2 Graphe associé à une matrice de transition

Pour visualiser l'évolution d'une chaîne de Markov homogène, il est souvent utile de représenter la matrice de transition Q de la chaîne de Markov par un graphe orienté : les noeuds du graphe sont les états possibles pour la chaîne de Markov, une flèche allant de l'état i à l'état j indique qu'il y a une probabilité strictement positive que le prochain état de la chaîne soit l'état j si elle est actuellement dans l'état i. On met le poids $Q(i,j)$ à la flêche allant de l'état i à l'état j.

Théorème 2.1. *(Équation de Chapman-Kolmogorov)*

Pour tout état $i,j \in \mathcal{X}$, pour tout $n,m \in \mathbb{N}$ et $k \in [0,n]$, on a l'égalité

$$p_{ij}^{(n+m)} = \sum_{k \in \mathcal{X}} p_{ik}^{(m)} p_{kj}^{(n)} \tag{14}$$

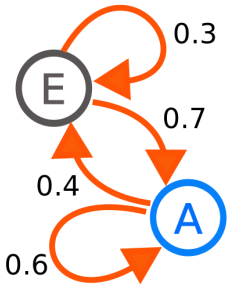
En notation matricielle, on a $P^{(n+m)} = P^{(n)} P^{(m)}$.

Démonstration :

Soient k l'état de la chaîne au temps m. On a

$$\begin{aligned}
p_{ij}^{n+m} &= \mathbb{P}(X_{n+m} = j | X_0 = i) \\
&= \sum_{k \in \mathcal{X}} \mathbb{P}(X_{n+m} = j, X_m = k | X_0 = i) \\
&= \sum_{k \in \mathcal{X}} \frac{\mathbb{P}(X_{n+m} = j, X_m = k, X_0 = i)}{\mathbb{P}(X_0 = i)} \\
&= \sum_{k \in \mathcal{X}} \frac{\mathbb{P}(X_{n+m} = j | X_m = k, X_0 = i)\mathbb{P}(X_m = k | X_0 = i)\mathbb{P}(X_0 = i)}{\mathbb{P}(X_0 = i)} \\
&= \sum_{k \in \mathcal{X}} \mathbb{P}(X_{n+m} = j | X_m = k, X_0 = i)\mathbb{P}(X_m = k | X_0 = i) \\
&= \sum_{k \in \mathcal{X}} \mathbb{P}(X_m = k | X_0 = i)\mathbb{P}(X_{n+m} = j | X_m = k) \\
&= \sum_{k \in \mathcal{X}} p_{ik}^{(m)} p_{kj}^{n}.
\end{aligned}$$

Exemple 2.5. *Élémentaire de chaîne de Markov, à deux états A et E. Les flèches indiquent les probabilités de transition d'un état à un autre.*



Exemple 2.6. ***la ligne téléphonique.***

On considère une ligne de téléphone. L'état X_n de cette ligne à l'étape n est 0 si elle est libre et 1 si elle occupée. Entre deux instants successifs, il y a une probabilité $1/2$ pour qu'un appel arrive. Si la ligne est occupée et qu'un appel arrive, cet appel est perdu. La probabilité pour que la ligne se libère entre l'instant n et l'instant $(n+1)$ est $1/3$.
Le graphe de transition de cette chaîne de Markov est donné figure 4.
La matrice de transition est la suivante :

$$P = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 2/3 \end{bmatrix}$$

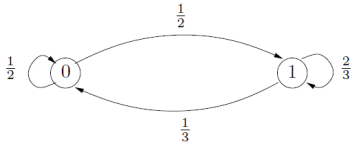


FIG. 4 – *Graphe de transition de la ligne téléphonique.*

Les probabilités de transition en n étapes sont en fait complètement déterminées par les probabilités de transition en un coup, c'est-à-dire par la matrice de transition. Ceci est explicité par les équations de Chapman-Kolmogorov, que nous allons voir maintenant.

Notation. La probabilité d'aller de l'état i à l'état j en n coups est notée

$$p_{ij}^{(n)} = \mathbb{P}(X_n = j \mid X_0 = i),$$

et la matrice de transition en n coups est notée :

$$P^{(n)} = \left[p_{ij}^{(n)}\right]_{1 \leq i,j \leq M}.$$

On adopte aussi la convention $P(0) = I_M$, matrice identité de taille M.

2.3.3 Classification des états

Il existe différents types d'états. Leur classification permet de mieux étudier les propriétés asymptotiques des chaînes de Markov. Maintenant, classifions les états possibles selon diverses caractéristiques.

2.3.3.1 Caractères récurrents et transitoires

Les définitions qui suivent sont tirées des ouvrages de Taylor et Karlin (1998) et de Lessard (2013).

Définition 2.6. Soit $i \in \mathcal{X}$, où $\mathcal{X}$ est l'espace des états.

1. Un état i est dit **récurrent** si, $\mathbb{P}(\exists n \geq 1, X_n = i | X_0 = i) = 1$. C'est à dire que la probabilité d'un éventuel retour à l'état i vaut 1, sachant que la chaîne a commencé à l'état i.
Sinon, on dit que l'état est **transitoire**.

2. Un état i est **récurrent nul** si, $\rho_i = E[R_i | X_0 = i] = \infty$ où $R_i = \min\{r : X_r = i\}$. Sinon, l'état i est **récurrent positif**.

Définition 2.7. Soient $i,j \in \mathcal{X}$ et $n \geq 1$ un entier.

La probabilité du premier temps de passage à l'état j, au n-ième pas, sachant que le processus démarre à l'état i est définie par :

$$f_{ij}^{(n)} = \mathbb{P}(X_n = j, X_k \neq j, k = 1,2,...,n-1 | X_0 = i)$$

$$= \mathbb{P}(R_i = n | X_0 = i) \qquad n = 1,2...$$

Par convention $f_{ij}^{(0)} = 0$.

Proposition 2.5. *La probabilité f_{ij} d'un possible passage de la chaîne à l'état j sachant que'elle démarre à l'état i vaut*

$$f_{ij} = \sum_{n=1}^{\infty} f_{ij}^{(n)}$$

Démonstration:

$$\begin{aligned} f_{ij} &= \mathbb{P}(\text{un possible passage à l'état } j | X_0 = i) \\ &= \mathbb{P}(\bigcup_{n=1}^{\infty} \{\text{le premier passage à 1 'état } j \text{ se fait au temps n}\} | X_0 = i) \\ &= \sum_{n=1}^{\infty} (\text{le premier passage à l'état } j \text{ se fait au temps } n | X_0 = i) \\ &= \sum_{n=1}^{\infty} f_{ij}^{(n)} \end{aligned}$$

Remarque 2.2. On peut donc affirmer qu'un état i est récurrent si $f_{ii} = 1$ et transitoire si $f_{ii} < 1$.

Théorème 2.2. *(1) L'état i est récurrent si seulement si $\sum_{n=0}^{\infty} P_{ii}^n = +\infty$.*

(2) L'état i est transient si et seulement si $\sum_{n=0}^{\infty} P_{ii}^n < +\infty$.

Ce théorème se démontre à partir du lemme suivant:

Lemme 2.2. *On a $\mathbb{E}_i(N_i) = \sum_{n=0}^{\infty} P_{ii}^n$.*

Démonstration.

$$\mathbb{E}_i(N_i) = \mathbb{E}_i\left(\sum_{n=0}^{\infty} \mathbb{1}_{X-n=i}\right) = \sum_{n=0}^{\infty} \mathbb{E}_i(\mathbb{1}_{X-n=i}) \text{ par Fubini-Tonelli,}$$
$$\sum_{n=0}^{\infty} \mathbb{P}_i(X - n = i) = \sum_{n=0}^{\infty} P_{ii}^n$$

Corollaire 2.1. *Soit i un point de E. Alors*

$$\mathbb{P}_i(N_i = \infty) = 1 \Longleftrightarrow \mathbb{P}_i(N_i = \infty) > 0,$$
$$\mathbb{P}_i(N_i = \infty) < 1 \Longleftrightarrow \mathbb{P}_i(N_i = \infty) = 0.$$

Corollaire 2.2. *Toute chaîne de Markov homogène sur un espace d'états fini admet (au moins) une classe récurrente.*

2.3.3.2 Classes irréductibles

Définition 2.8. Soient i et j deux états de $\mathcal{X}$. L'état j est accessible depuis l'état i, noté $i \longrightarrow j$, si :

$$\exists n \in \mathbb{N}, p_{ij}^{(n)} = \mathbb{P}(X_n = j | X_0 = i) > 0.$$

On dit que les états i et j **communiquent** si ils sont tous deux accessibles l'un de l'autre. On note $i \longleftrightarrow j$.
En d'autres termes, on dit que l'état j est accessible depuis l'état i si la probabilité d'atteindre j en n transitions depuis i est strictement positive.

Définition 2.9. La période d'un état $i \in \mathcal{X}$ notée $d(i)$ est l'entier définie par:

$$d(i) = PGCD\{n \geq 1 : p_{ii}^{(n)} > 0\}$$

On dit que i est **périodique** si $d(i) > 1$, sinon il est **apériodique**.

Définition 2.10. Soit $C \subseteq \mathcal{X}$ une classe d'états.

1. C est dite **fermée** si $\forall i \in C$, $j \notin C$ et $n \geq 1$, $p_{ij}^{(n)} = 0$.

2. C est dite **irréductible** si $\forall i,j \in C$, $i \longleftrightarrow j$.

Une classe d'état est clone fermée si aucun état hors d'elle n'est accessible depuis ses états intérieurs.
De plus, une chaîne de Markov est irréductible si elle n'est formée que d'une unique classe fermée.

Théorème 2.3. *(Décomposition de l'espace d'états Irwin (2006))*
Par la relation d'équivalence $\longleftrightarrow$, il existe une unique partition de l'espace d'états $\mathcal{X}$ telle que

$$\mathcal{X} = T \cup (\cup_{k=1}^{\infty} C_k) \tag{15}$$

où T est une classe uniquement constituée d'états transitoires et $\{C_k\}_{k\geq 1}$ est une suite de classes irréductibles fermées d'états récurrents.

Démonstration :

Soit $\{C_k\}_{k\geq 1}$ une suite de classe d'états récurrents pour la relation $\longleftrightarrow$.
Montrons que les C_k sont fermées :
On procède par l'absurde. Supposons qu'il existe $i \in C_k$ et $j \notin C_k$ tel que $p_{ij} > 0$.
On a j ne communique pas avec i donc :

$$\mathbb{P}(X_n \neq i, \forall n \geq 1 | X_0 = i) \geq \mathbb{P}(X_1 = j | X_0 = i) = p_{ij} \geq 0$$

Or i est un état récurrent, ce qui est contradictoire car si i est récurrent alors :$\exists n \geq 1$ tel que :

$$\mathbb{P}(X_n = i | X_0 = i) = 1$$

D'où:

$$\mathbb{P}(X_n \neq i, \forall n \geq 1 | X_0 = i) = 0$$

Ainsi, on vient de prouver que toutes les classes irréductibles d'états récurrents sont fermées. En outre, puisque la relation $\longleftrightarrow$ est une relation d'équivalence, on a directement la partition contenant la classe d'états transitoires T et la suite des classes irréductibles fermées d'états récurrents.
Exemple 2.7.

$$P = \begin{pmatrix} 1/4 & 1/4 & 1/2 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 1/3 & 1/3 & 1/3 \end{pmatrix}$$

Considérons l'espace d'états $\mathcal{X} = \{1,2,3,4,5\}$. Soit la chaîne de Markov de diagramme et matrice de transition.

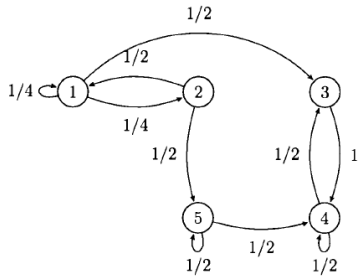


FIG. 5 – *Diagramme chaîne de Markov Exemple 2.8*

La chaîne de markov a trois classes d'états {1,2}.{3,4} *et* {5}. *Dans la classe* {1,2}, $p_{13} = p_{25} = \frac{1}{2} \neq 0$. *Alors la classe n'est pas fermée, donc transitoire et périodique car*

$$d(1) = d(2) = pgcd\{n \geq 1 : p_{11}^{(n)}) > 0 \text{ et } p_{22}^{(n)} > 0\} = pgcd\{2,4,6,8,...\} = 2.$$

Dans la classe {3,4}, *on remarque que* $\forall n \geq 1$ *et* $\forall j \neq \{3,4\}, p_{3j}^{(n)} = p_{4j}^{(n)} = 0$. *Elle est alors fermée sur un nombre d'états fini, donc récurrente positive et apériodique car*

$$d(3) = d(4) = 1.$$

Dans la classe {5}, $p_{54} = \frac{1}{2} \neq 0$. *Elle n'est pas fermée, donc transitoire et* 2 *apériodique car*

$$d(5) = pgcd\{n \geq 1 : p_{55}^{(n)} > 0\} = pgcd\{1,2,3,...\} = 1.$$

Définition 2.11. Si π une probabilité sur $\mathcal{X}$ vérifie $\pi(x)Q(x,y) = \pi(y)Q(y,x)$ alors π est dite **symétrique** (par rapport à Q) ou Q-symétrique.

Lemme 2.3. *Si* π *est* Q*-symétrique, alors elle est* Q*-invariante.*

Démonstration. Pour tout y,

$$\begin{aligned}(\pi Q)(y) &= \sum_{x \in \mathcal{X}} \pi(x)Q(x,y), \\ (\text{symétrie}) &= \sum_{x \in \mathcal{X}} \pi(y)Q(y,x), \\ &= \pi(y).\end{aligned}$$

Exemple 2.8. *(Marche simple dans $\mathbb{Z}$)*

Soit Q noyau de Markov de $\mathbb{Z}$ dans $\mathbb{Z}$ donné par

$$Q(x,x+1) = \frac{1}{2},\ Q(x,x-1) = \frac{1}{2},\ \forall x.$$

Pour tout x, $Q^2(x,x) = \mathbb{P}(X_2 = x|X_0 = x)$ (où (X_n) chaîne de Markov de transition Q). Nous avons donc

$$Q^2(x,x) \geq \mathbb{P}(X_2 = x, X_1 = x+1|X_0) = \left(\frac{1}{2}\right) > 0.$$

Par ailleurs, si $Q^n(x,x) > 0$ alors n est pair (on fait forcément un nombre pair de pas pour aller de x à x). Donc $d(x) = 2$. Donc le noyau Q n'est pas apériodique.

Définition 2.12. Soit Q un noyau de Markov. Si tous les x de $\mathcal{X}$ sont apériodique alors Q est dit **apériodique**.

Lemme 2.4. *Si Q est irréductible, $[\exists x \in \mathcal{X}$ qui est apériodique$] \Rightarrow [Q$ est apériodique$]$.*

Théorème 2.4. *Supposons que $\mathcal{X}$ est fini et que Q est une matrice de transition irréductible, apériodique et admettant une probabilité invariante π_0.*

(1) Il existe deux réels $\alpha \in [0,1[, M \in \mathbb{R}_+$ tles que pour tout $x,y \in \mathcal{X}$,

$$|Q^n(x,y) - \pi_0(y)| \leq M\alpha^n.$$

(2) Pour tout $x \in \mathcal{X}$ et toute fonction $f : \mathcal{X} \to \mathbb{R}$, si on appelle $(X_n)_{n\geq 0}$ une chaîne de Markov de loi initiale π_0 quelconque et de transition Q

$$\sqrt{n+1}\left(\frac{1}{n+1}\sum_{p=0}^{n} f(X_p) - \sum_{y\in E}\pi_0(y)f(y)\right) \xrightarrow[n\to+\infty]{loi} \mathcal{N}(0,\overline{\sigma}^2),$$

avec une variance $\overline{\sigma}^2 < +\infty$.

Remarque 2.3. Dans le cas d'un espace fini ou dénombrable, $\int_{x\in\mathcal{X}} f(x)\pi(dx) = \sum_{x\in\mathcal{X}} f(x)\pi(x)$ pour toute probabilité π et toute fonction f.

2.3.4 Loi des X_n

Le comportement d'une chaîne de Markov X dépend entièrement de sa matrice de transition Q, et de la position initiale X_0. On appelle μ_0 la **loi initiale** de X, c'est une mesure définie par

$$\mu_0(\{x\}) = \mathbb{P}(X_0 = x).$$

Mesure et notations. Toutes les mesures que l'on va voir dans ce cours sont sur un espace fini ou dénombrable $\mathcal{X}$. Cela veut dire qu'elles sont uniquement déterminées par leurs valeurs sur les singletons : $A \subset \mathcal{X}$,

$$\mu(A) = \sum_{x \in A} \mu(\{x\}).$$

On note en général abusivement $\mu(x) = \mu(\{x\})$ pour $x \in \mathcal{X}$, et ces valeurs déterminent entièrement la mesure. On peut aussi noter la mesure comme le vecteur (ou la suite si $\mathcal{X}$ est infini) de ses valeurs :

$$\mu = (\mu(x))_{x \in \mathcal{X}}.$$

Par exemple si $\mathcal{X}$ a 2 états et que la loi μ_0 de X_0 peut prendre indifféremment les deux valeurs avec probabilité 1/2,

$$\mu_0 = (1/2,1/2).$$

Connaissant μ_0 et Q, on peut calculer directement la loi de X_n.

Proposition 2.6. *Pour toute suite* $\{x_0,x_1,...,x_n\}$ *dans* $\mathcal{X}$, *on a*

$$\mathbb{P}(X_0 = x_0,X_1 = x_1,X_2 = x_2,...,X_n = x_n)$$

$$= \mu_0(x_0)Q(x_0,x_1)Q(x_1,x_2)...Q(x_{n-1},x_n)$$

Démonstration :

On a (en utilisant la propriété de Markov)

$$\begin{aligned}
&\mathbb{P}(X_0 = x_0,X_1 = x_1,X_2 = x_2,...,X_n = x_n)\\
&= \mathbb{P}(X_n = x_n,...,X_1 = x_1 \mid X_0 = 0)\mathbb{P}(X_0 = x_0)\\
&= \mathrm{P}(\mathrm{X}_n = x_n,...,X_1 = x_1 \mid X_0 = 0)\mu_0(x_0)\\
&= \mathbb{P}(X_n = x_n,...,X_2 = x_2 \mid X_1 = x_1,X_0 = 0)\mathbb{P}(X_1 = x_1 \mid X_0 = x_0)\mu_0(x_0)\\
&= \mathbb{P}(X_n = x_n,...,X_2 = x_2 \mid X_1 = x_1,X_0 = 0)Q(x_0,x_1)\mu_0(x_0)\\
&= \mathbb{P}(X_n = x_n,...,X_2 = x_2 \mid X_1 = x_1)Q(x_0,x_1)\mu_0(x_0)
\end{aligned}$$

en utilisant la **propriété de Markov**. En utilisant le même raisonnement, on montre que

$$\mathbb{P}(X_n = x_n,...,X_2 = x_2 \mid X_1 = x_1) = \mathbb{P}(X_n = x_n,...,X_3 = x_3 \mid X_2 = x_2)Q(x_1,x_2)$$

$$\mathbb{P}(X_0 = x_0,X_1 = x_1,X_2 = x_2,...,X_n = x_n) = \mathbb{P}(X_n = x_n,...,X_3 = x_3 \mid X_2 = x_2)Q(x_1,x_2)Q(x_0,x_1)\mu_0(x_0)$$

Par récurrence, on montre la formule proposée.

Pour une même chaîne X, on considère souvent plusieurs lois initiales différentes. Dans ce cas on précise la loi utilisée en notant

$$\mathbb{P} = \mathbb{P}_\mu,$$

dans chaque calcul de probabilité, et l'espérance est alors notée $\mathbb{E}_x$. Si la loi est un "Dirac" $\mu = \sigma_x$ pour un certain $x \in \mathcal{X}$ (ce qui veut dire $X_0 = x$ p.s.), alors on note plus simplement $\mathbb{P}_{\sigma_x} = \mathbb{P}_x, \mathbb{E}_{\sigma_x} = \mathbb{E}_x$.

Notation . Pour une mesure μ_0 et une matrice Q, on note la mesure

$$(\mu_0 Q)(y) = \sum_{x \in \mathcal{X}} \mu_0(x)Q(x,y).$$

Cela revient à multiplier (matriciellement) la mesure μ_0 vue comme un vecteur $\mu_0 = (\mu_0(x_1),\mu_0(x_2),...)$ par la matrice Q.

Proposition 2.7. *Si μ est la loi de X_0, alors (μQ) est la loi de X_1.*

Démonstration :

Soit $y \in \mathcal{X}$.

$$\mathbb{P}(X_1 = y) = \sum_x \mathbb{P}(X_1 = y, X_0 = x)$$
$$= \sum_x \mathbb{P}(X_1 = y, X_0 = x)\mathbb{P}(X_0 = 0) = \sum_x \mu(x)Q(x,y) = (\mu Q)(y).$$

Proposition 2.8. *Pour tout n, la loi de X_n est μQ^n.*

Démonstration :

D'après la proposition (2.6), la loi de X_1 est $\mu_1 = \mu Q$.
On peut alors "oublier" la variable X_0, et ne considérer que la chaîne qui part de X_1 (formellement, poser $X'_0 = X_1, X'_1 = X_2$, etc ...). La matrice de transition est toujours Q, par contre la loi initiale n'est plus μ, c'est μ_1.

La loi de X_2 (c'est-à-dire X'_1) est donc, en réutilisant la proposition (2.6),

$$\mu_2 = (\mu_1(Q) = ((\mu Q) * Q).$$

Comme le produit matriciel est associatif, $((\mu Q) * Q) = \mu * Q * Q = \mu * (Q * Q) = \mu * Q^2$.
La loi de X_2 est donc bien μQ^2, comme annoncé. En raisonnant par récurrence, on montre bien $\mu_3 = (\mu * Q^2) * Q = \mu Q^3$,...,... et $\mu_n = \mu Q^n$.

2.3.5 Distribution stationnaire et théorème ergodique

L'objectif de cette section est de trouver les conditions simples permettant d'approximer la loi d'une chaîne de Markov $\{X_n\}_{n\geq 1}$ sur une longue période, plus clairement sous quelles conditions pourrons nous trouver la limite $\lim_{n\to\infty} X_n$ afin de pouvoir identifier facilement la chaîne.

2.3.5.1 Distribution stationnaire

Définition 2.13. Un vecteur $\pi = \{\pi_i\}_{i\in\mathcal{X}}$ sur $\mathcal{X}$ est **stationnaire** pour la chaîne de Markov X si

$(i)\pi_i \geq 0, \forall_i \in \mathcal{X}$ et $\sum_{i\in\mathcal{X}} \pi_i = 1$,

$(ii)\pi_j = \sum_{i\in\mathcal{X}} \pi_i P_{ij}, \forall j \in \mathcal{X}$ ou en notation matricielle $\pi = \pi P$.

Une chaîne de Markov est donc stationnaire sous $\mathbb{P}$ si pour tout $k,n \in \mathbb{N}$, la distribution du vecteur aléatoire $(X_1,X_2,...,X_n)$ est identique à celle du vecteur $(X_k,X_{k+1}...,X_{n+k})$.

Remarque 2.4. La stationnarité de cette distribution se voit en itérant l'égalité de l'assertion (ii),
soit $\pi P^2 = (\pi P)P = \pi P = \pi$.

De la même manière, on a $\pi P^n = \pi \qquad n \in \mathbb{N}$.

Théorème 2.5. *Soit $P = p_{jk}$ la matrice de transition $m \times m$ d'une chaîne de Markov régulière sur m états. Alors*

- *pour $n \to \infty$, les puissances P^n approchent une matrice de transition P^∞ de la forme*

$$P^\infty = \begin{pmatrix} \pi_1 & \pi_2 & \dots & \pi_{m-1} & \pi_m \\ \pi_1 & \pi_2 & \dots & \pi_{m-1} & \pi_m \\ \dots & \dots & \dots & \dots & \dots \\ \pi_1 & \pi_2 & \dots & \pi_{m-1} & \pi_m \\ \pi_1 & \pi_2 & \dots & \pi_{m-1} & \pi_m \end{pmatrix}$$

avec $\pi_j > 0 \quad et \sum_{j=1}^{m} \pi_j = 1$

- *pour la distribution $\pi = (\pi_1,\pi_2,...,\pi_m)^T$ est la seule solution de l'équation*

$$\sum_{j=1}^{m} \pi_j p_{jk} = \pi_k \ \forall k \qquad \text{c'est-à-dire} \quad P^T\pi = \pi.$$

obéissant à la condition de normalisation

$$\sum_{j=1}^{m} \pi_j = 1$$

La distribution π est qualifiée de distribution stationnaire ou d'équilibre associée à la chaîne P.

2.3.5.2 Théorème ergodique

Théorème 2.6. *(Théorème ergodique)*
Soit Q un noyau de Markov irréductible (sur $\mathcal{X}$). Alors, il existe une probabilité Q-invariante π_0 et de plus :

(1) π_0 est l'unique probabilité Q-invariante.

(2) Tous les états sont récurrents (pour Q).

(3) Si$(X_n)_{n\geq 0}$ est une chaîne de Markov de loi initiale π quelconque et de transition Q alors

$$\frac{1}{n+1}\sum_{p=0}^{n} f(X_p) \xrightarrow[n\to+\infty]{ps} \int f(x)\pi_0(dx),$$

pour toute fonction f qui est π_0-intégrable.
Ce théorème nous permet d'approcher une intégrale par une moyenne empirique. Ici, la suite des (X_p) n'est pas i.i.d. mais est une chaîne de Markov. On parle alors de méthode de Monte-Carlo par chaîne de Markov (« MCMC » pour « Monte Carlo Markov Chain » en anglais).

2.4 Pratique des méthodes MCMC

L'utilisation intensive des méthodes de simulation MCMC est due à leur adaptabilité pour une vaste classe de problèmes et de modèles mais également à l'amélioration des technologies informatiques. Les algorithmes sont faciles à implanter et un simple ordinateur de bureau est suffisant pour mettre en oeuvre la réalisation des simulations. La part la plus importante du travail, et sans doute une des plus intéressantes, est l'élaboration du schéma de simulation pour un modèle donné: distribution a priori des paramètres du modèle (Robert, 1992), distribution candidate pour l'algorithme de Metropolis-Hastings et distributions conditionnelles pour l'échantillonnage de Gibbs (Robert, 1992) par exemple. Toutefois, un aspect important de l'implantation de ce type de méthode est qu'il faut utiliser plusieurs réalisations ou simulations sur un ensemble de données pour obtenir des résultats statistiquement significatifs, étant donné la nature aléatoire des algorithmes de simulation. Les chaînes de Markov sont généralement très riches en informations statistiques, plus qu'il n'en faut pour calculer seulement les estimateurs de type (8)(chapitre 1).

Deux algorithmes MCMC sont présentés dans la section suivante: l'algorithme de Metropolis-Hastings et l'échantillonnage de Gibbs. Ces méthodes permettent de construire un noyau de transition en fonction d'une distribution d'intérêt donnée telle que les séquences d'échantillons simulés qui forment une chaîne de Markov convergente vers la distribution stationnaire voulue.

2.4.1 Algorithme de Metropolis-Hastings

La technique de Metropolis-Hastings a été développée par Metropolis et al.(1953), au départ pour la physique particulaire, et généralisée par Hastings (1970) dans un cadre plus statistique.

L'algorithme de Metropolis-Hastings est un schéma de simulation permettant de générer des échantillons suivant une distribution de $\pi(x)$ dont une forme analytique est disponible. Cette méthode a d'abord été utilisée pour la résolution de problèmes en mécanique statistique avant d'être élargie à un cadre plus général de simulation statistique. Bien que les premiers résultats sur cette méthode de simulation soient connus depuis les années 1950, son utilisation pratique n'est devenue possible que depuis une trentaine d'années car elle nécessite de puissants moyens de simulation informatique.

Schéma de simulation: l'algorithme de Metropolis-Hastings est étudié et présenté dans de nombreux ouvrages consacrés aux méthodes de simulation MCMC (voir (Robert, 1992)). Le schéma de simulation est décrit ci-dessous:

2.4.1.1 Algorithme de Metropolis-Hastings

On se fixe une loi π suivant laquelle on aimerait simuler ou dont on voudrait calculer une intégrale $\int_{\mathcal{X}} f(x)\pi(dx))$. Nous appellerons π la **loi cible**. On se donne un noyau de Markov Q. Nous appellerons Q le **noyau de proposition**. Nous allons construire une chaîne de Markov $(X_n)_{n\geq 0}$.

- Nous prenons $X_0 = x_0$ tel que $\pi(x_0) > 0$.
- Si $X_n = x$, nous simulons Y_{n+1} et U_{n+1} indépendants (et indépendants des simulations passées) avec

$$Y_{n+1} \sim Q(x,.),\ U_{n+1} \sim \mathcal{U}([0;1]).$$

La variable Y_{n+1} s'appelle une proposition. Posons, pour tout x,y

$$\alpha(x,y) = \min\left(1, \frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}\right).$$

Alors : $X_{n+1} = \begin{cases} Y_{n+1}, & \text{si } U_{n+1} \leq \alpha(X_n,Y_{n+1}), \\ X_n, & \text{sinon.} \end{cases}$

Dans le cas où $U_{n+1} \leq \alpha(X_n,Y_{n+1})$, on dit qu'on accepte la proposition, et dans le cas contraire on dit qu'on refuse la proposition.

Proposition 2.9. *La suite aléatoire $(X_n)_{n\geq 0}$ construite ci-dessus est une chaîne de Markov de transition P avec*

$$\begin{cases} P(x,y) = Q(x,y)\alpha(x,y), & \text{si } x \neq y, \\ P(x,y) = 1 - \sum_{y\neq x} P(x,y). \end{cases}$$

La loi π est P-invariante

Démonstration :

Pour tout x, $\mathbb{P}(X_{n+1} = x|X_0,...,X_n) = \mathbb{P}(X_{n+1} = x|X_n)$ (d'après la construction ci-dessus).

Donc $(X_n)_{n\geq 0}$ est bien une chaîne de Markov. Calculons, pour tout $x \neq y$

$$\begin{aligned} \mathrm{P}(\mathrm{X}_{n+1} = y|X_n = x) &= \mathrm{P}(\mathrm{Y}_{n+1} = y, U_{n+1} \leq \alpha(X_n,Y_{n+1})|X_n = x) \\ &= \mathrm{P}(\mathrm{Y}_{n+1} = y, U_{n+1} \leq \alpha(x,Y_{n+1})|X_n = 1) \\ &= \mathrm{P}(\mathrm{Y}_{n+1} = y, U_{n+1} \leq \alpha(x,y)|X_n = x) \\ &= \mathrm{Q}(\mathrm{x},\mathrm{y})\alpha(x,y). \end{aligned}$$

Nous en dédusions $\mathbb{P}(X_{n+1} = y|X_n = x) = 1 - \sum_{y\in\mathcal{X}} P(x,y) = 1 - \sum_y Q(x,y)\alpha(x,y)$.

Pour tout $x \neq y$,

$$\begin{aligned} \pi(x)P(x,y) &= \pi(x)Q(x,y)\alpha(x,y) \\ &= \pi(x)Q(x,y)\min\left(1, \frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}\right) \\ &= \min(\pi(x)Q(x,y), \pi(y)Q(y,x)) \\ &= \min(\pi(y)Q(y,x), \pi(y)Q(x,y)) \end{aligned}$$

(on refait le calcul en inversant x et y)

$$= \pi(y)P(y,x)$$

Donc π est symétrique par rapport à P. D'après le lemme 2.3, nous avons donc $\pi P = \pi$.

Exemple 2.9. *Soit Q défini dans l'exemple 2.8 et π une loi sur $\mathbb{Z}$ telle que $\pi(x) = C\exp(-\sqrt{|x|})$ pour tout x (C est alors la constante telle que $\sum_{x\in\mathbb{Z}} \pi(x) = 1$. Le programme 2.2 fait une simulation d'une marche de Metropolis de noyau de proposition Q et de loi-cible π. On remarque qu'il n'est pas nécessaire de connaître C pour implémenter l'algorithme*

Programme 2.2 Chain de Metropolis

```
n=100
f<-function(k)
{
return(exp(-sqrt(abs(k))))
}
liste=c()
x=0
for (k in 1:n)
{
liste=c(liste,x)
v=runif(1,0,1)
if (v<0.5)
{ y=x+1 }
else
{ y=x-1 }
u=runif(1,0,1)
alpha=min(1,f(y)*0.5/(f(x)*0.5))
     if (u<alpha)
     { x=y }
}
```

2.4.1.2 Algorithme de Metropolis simple

Si $Q(x,y) = Q(y,x)$ pour tout x,y (on dit que le noyau Q est **symétrique**). Dans la version originale de l'algorithme de Metropolis, le noyau Q est symétrique. Dans ce cas, α se simplifie en $\alpha(x,y) = min\left(1, \frac{\pi(y)}{\pi(x)}\right)$ pour tout x,y.

Proposition 2.10. *Si Q est un noyau de Markov irréductible, symétrique et si π est une probabilité non constante telle que $\pi(x) > 0$ pour tout x, alors la chaîne de Markov de Metropolis de noyau de proposition Q et de loi cible π est irréductible, apériodique, de loi invariante π.*

Démonstration :

Pour tout x,y, $\alpha(x,y) > 0$. De plus,pour tout x,y, il existe P $\in \mathbb{N}$ et une suite $x_1,...,x_p$ de $\mathcal{X}$ telle que $Q(x,x_1)Q(x_1,x_2)...Q(x_{p-1},x_p)Q(x_p,y) > 0$.
Donc :

$$P(x,x_1)P(x_1,x_2)...P(x_p,y) = Q(x,x_1)\alpha(x,x_1)...Q(x_p,y)\alpha(x_p,y) > 0.$$

Pour prouver l'apériodicité de P, il suffit de montrer qu'il existe x tel que $P(x,x) > 0$. Supposons que $P(x,x) = 0$ pour tout x. Or, pour tout x,

$$
\begin{aligned}
P(x,x) &= 1 - \sum_{y:y\neq x} P(x,y) \\
&= \textstyle\sum_y Q(x,y) - \displaystyle\sum_{y:y\neq x} Q(x,y)(1-\alpha(x,y) \\
&= Q(x,x) + \sum_{y:y\neq x} Q(x,y)(1-\alpha(x,y)) \\
&= Q(x,x) + \sum_{y:y\neq x} Q(x,y)\left(1-\frac{\pi(y)}{\pi(x)}\right)_+ ,
\end{aligned}
$$

donc $$P(x,x) - Q(x,x) = \sum_{y:y\neq x} Q(x,y)\left(1-\frac{\pi(y)}{\pi(x)}\right)_+ .$$

Fixons $x \in \mathcal{X}$. Les deux termes de l'équation ci-dessus sont de signes opposés, ils valent donc 0. Nous avons donc, pour tout y tel que $y \neq x$ et $Q(x,y) \neq 0$ (il en existe car Q est irréductible),

$$\left(1-\frac{\pi(y)}{\pi(x)}\right)_+ = 0, \tag{16}$$

c'est-à-dire $\pi(y) \geq \pi(x)$. Comme π n'est pas constante alors il existe y tel que $\pi(y) > \pi(x)$. Mais en inversant x et y dans (16), nous avons $\pi(x) \geq \pi(y)$. Nous aboutissons donc à une contradiction.

Donc il existe x tel que $P(x,x) > 0$. Donc le noyau P est apériodique.

Sous les hypothèse de cette proposition, nous pouvons appliquer le théorème vitesse de convergence.

2.4.2 L'échantillonnage de Gibbs

Cette méthode a été utilisée par Geman S et Geman D (1984) pour générer des observations à partir d'une distribution de Gibbs (distribution de Boltzmann). Il s'agit d'une forme particulière de la méthode de Monte-Carlo par chaîne de Markov qui, du fait de son efficacité, est largement utilisée dans de nombreux domaines d'analyse statistique bayésienne.
Dans la méthode de Gibbs, après avoir choisi un point départ, les d composantes du vecteur de covariables (θ) sont générées les unes après les autres conditionnellement à toute les autres composantes. Si $\pi(\theta \mid x)$ est la densité des d composantes du vecteur θ, conditionnellement aux données observées (x), nous utilisons alors les densités conditionnelles $\pi(\theta_1 \mid \theta_2,\theta_3,...,\theta_d,x)$, $\pi(\theta_2 \mid \theta_1,\theta_3,...,\theta_d,x)$ et ainsi de suite. À chaque k^e étape, la distribution conditionnelle utilise les valeurs générées les plus récentes parmi toutes les autres composantes. Par la théorie des chaînes de Markov, on a que lorsque $k \to \infty$ la densité des réalisations obtenues converge vers $\pi(\theta \mid x)$. Le schéma de simulation est décrit ci-dessous :

Algorithme d'échantillonnage de Gibbs

1.Fixer $k = 0$.

2. - Générer $\theta_1^{(k+1)}$ à partir de $\pi(\theta_1 \mid \theta_2^{(k)},\theta_3^{(k)},...,\theta_d^{(k)},x)$

- Générer $\theta_2^{(k+1)}$ à partir de $\pi(\theta_2 \mid \theta_1^{(k+1)},\theta_3^{(k)},...,\theta_d^{(k)},x)$

...

- Générer $\theta_{d-1}^{(k+1)}$ à partir de $\pi(\theta_{d-1} \mid \theta_1^{(k+1)},\theta_2^{(k+1)},...,\theta_d^{(k)},x)$

- Générer θ_d^{k+1} à partir de $\pi(\theta_d \mid \theta_1^{(k+1)},\theta_2^{(k+1)},...,\theta_{d-1}^{(k+1)},x)$.

3. Si la convergence est obtenue, alors

   - 1.a) retenir $\theta = \theta^{(k+1)}$

     sinon

   - 2.b) fixer $k = k + 1$, et retourner à 1.

Le schéma de simulation est d'autant plus efficace lorsque les distributions conditionnelles

$$\pi_i(\theta_i|\theta_1,...,\theta_{(i-1)},\theta_{(i+1)},...,\theta_d),$$

sont simulables rapidement.
Un des points remarquables de l'algorithme de Gibbs est qu'il rend possible la simulation de la distribution jointe $\pi(\theta) = \pi(\theta_1,...,\theta_d)$ à partir des distributions conditionnelles $\pi_i(\theta_i|\theta_{j\neq i})$.
Notons que l'échantillonnage de Gibbs s'applique particulièrement bien aux modèles hiérarchiques, c'est-à-dire à un modèle bayésien $(\pi(x \mid \theta),\pi(\theta))$, où la loi a priori $\pi(\theta)$ est décomposée en distributions conditionnelles

$$\pi_1(\theta \mid \theta_1),\pi_2(\theta_1,\theta_2),...,\pi_d(\theta_{d-1} \mid \theta_d),$$

pour des raisons structurelles ou calculatoires. De tels modèles apparaissent naturellement dans l'analyse bayésienne de structures complexes, où la diversité des informations a priori et la variabilité des observations nécessitent l'introduction de plusieurs niveaux de lois a priori (Robert, 1992).
Dans le cas particulier à deux composantes par exemple, la distribution jointe s'écrit d'après la formule de Bayes suivant les expressions

$$\pi(\theta_1,\theta_2) = \pi_1(\theta_1 \mid \theta_2) \times \pi^1(\theta_2) = \pi_2(\theta_2 \mid \theta_1) \times \pi^2(\theta_1), \tag{17}$$

où $\pi^1(\theta_2)$ et $\pi^2(\theta_1)$ désignent les distributions marginales. Il s'ensuit alors

$$\pi^1(\theta_2) = \frac{\pi_2(\theta_2 \mid \theta_1) \times \pi^2(\theta_2 \mid \theta_1)}{\pi_1(\theta_1 \mid \theta_2)} \propto \frac{\pi_2(\theta_2 \mid \theta_1)}{\pi_1(\theta_1 \mid \theta_2)}.$$

Ainsi, il est possible d'exprimer les densités des probabilités des distributions marginales à partir de celles des distributions conditionnelles, par exemple:

$$\pi^1(\theta_2) = \left(\int \frac{\pi_2(\theta_2 \mid x)}{\pi_1(x \mid \theta_2)} dx \right)^{-1} \frac{\pi_2(\theta_2 \mid \theta_1)}{\pi_1(\theta_1 \mid \theta_2)}.$$

La densité de la distribution jointe (17) s'écrit donc

$$\pi(\theta_1,\theta_2) = \left(\int \frac{\pi_2(\theta_2 \mid x)}{\pi_1(x \mid \theta_2)} dx \right)^{-1} \pi_2(\theta_2 \mid \theta_1).$$

Le cas général pour d composantes nécessite certaines conditions sur les lois conditionnelles.

Exemple 2.10. *Considèrons une loi jointe de la forme :*

$$f(x,y) \propto C_n^n y^{x+\alpha-1}(1-y)^{n-x+\beta-1},\ x = 0,...,n,\ 0 \leq y \leq 1.$$

On remarque que l'on peut proposer une loi binomiale de paramètres (n,y) pour $f(x \mid y)$ et une loi bêta de paramètres $x+\alpha, n-x+\beta$ pour $f(y \mid x)$.
On peut donc appliquer un algorithme de Gibbs pour obtenir des réalisations de $f(x)$, la loi marginale de x en simulant alternativement une réalisation x^* d'une binomiale de paramètres n,y^* où y^* est la valeur courante de y obtenu à l'étape précédente, puis une nouvelle réalisation x d'une bêta de paramètre $(x^*+\alpha, n-x^*)$.
Il se trouve qu'ici la loi marginale est accessible. En effet,

$$f(x) \propto \int_0^1 f(x,y)dy \propto \int_0^1 C_n^x y^{x+\alpha-1}(1-y)^{n-x+\beta-1}dy \propto \frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(\alpha+\beta+n)}\ .$$

La figure (6) donne une représentation de l'histogramme d'un échantillon de taille 2000 *simulé par l'algorithme de Gibbs et de la "vraie" densité f(x) qui est une bêta-binomiale pour $n = 16$, $\alpha = 2$ et $\beta = 4$.*

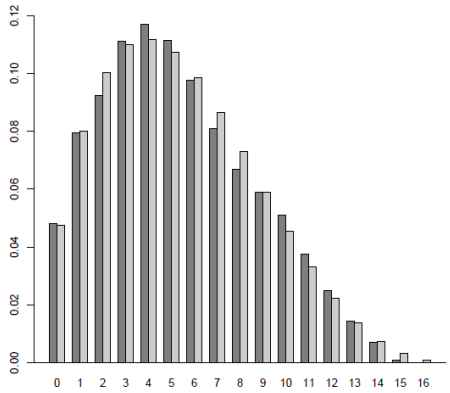


FIG. 6 – *Représentations de l'exacte bêta-binomiale et de l'histogramme d'un échantillon de taille 2000 obtenu par l'algorithme de Gibbs (séquences de 10 itérations) pour $n = 16$, $\alpha = 2$ et $\beta = 4$.*

Dans le cadre bayésien, l'algorithme de Gibbs va permettre d'obtenir une réalisation du paramètre $\theta = (\theta_1,...,\theta_m)$ suivant la loi a posteriori $\pi(\theta \mid x)$ dès que l'on est capable d'exprimer les lois
conditionnelles : $\pi(\theta_i \mid \theta_j \ ; \ x), j \neq i$.
L'échantillonnage de Gibbs consiste à :

Partant d'un vecteur initial $\theta^{(0)} = (\theta_1^{(0)},...,\theta_m^{(0)})$
A la $(p+1)^{\text{ième}}$ étape, disposant du vecteur $\theta^{(p)} = (\theta_1^{(p)},...,\theta_m^{(p)})$,
simuler

$$\begin{aligned}
\theta_1^{(p+1)} &= \pi(\theta_1 \mid \theta_2^{(p)},\theta_3^{(p)},...,\theta_m^{(p)}; x) \\
\theta_2^{(p+1)} &= \pi(\theta_2 \mid \theta_1^{(p)},\theta_3^{(p)},...,\theta_m^{(p)}; x) \\
&... \\
\theta_m^{(p+1)} &= \pi(\theta_m \mid \theta_1^{(p)},\theta_2^{(p)},...,\theta_{m-1}^{(p)}; x)
\end{aligned}$$

Les itérations successives de cet algorithme génèrent successivement les états d'une chaîne de Markov $\{\theta^{(p)}, p > 0\}$ à valeurs $N^{\otimes m}$.
La probabilité de transition de θ' vers θ a pour expression :

$$\begin{aligned}
K(\theta',\theta) = \pi(\theta_1 \mid \theta'_2,...,\theta'_m) \times \pi(\theta_2 \mid \theta_1,\theta'_3,...,\theta'_m) \\
\times \pi(\theta_3 \mid \theta_1,\theta_2,\theta'_4,...\theta'_m) \times ... \times \pi(\theta_m \mid \theta_1,...,\theta_{m-1}).
\end{aligned}$$

On montre que cette chaîne admet une mesure invariante qui est la loi a posteriori. Pour un nombre d'itérations suffisamment grand, le vecteur θ obtenu peut donc être considéré comme étant une réalisation de la loi a posteriori.
Exemple 2.11. *(Loi normale bivariée)*
On désire simuler une loi normale bivariée

$$(X,Y) = x \sim \mathcal{N}(0,\Lambda) \text{ avec } \Lambda = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

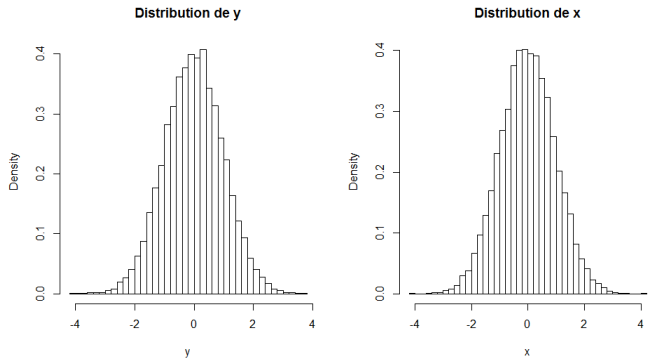
Les lois conditionnelles :

$$Y|X = x \sim \mathcal{N}(\rho x, 1-\rho^2) \text{ et } X|Y = y \sim \mathcal{N}(\rho y, 1-\rho^2)$$

L'implémentation en R

Programme 2.3 Loi normale bivariée

```
N=20000; a=0.2; bn=40; mu=0;
f=function(x){(1/sqrt(pi*2))*\exp(-(1/2)*(x-mu)^2)}
y=x=numeric(1); y[1]=0.5;
x[1]=rnorm(1,a*y[1],1-a^2)
for(i in 2:N){
y[i]=rnorm(1,a*x[i-1],1-a^2)
x[i]=rnorm(1,a*y[i],1-a^2)}
par(mfrow=c(1,2))
hist(y,breaks=bn,prob=TRUE,main=Distribution de Y")
curve(f,add=TRUE,col=blue",lty=1,lwd=2)
hist(x,breaks=bn,prob=TRUE,main=Distribution de X")
```



Exemple 2.12. *Modèle Auto-exponentiel de Besag (1974)*

Soit une densité f, appelée auto-exponentielle définie comme suit

$$f(x_1,x_2,x_3) \propto \exp\{-(x_1 + x_2 + x_3 + \theta_1 x_1 x_2 + \theta_2 x_2 x_3 + \theta_3 x_1 x_3)\}$$

avec θ_1,θ_2,θ_3 connus :
Toutes les lois conditionnelles sont exponentielles :

$$X_1 \mid x_2,x_3 \sim Exp(1 + \theta_1 x_2 + \theta_3 x_3).$$

$$X_2 \mid x_1,x_3 \sim Exp(1 + \theta_1 x_1 + \theta_2 x_3).$$

$$X_3 \mid x_2,x_1 \sim Exp(1 + \theta_2 x_2 + \theta_3 x_1).$$

Nous prenons pour notre cas $\theta_1 = 0.1$, $\theta_2 = 2$ et $\theta_3 = 20$.

Programme 2.4 Modèle Auto-exponentiel

```
N=20000; a=0.1; b=2; c=20; bn=40;
X1=X2=X3=numeric(1);
X3[1]=X2[1]=0.5;
for (i in 2:N){ X1[i]=rexp(1,1+a*X2[i-1]+c*X3[i-1])
X2[i]=rexp(1,1+a*X1[i-1])+b*X3[i-1])
X3[i]=rexp(1,1+c*X1[i-1])+b*X3[i-1]) }
par (mfrow=c(2,2))
hist(X1,breaks=bn,prob=TRUE,main=Distribuction de X1")
hist(X2,breaks=bn,prob=TRUE,main=Distribuction de X2")
hist(X3,breaks=bn,prob=TRUE,main=Distribuction de X2")
```

2.5 Conclusion

L'objet principal de ce chapitre est la présentation des algorithmes basés sur les différentes méthodes et techniques de Monte Carlo, en plus d'étudier les chaînes de Markov.

Nous avons suivi l'ordre chronologique des événements dans l'évolution et l'introduction des différentes approches d'approximation des intégrales, soit analytiquement, soit numériquement ou par simulation. Les critiques faites aux deux premières approches nous orientent sur l'approche simulation dont les méthodes d'échantillonnage préférentiel adaptatif et les méthodes MCMC l'emportent sur les autres techniques que nous avons présentées.

Le chapitre suivant est consacrée à l'adaptation et l'implémentation des applications de la MCMC.

Chapitre 3

3 Application des méthodes MCMC à la statistique Bayésienne

3.1 Introduction

L'émergence des probabilités remonte au 17ème siècle tandis que les premiers travaux de statistique datent du 18ème siècle avec Bayes et Laplace. Il s'agit alors de statistique bayésienne. Au cours du 19ème siècle et du 20ème siècle les méthodes fréquentistes supplantent largement les méthodes bayésiennes. Depuis le début des années 1980, on note un retour trés important de la recherche et des applications des méthodes bayésiennes. On peut se demander pourquoi il a fallu attendre si tard pour que la statistique bayésienne revienne au premier plan. La raison est simple : la statistique bayésienne nécessite souvent des calculs potentiellement lourds ou infaisable lorsque l'on sort des exemples simples, il a donc fallu attendre que des méthodes de résolution numérique soient suffisamment performantes pour permettre d'obtenir des approximations numériques en des temps raisonnables. L'utilisation de méthodes de Monte-Carlo et MCMC pour approximer numériquement les intégrales a permis de sortir du cadre simple des lois conjuguées et d'élargir considérablement le spectre d'application des méthodes bayésiennes.

3.2 Application de l'échantillonneur de Gibbs à l'éstimation Bayésienne dans un cas gaussien

Nous nous intéressons à une analyse Bayésienne d'une suite de variables aléatoires Gaussienne, indépendantes et identiquement distribuées.
En utilisant l'algorithme de Gibbs Sampler, nous retrouvons les densités a posteriori de la moyenne et de la variance.

Présentation du modèle :

Soient X_1,...,X_n une suite de variables aléatoires i.i.d. de moyenne μ et de variance σ^2.

Lois a priori :

Supposons que les lois a priori assignées aux paramètres μ *et* σ^2 sont données respectivement par :

$$f(\mu) \propto \text{constante sur } \mathbb{R}$$

$$f(\sigma^2) \propto \frac{1}{\sigma^2}$$

L'analyse a posteriori :

La fonction de vraisemblance est donnée par :

$$L(\mu,\sigma^2|X) \propto (\sigma^2)^{\frac{-n}{2}} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

La densité jointe a priori est donnée par:

$$f(\mu,\sigma^2) = f(\mu).f(\sigma^2) \propto \frac{1}{\sigma^2}$$

D'où la densité jointe a posteriori:

$$f(\mu,\sigma^2/X) \propto (\sigma^2)^{\frac{-n}{2}-1} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

Les densités a posteriori conditionnelles :

Posons $\tau = \frac{1}{\sigma^2}$.

$$f(\mu|\tau,x) \sim N(\bar{x},\frac{1}{n\tau})$$

$$f(\tau|\mu,x) \sim Gamma\left(\frac{n}{2},\frac{2}{(n-1)S^2+n(\mu-\bar{x})^2}\right).$$

Le programme sous le logiciel R suivant nous donné les densités de μ et σ^2 obtenues par l'algorithme de Gibbs.

Programme 3.1 La loi gaussienne

```
n=30
mu=1
tau=0.5
x=rnorm(n,mu,1/tau)
 xbar=mean(x)
 s2=mean((x-mean(x))^2)
s2
T  = 1000
tau[1] = 1
for(i in 2:11000) {
mu[i]  = rnorm(n = 1, mean = xbar, sd = sqrt(1 / (n * tau[i - 1])))
tau[i] = rgamma(n = 1, shape = n / 2, scale = 2 / ((n - 1) * s2 + n * (mu[i] - xbar)^2))
}
mu  <- mu[-(1:T)]
tau <- tau[-(1:T)]
hist(mu)
hist(tau)
```

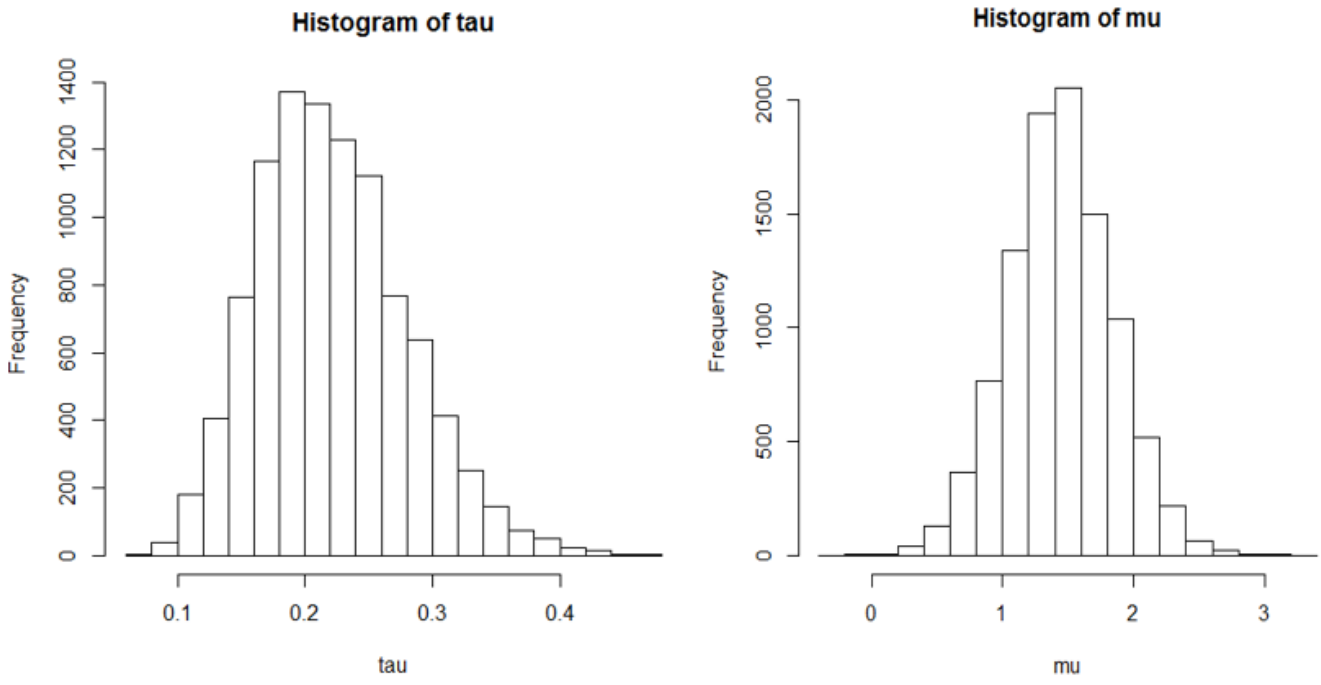


FIG. 7 – *Les distributions de τ et μ obtenues par l'algorithme de Gibbs.*

3.3 Application des méthodes MCMC à la détection des points de rupture

3.3.1 Introduction

L'étude de rupture de modèle est devenue l'un des passages incontournables pour les économètres et les statisticiens. En effet, la stabilité d'un modèle si elle venait à faire défaut doit être décelée sinon elle risque de conduire à des résultats erronés et entrainer des interprétations fausses. Il est donc impératif de s'assurer de la stabilité ou non du modèle étudié. Pour cela, le statisticien est appelé à faire des tests pour détecter un eventuel changement des paramètres de

son modèle, estimer s'il y a lieu le ou les points de rupture, identifier les paramètres sujets à un changement et enfin estimer leur amplitude. L'approche bayésienne à été largement sollicitée dans ce domaine.

3.3.2 Définition d'un modèle de rupture

On dira qu'il y a rupture dans une suite de variables aléatoires X_1,X_2,...X_n s'il existe un entier m; $(1 < m \leq n-1)$; tel que les variables aléatoires X_1,X_2,...,X_m suivent une loi de probabilité de fonction de répartition $F_1(x/\theta_1)$ et les variables aléatoires X_{m+1},...,X_n suivent une autre loi de probabilité de fonction de répartition $F_2(x/\theta_2)$, où θ_1 et θ_2 sont des paramètres inconnus réels ou vectoriels $(\theta_1 \neq \theta_2)$. Le point m est appelé point de rupture ou point de changement. Dans la litterature anglo-saxonne il est connu sous le nom de shift point ou change point.

3.3.3 Détection de rupture dans un modèle de poisson

Nous nous intéressons dans cette section aux travaux de Ilker Yildrim (2012). L'auteur considére une suite de variables aléatoires indépendantes de loi de poisson ayant subit un changement à un instant inconnu. Soit X_1,...,X_n cette suite de paramètre λ_1 pour $i = 1 : r$ et λ_2 pour $i = r+1 : n$.

Le changement concérne le paramètre de la loi, le modèle peut donc s'écrire comme suit:

$$X_i \sim \begin{cases} P(\lambda_1), & \text{i = 1,...,r,} \\ P(\lambda_2), & \text{i = r+1,...,n.} \end{cases}$$

L'analyse bayésienne :

Les lois a priori assignées par l'auteur aux paramètres r,λ_1,λ_2 sont données respectivement par:

$r \propto$ Uniforme(1,n)

$\lambda_i \propto$ Gamma(a,b)

Les lois a posteriori conditionnelles retrouvées par l'auteur sont données par :

$$\log P(\lambda_1|n,\lambda_2) \sim \log Gamma(a + \sum_{i=1}^{r} x_i, r + b).$$

$$\log P(\lambda_2|n,\lambda 1) \sim \log Gamma(a + \sum_{i=r+1}^{n} x_i, n - r + b).$$

$$\log(r|\lambda_1,\lambda_2) \propto \left(\sum_{i=1}^{r} x_i\right) \log \lambda_1 - r\lambda_1 + \left(\sum_{i=r+1}^{n} x_i\right) \log \lambda_2 - (n-r)\lambda_2.$$

Simulation MCMC:

Pour détecter le point de la rupture, l'auteur s'intéresse à la densité a posteriori marginale de r, dont le mode nous donne le point où a eu lieu le changement.
Afin de retrouver cette densité, l'auteur utilise l'echantillonneur de Gibbs. Il considère un échantillon de taille $n = 50$, avec $a = 2$ et $b = 1$.

Les résultats obtenus sont donnés par le graphe suivant:

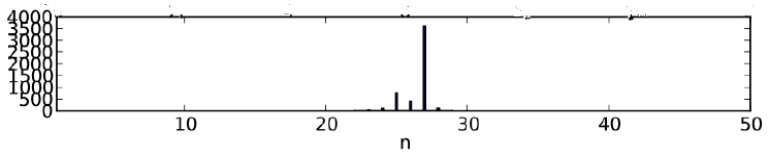


FIG. 8 – *Densité a posteriori de r*

La figure nous donne la densité a posteriori marginale de r. Nous remarquons que le mode est égale à 26. Donc nous avons une rupture au point $r = 26$.

3.3.4 Exemple de rupture sur des données réelles

[Antoch Jaromir, Praha, David Legat, Praha, Application of MCMC To Change Point Detection]

Klementinum est un bâtiment historique dans la partie centrale de Prague où la température a été mesurée depuis (1775), grâce à une station météorologique, placée dans l'une de ses tours.Les auteurs s'interessent aux moyennes annuelles de ses températures durant la période de 1775 à 1992.

Tour astronomique de Klementinum (2008).

A fin de détecter un éventuel changement dans la moyenne des températures, les auteurs considèrent le modele suivant :

Modèle: Soit Z_1,...,Z_N une suite de variables aléatoires de moyenne μ_i pour $i = 1 : N$et de variance σ^2.

Le modèle peut donc s'écrire comme suit :

$$(1)\ Z_i \sim \begin{cases} \mathrm{N}(\mu_1,\sigma^2), & 1 \leq i \leq r, \\ \mathrm{N}(\mu_2,\sigma^2), & r < i \leq N, \end{cases}$$

où r,μ_1,μ_2 et σ^2 sont des paramètres inconnus

Les lois a priori

Les lois a priori proposées par les auteurs sont :

On pose $\gamma = \dfrac{1}{\sigma^2}$.
Pour $r \sim u[1,N]$

$\pi_0(r) \sim u[1,N]$

$\pi_0(\mu_1) \sim N(\nu_1,\xi_1^2)$

$\pi_0(\mu_2) \sim N(\nu_2,\xi_2^2)$

$\pi_0(\gamma) \sim Ga(1,1)$.

Avec $\nu_1 = \nu_2 = 9.5$ et $\xi_1^2 = \xi_2^2 = 1$.

La densité jointe a priori :

μ_1,μ_2,γ,r est donnée par

$$f(\mu_1,\mu_2,\gamma,r) \propto \exp\{-\frac{1}{2}(\mu_1-\nu_1)^2/\xi_1^2 - \frac{1}{2}(\mu_2-\nu_2)^2/\xi_2^2 - \gamma\}.$$

La fonction de vraisemblance des Z_i

de la séquence Z_1,...,Z_N pour des valeurs données de paramètres est

$$f(z_1,...,z_N|\mu_1,\mu_2,\gamma,r) \propto \gamma^{N/2}\exp\left(-\frac{\gamma}{2}\left[\sum_{i=r}^{r}(z_i-\mu_1)^2 + \sum_{i=r+1}^{N}(z_i-\mu_2)^2\right]\right),$$

La densité jointe a posteriori

Par le théorème de Bayes, on retrouve la densité a posteriori jointe des paramètres

$$f_{\mu_1,\mu_2,\gamma,r|z} \propto \gamma^{N/2}\exp\left(-\frac{(\mu_1-\nu_1)^2}{2\xi_1^2} - \frac{(\mu_2-\nu_2)^2}{2\xi_2^2} - \gamma - \frac{\gamma}{2}\left[\sum_{i=1}^{r}(z_i-\mu_1)^2 + \sum_{i=r+1}^{N}(z_i-\mu_2)^2\right]\right),$$

Après avoir calculer les densités conditionnelles, les auteurs utilisent l'échantillonneur de Gibbs pour trouver la distribution a posteriori du paramètre r.

Les résultats obtenus sont donnés par la figure suivante :

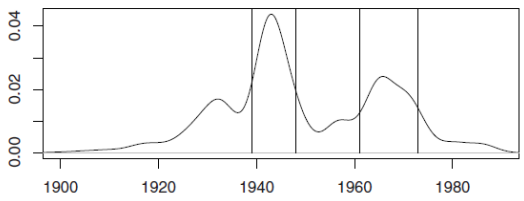


FIG. 9 – *La densité a posteriori de r.*

Les auteurs se sont intéressé, a la distribution a posteriori du paramètre r. Le mode de cette dérnière nous détérmine le point ou a eu lieu le changement. Nous remarquons que la rupture s'est produite en 1943.

Conclusion générale

L'analyse bayésienne constitue un développement important en statistique ces dernières années. L'utilisation de méthodes bayésiennes peut présenter certains avantages. En effet, ces méthodes utilisent des connaissances antérieures, exprimées sous forme de distribution de probabilité, afin de modifier une information nouvelle. L'utilisation de lois de distributions, et surtout de leurs propriétés.

Le développement des algorithmes itératifs possédant des propriétés markoviennes a permis de surmonter cet obstacle dans la mesure où ceux-ci permettent d'échantillonner à partir de la distribution a posteriori, sans qu'il soit nécessaire de la spécifier explicitement, et assurent la convergence en distribution vers la distribution a posteriori. L'échantillonneur de Gibbs (1984) et l'algorithme de Metropolis-Hastings (1953) sont les algorithmes stochastiques de Monte-Carlo par Chaîne de Markov les plus utilisés dans les méthodes d'analyse bayésienne.

Cependant, l'inférence bayésienne est largement sollicitée et donne de bons résultats dans beaucoup de disciplines, notamment en physiques, en économie, en finances et en medecine. Les méthodes Monte Carlo par chaines de Markov apportent une grande souplesse à la méthodologie bayésienne.

Résumé

Dans ce mémoire, nous étudions les méthodes de Monte-Carlo, application de bayésienne.

Dans le premier chapitre, nous présentons l'analyse statistique bayésienne, nous montrons dans ce chapitre la théorie de Bayes, puis les lois a priori, en plus les bases de la théorie de la décision qui contient l'estimation de Bayes.

Dans le chapitre 2, les méthodes d'estimation Monte-Carlo avec un bref aperçu des méthodes de simulation par chaînes de Markov (MCMC). Nous nous intéressons en particulier aux deux algorithmes qui sont utiles pour ces méthodes: il s'agit de l'algorithme Metropolis-Hastings et la méthode d'échantillonnage de Gibbs. Cette dernière a été utilisée par Geman et Geman (84) pour générer des observations à partir d'une distribution de Gibbs . Il s'agit d'une forme particulière de méthode de Monte-Carlo par chaîne de Markov qui, du fait de son efficacité, est largement utilisée dans de nombreux domaines d'analyse statistique bayésienne.

Dans le chapitre 3, nous verrons l'application des méthodes MCMC à la détection des points de rupture, la détection des points de changement est un sujet d'intérêt pour de nombreuses statistiques appliquées et théoriques. depuis les années soixante-dix.

Mots-clés: L'analyse statistique Bayésienne, méthodes de Monte-Carlo par chaîne de Markov, échantillonnage de Gibbs, algorithme de Metropolis-Hastings.

Références

[1] Antoch Jaromir, Praha, David Legat, Praha,*Application of MCMC To Change Point Detection* journal APPLICATION OF MATHEMATICS(2008)

[2] Brown R.L., *Notes of statistical decision theory of conditional confidence procedures*, Ann. statist.,6, 56-71, (1976).

[3] Eckhardt. R. Stam Ulam, John Von Neumann and the Monte Carlo méthod. *Los Alamos Science,*Special Issue :131-136, (1987).

[4] Gauss C., *Méthodes des moindres carrés*, Mémoire sur la combinaison des obsérvations, Mallet-Bachelier, Paris. transl. J.Bertrand, (1810).

[5] Geman, S. et Geman, D. 1984. Stochastiic relaxation, Gibbs distribution and the Bayesian restoration of images., *IEEE Transaction on Pallern Analysis and Machine Intelligence*, 6 : 721·741.(1984).

[6] Hasting W., *Monte Carlo simpling methods using Markov chains and their application*, Biometrika, 57,97-109, (1970).

[7] Hwang J.et Pamantle R., *Confidence sets recenterd at janes- stein estimators- a surprise concerning the unknown varianece case*, Econometrics, 60(1-2), 145-156, (1994).

[8] Iker yaldinim., *Bayésienne inference, Gibbis Sampling*, Springer-New york (2012).

[9] Irwin. M.E. *Markov chains*, URLhttp://www.markirwin.net /statllO/Lecture/MarkovChains.pdf.(2006).

[10] Le cam L., *Asymptotic Methode in statistical Decision Theory* , Springer- Verlag, New york, (1986).

[11] Legendre A., *Nouvelles méthodes pour la determination des orbites des cométes*, Courcier,Paris, (1805).

[12] Lessard., *Cours de processus stochastiques*, (2013). URL http://www. dms. umontreal.ca/-lessards/Processus Stochastiques Lessard (2014).pdf.

[13] Lindley D., *Marking Decision*, John Wiley, New york, (1985).

[14] Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. (1953) Equations of state calculations by fast computing machines., *Journal of Chemical Physics*, 21 : 1087-1092, (1953).

[15] Nicholas Metropolis, *The Beginning of the Monte Carlo Method*, Los Alamos Science, p. 125-130, (1947).

[16] Nicholas Metropolis et Stanislaw Ulam, *The Monte Carlo Method*, Journal of the American Statistical Association, p. 335-341, septembre (1949).

[17] Raiffa H., Schalaifer R., *Applied statistical decision theory*, Technical report, Division of Research, Graduate School of Buisiness Administration, Havard Univ, (1961).

[18] Robert C.P., *L'Analyse Stalislique bayésienne*, Statistique mathématique et probabilité. Economica, (1992).

[19] Robert C.P., *le Choix Bayésien - Principe et partique*, Saringer, (2006).

[20] Strasser H., *Mathematical Theory of Statistics*, W. de Grayter, Berlin, (1985).

[21] Taylor. H. M et S. Karlin., *An Introduction to Stochastic Modeling. Academie Press*, (1998).