

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA
RECHERCHE SCIENTIFIQUE
UNIVERSITÉ Abderrahmane Mira - Béjaïa
FACULTÉ DES SCIENCES EXACTES
Département de Mathématiques

Mémoire présenté pour l'obtention du diplôme de Master en Mathématiques

Option : Statistique et Analyse Décisionnelle



THÈME

Méthodes de Monte-Carlo par chaînes de Markov,
application à l'estimation bayésienne

Présenté par:

- M^r KADDOUR Akhalaf.
- M^{elle} MOUMEN Aïcha.

Soutenu publiquement, le 27/06/2016 devant le jury composé de :

M ^{elle}	L. BAICHE	M.A.B	Université de Béjaïa.	Présidente.
M ^{me}	L. BOURAINE	M.C.A	Université de Béjaïa.	Promotrice.
M ^f	Y. BOUMZAID	M.A.A	Université de Béjaïa.	Examineur.
M ^{elle}	H. BRAHAM	Doctorante	Université de Béjaïa.	Examinatrice.

Promotion: 2015/2016

Remerciements

Nous tenons tout d'abord à remercier le bon Dieu le tout puissant de nous avoir accordé la force et le courage afin d'accomplir ce modeste travail.

Nous exprimons notre gratitude à notre promotrice M^{me} L. BOURAINE pour son aide très précieuse, ses conseils et sa disponibilité qui ont contribué à faire de ce mémoire ce qu'il est aujourd'hui.

Nos remerciements sont aussi adressés à M^r Y. BOUMZAID et M^{elle} L. BAICHE et H. BRAHAM qui nous font l'honneur de juger notre travail. Nous remercions aussi tous les enseignants du département de Mathématiques qui ont contribué à notre formation.

Enfin, nos profonds remerciements vont à nos chers parents, frères et sœurs, pour leur soutien et leur confiance en nous, sans oublier nos amis et nos camarades de tous les niveaux, et tous ceux qui ont contribué, de près ou de loin à notre formation.

Grand Merci.....

Dédicaces

Je dédie ce travail à :

- Mes chers parents qui m'ont toujours soutenu durant mes études.
- Mon frère Abderezak.
- Mes deux sœurs Meriem et Aicha.
- Ma belle aimée sans citer de nom.
- Mon beau frère Kousaila et sa famille.
- Mes amis.

A'ALAF

Dédicaces

A MES PARENTS (MA MÈRE, MA MÈRE, MA MÈRE, MON PÈRE)

A MES FRÈRES ET SOEURS

A MES NEVEUX ET MES NIÈCES

A TOUTE MA FAMILLE

A MES AMIS

AICHA

Liste des abréviations

$L(\underline{x}|\theta)$: La loi des observations (vraisemblance).

$\pi(\theta)$: La loi a priori.

$\pi(\theta|\underline{x})$: La loi a posteriori.

F : Distribution cible de densité f .

G : Distribution instrumentale de densité g .

i, j : Indices généraux.

i : Indice des itérations d'un algorithme.

n : Nombre total d'itérations.

X : Variable aléatoire.

x : Réalisation de la variable aléatoire X .

χ : Domaine de la variable aléatoire X .

\underline{x} : Vecteur des observations.

$\mathcal{U}_{[0,1]}$: Distribution uniforme défini sur l'intervalle unité.

U : Variable aléatoire distribuée selon $\mathcal{U}_{[0,1]}$.

$\mathcal{N}(0, 1)$: Distribution gaussienne centrée réduite.

$\mathcal{N}(\mu, \sigma^2)$: Distribution gaussienne de moyenne μ et de variance σ^2 .

$\mathcal{Be}(\alpha, \beta)$: Distribution d'une loi Beta de paramètres α et β .

$\mathcal{IG}(\alpha, \beta)$: Distribution d'une loi inverse Gamma de paramètre α et β .

Θ : Espace des paramètres, espace des états d'une chaîne de Markov.

$\theta \in \Theta$: Paramètre, état d'une chaîne de Markov

$Supp f$: Support de la fonction f .

MCMC : Méthodes de Monte-Carlo par chaînes de Markov.

M-H : Algorithme de Metropolis-Hastings.

NC : Longueur de la période de chauffage de la chaîne de Markov.

NI : Nombre d'itérations de la chaîne de Markov construite pour réaliser l'estimation.

Liste des tableaux

Tableau 1-1-1 : Lois a priori conjuguées naturelles	p 6.
Tableau 4-1-1 : Facteur d'échelle pour la probabilité p	p 49.
Tableau 4-2-1 : Taille (cm) de 15 garçons âgés de 10 ans	p 50.
Tableau 4-2-2 : Facteur d'échelle pour la moyenne m	p 54.
Tableau 4-2-3 : Facteur d'échelle pour la variance σ^2	p 55.

Liste des figure

Figure 2-1-1 : Distribution d'un échantillon généré à partir de la loi $\mathcal{E}(0.5)$	p 17
Figure 2-1-2 : Simulation d'une loi normale à partir de la loi de Laplace.....	p 19
Figure 3-1-1 : L'évolution de la chaîne suivant la loi de $\mathcal{P}(10)$	p 35
Figure 3-1-2 : Distribution et histogramme de la chaîne simulée suivant $\mathcal{P}(10)$	p 35
Figure 4-1-1 : Distribution de la loi de vraisemblance $\mathbb{P}(x = 7 p)$	p 47
Figure 4-1-2 : L'évolution de la chaîne du paramètre p	p 48
Figure 4-1-3 : Distribution et histogramme de la chaîne du paramètre p	p 49
Figure 4-1-4 : Convergence de l'estimateur de la probabilité p	p 49
Figure 4-2-1 : Distribution de la taille d'un garçon agés de 10 ans.....	p 50
Figure 4-2-2 : Evolution de la moyenne m	p 53
Figure 4-2-3 : Densité et histogramme de la moyenne m	p 53
Figure 4-2-4 : Evolution de la chaîne de la variance σ^2	p 54
Figure 4-2-5 : Densité et histogramme de la chaîne de la variance σ^2	p 54
Figure 4-2-6 : Convergence de l'estimateur de la moyenne m	p 55
Figure 4-2-7 : Convergence de l'estimateur de la variance σ^2	p 55

Table des matières

Introduction générale	1
1 Rappels sur la statistique bayésienne	4
1.1 Modèle bayésien	5
1.1.1 Distribution a priori	5
1.1.2 Distribution a posteriori	7
1.2 Modèle bayésien hiérarchique	8
1.3 Inférence bayésienne décisionnelle	9
1.3.1 Fonction de perte et risque bayésien	9
1.3.2 Estimateur de Bayes	10
1.4 Conclusion	11
2 Méthodes de Monte-Carlo	13
2.1 Méthodes usuelles de simulation	13
2.1.1 Générateur de la loi uniforme[0,1] ($\mathcal{U}_{[0,1]}$)	14
2.1.2 Simulation de lois non uniformes (Méthodes générales)	15
2.2 Méthodes de Monte-Carlo	20
2.2.1 Intégration de Monte-Carlo	20
2.3 Réduction de variance	21
2.3.1 L'échantillonnage préférentiel	21
2.3.2 Variable de contrôle	23
2.4 Conclusion	24

3	Les méthodes de Monte-Carlo par Chaînes de Markov (MCMC)	25
3.1	Généralités sur les chaînes de Markov	25
3.1.1	Matrice de transitions	26
3.1.2	Noyau d'une chaîne de Markov	27
3.1.3	Propriétés fondamentales	27
3.1.4	Classification des chaînes de Markov	30
3.2	Méthodes de Monte-Carlo par Chaînes de Markov (MCMC)	31
3.2.1	Algorithme de Metropolis-Hastings	32
3.2.2	Echantillonnage de Gibbs	38
3.3	Contrôle de convergence des méthodes MCMC	41
3.3.1	Méthodes de comparaison inter-intra chaînes	41
3.3.2	Méthodes non paramétriques	43
3.3.3	Méthodes graphiques.	44
3.4	Difficultés des méthodes MCMC	45
3.5	Conclusion	45
4	Application des méthodes MCMC à l'estimation bayésienne	46
4.1	Estimation de la probabilité de tomber sur <i>face</i>	46
4.1.1	Algorithme	48
4.1.2	Résultats	48
4.1.3	Etude de convergence	49
4.1.4	Discussion des résultats	50
4.2	Estimation de la taille d'un garçon âgé de 10 ans	50
4.2.1	Test d'ajustement	50
4.2.2	Approche bayésienne	51
4.2.3	Estimation des paramètres	52
4.2.4	Discussion des résultats	56
	Conclusion générale	57
	Annexe	59

Bibliographie

65

Introduction générale

L'homme est curieux et c'est sans doute ce qui explique le mieux son cheminement depuis le début de l'humanité jusqu'à nos jours. Ce besoin de comprendre les phénomènes observés et le désir de les anticiper est au cours de ses préoccupations. C'est ce qui explique l'émergence et le succès de la statistique, une discipline scientifique en plein essor. Pour définir son objet, on ne saurait mieux faire que Christian Robert (Robert, 2006) : *"L'objet principal de la statistique est de mener, grâce à l'observation d'un phénomène aléatoire, une inférence sur la distribution probabiliste à l'origine de ce phénomène, c'est-à-dire de fournir une analyse (ou une description) d'un phénomène passé, ou une prédiction d'un phénomène à venir de nature similaire "* [2].

En statistique, un défi récurrent est d'obtenir des descriptions adéquates d'une variable aléatoire étant donné sa fonction de densité. En effet, il arrive souvent qu'il y ait un intérêt pour l'espérance d'une variable aléatoire, pour sa variance, ses quantiles ou toute autre mesure significative. Bien certainement, il existe une multitude d'approches à ces problèmes et chaque contexte particulier déterminera la meilleure solution.

Afin de répondre à certaines de ces questions, il est possible d'utiliser des méthodes d'intégration numérique, aussi appelées quadratures. Dans le cas unidimensionnel, il existe des méthodes classiques telles que l'algorithme de Newton-Cotes ou des approches plus puissantes telles que la quadrature de Gauss-Legendre. Ces méthodes peuvent être modifiées afin d'accommoder les cas multidimensionnels qui sont beaucoup plus fréquents, par exemple en traitant une intégrale multiple comme une série d'intégrales unidimensionnelles. D'un autre côté, il est possible d'utiliser certains algorithmes développés récemment tel que la méthode introduite par Genz (1972) [11]. Ces approches se comportent

de façon excellente lorsque la dimension de l'intégrale reste relativement petite, mais sont peu utiles en dimension modérée ou grande. En effet, la quantité de ressources nécessaires augmente de façon exponentielle avec la dimension du problème, menant à une durée d'exécution déraisonnable au-delà d'une dimension de 4 ou 5.

Il est également possible d'employer des méthodes de simulation, aussi appelé méthodes de Monte Carlo. L'idée de base est d'obtenir un échantillon aléatoire de la distribution d'intérêt et ensuite d'estimer les quantités voulues de façon empirique en se servant de l'échantillon généré. Cette approche repose sur deux conditions qui sont la facilité de génération des valeurs de la distribution en question et la capacité de produire un grand échantillon afin d'obtenir des résultats fiables. Il existe un nombre considérable d'algorithmes de simulation, comme entre autres, la méthode du rejet ou la méthode d'échantillonnage d'importance (échantillonnage préférentiel).

Ces méthodes consistent donc à simuler une variable aléatoire de loi donnée, mais il y a des cas où cette loi n'est pas facilement simulable. La méthode de rejet peut aussi conduire à des simulations lentes si la probabilité de rejet est grande. Les méthodes de Monte-Carlo par chaînes de Markov vont permettre de pallier à ces problèmes, ces méthodes constituent une des approches les plus utilisées dans la communauté statistique. Elle emploient une chaîne de Markov auxiliaire dont la distribution stationnaire est la distribution d'intérêt.

L'algorithme Metropolis-Hastings, introduit par Metropolis et al. (1953) [18] et généralisé par Hastings (1970) [14], est considéré comme l'une des premières méthodes MCMC. Depuis, le nombre de tels algorithmes ainsi que le nombre de publications portant sur leur convergence et leur application a augmenté de façon remarquable. Le principal attrait des approches MCMC est leur facilité d'application à des distributions d'intérêt complexes et/ou en grandes dimensions. En plus, un autre grand avantage est le fait que la constante de normalisation de la densité d'intérêt ne doit généralement pas être spécifiée. Cet aspect représente un attrait important pour la communauté bayésienne, car cette constante est souvent difficilement obtainable pour une fonction de densité a posteriori. Les algorithmes de simulation autre que les méthodes MCMC requièrent, en général, au moins une approximation ou des bornes pour cette valeur, rendant leur application

plus ardue. Pour ces raisons, les algorithmes MCMC ont largement gagné en popularité dans les cinquante dernières années, laissant quelque peu en arrière-plan les méthodes d'échantillonnage classiques ou bien les méthodes par quadrature.

Parmi les questions auxquelles on va répondre à travers ce travail, nous citons :

- Etant donnée un échantillon aléatoire dont la loi de probabilité dépend d'un ou de plusieurs paramètres inconnus, le problème est comment estimer ces paramètres ?
- Ya-t-il un moyen de s'assurer de la convergence vers la valeur exacte ?

Ce travail est une application des méthodes de Monte-Carlo par chaînes de Markov à la statistique bayésienne pour l'estimation des paramètres, après cette introduction générale, ce mémoire est organisé comme suit

Le premier chapitre donne un rappel de quelques notions préliminaires de la statistique bayésienne telles que la loi a posteriori et les différentes approches de construction d'une loi a priori dans le cas de la disponibilité et la non disponibilité de l'information a priori.

Le deuxième chapitre concerne quelques méthodes de simulation classiques, les méthodes de Monte-Carlo et quelques méthodes de réduction de variance.

Le troisième chapitre est une description générale de l'approche MCMC. Il sera question de la construction de tels algorithmes, de leur convergence. Certains algorithmes de base seront décrits en détails et plusieurs méthodes de contrôle de convergence seront présentées.

Le dernier chapitre porte sur une application des algorithmes MCMC à l'estimation de Bayes, en donnant quelques exemples théoriques et une application sur des données réelles.

La conclusion synthétise les points essentiels abordés dans ce mémoire et donne quelques perspectives de recherche future.

Les programmes de simulation élaborés sous le logiciel **R** sont donnés en annexe.

Rappels sur la statistique bayésienne

Introduction

Soit (X_1, X_2, \dots, X_n) un n -échantillon issu d'une variable aléatoire (*v.a*) X dont la loi dépend d'un paramètre inconnu θ .

Le but de l'analyse statistique est de faire de l'inférence sur θ (utiliser les observations pour améliorer l'information sur le paramètre) et ce afin de pouvoir prendre une décision.

Si toute l'information à propos de θ , est celle qui est apportée par les observations, on peut appliquer la méthode du maximum des vraisemblance, et on interprète la loi des observations (vraisemblance) comme la loi conditionnelle des observations $\underline{x} = (x_1, x_2, \dots, x_n)$ sachant θ . On note sa densité par $L(\underline{x}|\theta)$ qui est définie comme suit :

$$L(\underline{x}|\theta) = \begin{cases} f_{\theta}(\underline{x}) = f(\underline{x}|\theta) = \prod_{i=1}^n f(x_i, \theta) & \text{cas continu} \\ \mathbb{P}_{\theta}(\underline{x}) = \mathbb{P}(\underline{X} = \underline{x}|\theta) = \prod_{i=1}^n \mathbb{P}(X_i = x_i, \theta) & \text{cas discret} \end{cases}$$

L'approche bayésienne consiste à utiliser cette information (vraisemblance) pour l'estimation du paramètre inconnu θ . On considère le cas où on a des connaissances a priori sur le paramètre θ inconnu.

1.1 Modèle bayésien

Définition 1.1.1 *Un modèle statistique bayésien est la double donnée d'un modèle paramétrique $\{f_\theta(\underline{x}), \theta \in \Theta\}$ et une loi de probabilité, de densité π , dite aussi loi a priori qui est la loi marginale de la variable aléatoire θ [21].*

1.1.1 Distribution a priori

La spécification d'un modèle statistique bayésien se décompose en général en deux éléments: le choix de la famille des probabilités d'échantillonnage et celui de la distribution a priori sur les paramètres. La détermination du modèle d'échantillonnage n'est pas spécifique à l'approche bayésienne. La détermination de la distribution a priori est, au contraire, l'essence de l'analyse bayésienne.

Très schématiquement, on peut mettre en évidence deux modes de pensée bayésiens. Le premier est *subjectif* et qui considère que la distribution a priori traduit les connaissances avant l'observation des données. Cette information est exprimée par une loi de probabilité dite "*distribution a priori informative*".

Cependant, la théorie bayésienne peut être appliquée même dans le cas où on ne dispose pas d'information a priori; c'est le deuxième mode de pensée qui est plutôt "*objectif*". Là encore, le paramètre est considéré comme une valeur d'une variable aléatoire et on lui assigne une loi de probabilité dite "*distribution a priori non informative*".

a) Distribution a priori informative

La modélisation a priori informative est le point le plus délicat de l'analyse bayésienne. Il existe plusieurs procédés pour obtenir des lois informatives. On donne la description de l'un des plus intéressants des procédés qui est celui des familles naturelles conjuguées.

Familles naturelles conjuguées

Définition 1.1.2 *Une famille G de lois a priori pour θ est dite conjuguée pour la vraisemblance $L(\underline{x}, \theta)$ si pour tout $\pi(\theta) \in G$, la loi a posteriori $\pi(\theta|\underline{x})$ est un élément de G .*

L'approche conjuguée reste la solution la plus standard dans le cadre informatif. Une famille naturelle conjuguée est un concept qui simplifie considérablement le calcul des distributions a posteriori. C'est la raison du développement de ce type de lois a priori.

Il faut noter que seuls les modèles à structures exponentielles admettant une famille conjuguée.

Exemple 1.1.1 *Lois a priori conjuguées naturelles pour quelques familles exponentielles usuelles.*

$f(x \theta)$	$\pi(\theta)$	$\pi(\theta x)$
Normale $\mathcal{N}(\theta, \sigma^2)$	Normale $\mathcal{N}(\mu, \tau^2)$	$\mathcal{N}(\varrho(\sigma^2\mu + \tau^2x), \varrho\sigma^2\tau^2)$ $\varrho^{-1} = \sigma^2 + \tau^2$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	$\mathcal{G}(\alpha + x, \beta + 1)$
Gamma $\mathcal{G}(\nu, \theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	$\mathcal{G}(\alpha + \nu, \beta + x)$
Binomiale $\mathcal{B}(n, \theta)$	Bêta $\mathcal{Be}(\alpha, \beta)$	$\mathcal{Be}(\alpha + x, \beta + n - x)$
Binomiale Négative $\mathcal{Neg}(m, \theta)$	Bêta $\mathcal{Be}(\alpha, \beta)$	$\mathcal{Be}(\alpha + m, \beta + x)$
Multinomiale $\mathcal{M}_k(\theta_1, \dots, \theta_k)$	Dirichlet $\mathcal{D}(\alpha_1, \dots, \alpha_k)$	$\mathcal{D}(\alpha_1 + x_1, \dots, \alpha_k + x_k)$
Normale $\mathcal{N}(\mu, 1/\theta)$	Gamma $\mathcal{Ga}(\alpha, \beta)$	$\mathcal{G}(\alpha + 0.5, \beta + (\mu - x)^2/2)$

Tableau 1.1.1 : Lois a priori conjuguées naturelles

b) Distribution a priori non informative

Ces distributions sont conçues dans le but de faire de l'analyse bayésienne et ceci lorsqu'il y a absence de l'information a priori sur le paramètre d'intérêt ou dans le cas où c'est difficile de traduire en terme de loi a priori l'information disponible sur les paramètres par une loi de probabilité.

Le choix d'une loi a priori non informative conduit souvent à la spécification d'une mesure et non d'une probabilité, c'est-à-dire une loi impropre .

Une loi impropre est une loi qui est définie sur l'espace \mathbb{R} de densité π non intégrable.

$$\int_{\mathbb{R}} \pi(\theta) d\theta = +\infty.$$

Les lois a priori impropres forment une extension nécessaire pour les lois a priori propres. Dans un contexte non informatif, elles permettent en général une meilleure modélisation de l'absence d'information a priori.

Mesures a priori de Jeffreys [7]

Le mode de spécification d'une loi a priori non informative connue sous le nom de la mesure a priori de Jeffreys consiste à assigner à un modèle d'échantillonnage caractérisé par sa vraisemblance $L(\underline{x}|\theta)$, la mesure a priori de densité :

$$\pi_J(\theta) \propto [\det I(\theta)]^{\frac{1}{2}}.$$

où θ est un vecteur de paramètres inconnus et $I(\theta)$ est la matrice de l'information de Fischer définie par:

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j} \ln L(\underline{x}|\theta)\right].$$

1.1.2 Distribution a posteriori

Soit $X = (X_1, \dots, X_n)$ un n -échantillon de variables aléatoires indépendantes et identiquement distribuées (*i.i.d*) de densité $f(\underline{x}|\theta)$. On suppose que θ est une valeur d'une variable aléatoire θ de densité $\pi(\theta)$ et soit la fonction de vraisemblance des observations.

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

la distribution a posteriori de θ notée $\pi(\theta|\underline{x})$ représente la distribution conditionnelle de θ sachant les observations $\underline{x} = (x_1, \dots, x_n)$.

$$\pi(\theta|\underline{x}) = \frac{L(\underline{x}|\theta)\pi(\theta)}{\int_{\Theta} L(\underline{x}|\theta)\pi(\theta)d\theta} = \frac{\prod_{i=1}^n f(x_i|\theta)\pi(\theta)}{\int_{\Theta} \prod_{i=1}^n f(x_i|\theta)\pi(\theta)d\theta}.$$

Où encore, on peut écrire :

$$\pi(\theta|\underline{x}) \propto \prod_{i=1}^n f(x_i|\theta)\pi(\theta),$$

où \propto signifie "proportionnelle à".

La distribution a posteriori $\pi(\theta|\underline{x})$ donne alors l'information dont on dispose sur θ après observation. Elle représente un compromis entre l'information a priori (donnée par π) et l'information tirée de l'échantillon X (donnée par $L(\underline{x}|\theta)$).

1.2 Modèle bayésien hiérarchique

Définition 1.2.1 [21] *Un modèle bayésien hiérarchique est un modèle statistique bayésien $(f_{\theta}(\underline{x}), \pi(\theta))$, dans lequel la loi a priori $\pi(\theta)$ est décomposée en plusieurs lois conditionnelles $\pi_1(\theta|\theta_1)$, $\pi_2(\theta_1|\theta_2)$, ... , $\pi_n(\theta_{n-1}|\theta_n)$, et une loi marginale $\pi_{n+1}(\theta_n)$ telle que:*

$$\pi(\theta) = \int_{\Theta_1 \times \dots \times \Theta_n} \pi_1(\theta|\theta_1)\pi_2(\theta_1|\theta_2)\dots\pi_{n+1}(\theta_n)d\theta_1\dots d\theta_{n+1}.$$

Les paramètres θ_i sont appelés hyperparamètres de niveau i ($1 \leq i \leq n$).

Cette réduction illustre l'aspect « classique » des modèles hiérarchiques et montre donc que cette approche particulière bénéficiera des propriétés générales de l'approche bayésienne avec, en plus, des avantages propres liés à la décomposition de la loi a priori. L'approche bayésienne hiérarchique tend à modéliser le manque d'information sur les paramètres de la loi a priori par une nouvelle distribution sur ces paramètres. Bien que conceptuellement plus abstraite, cette méthode est souvent préférée à une approche bayésienne empirique qui utilise les observations et la loi marginale pour estimer les paramètres de la loi a priori.

La technique bayésienne d'échantillonnage qui utilise les structures hiérarchiques du modèle bayésien est fondée sur les travaux de Gelfand et Smith [10]. On évalue $\pi(\theta|\underline{x})$ de la manière suivante :

$$\pi(\theta|\underline{x}) = \int_{\Theta_1} \pi_1(\theta|\underline{x}, \theta_1) \pi_2(\theta_1) d\theta_1.$$

Si les lois $\pi_1(\theta|\underline{x}, \theta_1)$, $\pi_2(\theta_1)$ sont connues, générer θ suivant $\pi(\theta|\underline{x})$ revient en fait à générer θ_1 suivant $\pi_2(\theta_1)$ puis θ suivant $\pi_1(\theta|\underline{x}, \theta_1)$.

1.3 Inférence bayésienne décisionnelle

L'ensemble des décisions possibles D est appelé espace des décisions et la plupart des exemples théoriques se concentrent sur le cas $D = \Theta$ qui représente le cadre d'estimation standard.

Une inférence bayésienne est fondée sur la détermination rigoureuse de trois facteurs:

1. La loi des observations $L(\underline{x}|\theta)$.
2. La distribution a priori des paramètres $\pi(\theta)$.
3. Le coût associé aux décisions $l(\theta, d)$.

On note que les critiques fréquentes de l'approche bayésienne se font souvent sur le point 2, alors que, conceptuellement, les points 1 et 3 se trouvent sur le même rang et que le point 3, qui est un point commun entre l'approche classique et l'approche bayésienne, est encore plus subtil. Le choix d'une fonction de coût dépend de l'application concernée. Lorsque le contexte ne conduit pas à la détermination de la fonction d'utilité qui permet d'en déduire la fonction de coût, on peut avoir recours à des fonctions de coûts classiques, qui sont à la fois simples et bien étudiées.

1.3.1 Fonction de perte et risque bayésien

Fonction de perte

C'est la fonction notée l de $\Theta \times D$ dans $[0, +\infty[$

$$l : \Theta \times D \rightarrow \mathbb{R}^+ \\ (\theta, d) \rightarrow l(\theta, d)$$

$l(\theta, d)$ évalue le coût d'une décision d lorsque le paramètre vaut θ .

Coût quadratique Parmi les fonctions de perte les plus utilisées on peut noter le coût quadratique

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

qui pénalise trop fortement les grandes erreurs, du fait de sa convexité stricte.

Fonction de risque

La fonction de risque est l'espérance mathématique de fonction de perte composée avec une fonction de décision [21].

$$R(\theta, d) = \mathbb{E}_\theta[l(\theta, d(x))]$$

Risque bayésien

Définition 1.3.1 Soit $T = T(x_1, \dots, x_n)$ un estimateur de θ et $l(t, \theta)$, la fonction de perte associée. On appelle risque bayésien moyen noté $R_T(\theta)$, la quantité:

$$R_T(\theta) = \mathbb{E}_\theta[l(\theta, t)] = \int_{\Theta} l(\theta, t) L(\underline{x}|\theta) d\underline{x}.$$

1.3.2 Estimateur de Bayes

Définition 1.3.2 La règle de Bayes associées à la loi a priori π est donnée par la solution du système:

$$\min_d \mathbb{E}^\pi[l(\theta, d)|\underline{x}]. \quad (1.3.1)$$

On appelle estimateur de Bayes toute fonction δ^π qui associe à \underline{x} une solution de (1.3.1).

Admissibilité

Définition 1.3.3 *Un estimateur δ est inadmissible s'il existe un estimateur δ' tel que $R(\theta, \delta') \leq R(\theta, \delta)$ pour tout $\theta \in \Theta$ et $R(\theta_0, \delta') < R(\theta_0, \delta)$ pour au moins un $\theta_0 \in \Theta$. Dans le cas contraire, δ est admissible[7].*

Propriétés [7]

- ◇ L'estimateur de Bayes est admissible;
- ◇ L'estimateur de Bayes est biaisé;
- ◇ L'estimateur de Bayes converge en probabilité pour la taille de l'échantillon $n \rightarrow \infty$;

Exemple 1.3.1 *L'estimateur de Bayes δ de $\theta \in \Theta$, associé à la loi a posteriori $\pi(\theta|\underline{x})$ et au coût quadratique $l(\theta, d) = (\theta - d)^2$, est la moyenne a posteriori*

$$\theta_m = \mathbb{E}[\theta|\underline{x}] = \int_{\Theta} \theta \pi(\theta|\underline{x}) d\theta. \quad (1.3.2)$$

En effet, on a :

$$\mathbb{E}[l(\theta, d)] = d^2 - 2d\mathbb{E}[\theta] + \mathbb{E}[\theta^2],$$

qui est minimisé lorsque $d = \mathbb{E}[\theta]$.

Un tel estimateur prend aussi la forme de l'extremum de la distribution a posteriori :

$$\hat{\theta}_{MAX} = \arg \max_{\theta \in \Theta} \{\pi(\theta|\underline{x})\} \quad (1.3.3)$$

1.4 Conclusion

L'approche bayésienne nécessite le calcul d'estimateurs suivant une distribution a posteriori d'intérêt. Les estimateurs de type maximum a posteriori et moyen a posteriori sont donc calculables par des techniques d'intégration et d'optimisation (voir les expressions (1.3.2) et (1.3.3)). La mise en pratique de l'approche bayésienne est toutefois problématique car il est souvent impossible de trouver une expression analytique directement exploitable de ces estimateurs. De plus, les dimensions des variables considérées

rendent souvent très difficiles les calculs d'intégration et d'optimisation, même dans le cas où une forme analytique des estimateurs est disponible explicitement. Des solutions classiques pour résoudre ces problèmes d'implantation sont basées soit sur des méthodes d'approximations stochastique, soit sur des méthode de calculs numériques. Dans ce manuscrit nous nous intrissons aux méthode de calcule stochastique à savoir les méthode de Monte-Carlo. Ces méthodes sont présentées dans le chapitre suivant.

Méthodes de Monte-Carlo

Introduction

Les méthodes de Monte-Carlo sont des techniques statistiques ayant pour but d'estimer la valeur des différents moments d'une distribution à l'aide d'un échantillon aléatoires $(X_1, \dots, X_i, \dots, X_n)$ *i.i.d* selon une fonction de répartition F définie sur un le domaine χ .

Les méthodes de Monte-Carlo sont des méthode d'aproximation basées sur la simulation d'une suite de variables aléatoires $(X_n)_{n \geq 1}$ indépendantes et identiquement distribuées (*i.i.d*) selon une loi donnée.

Ce chapitre se veut une introduction aux méthodes de Monte-Carlo. La première section présente quelques techniques usuelles de simulation comme les méthodes d'inversion et d'acceptation-rejet. Le reste du chapitre est consacrée à l'étude de la méthodes de Monte-Carlo et quelques méthodes de réduction de la variance.

2.1 Méthodes usuelles de simulation

La simulation est une expérience faite dans un environnement de hasard. A l'origine, cette expérience repose sur le tirage des boules dans une urne, le tirage de carte ou le jet d'un dé pour obtenir des tables de nombres aléatoires. Avec l'évolution du temps et

l'apparition des machines différents logiciels mathématiques ont remplacé ces jeux-là pour obtenir des nombres aléatoires à simuler.

L'objectif de la simulation est de générer des échantillons *i.i.d* selon une loi de probabilités. En effet la simulation permet d'estimer les moments de tout ordre de distribution de variables aléatoires, trop complexes pour être étudiés de façon analytique en utilisant des algorithmes itératifs de génération.

2.1.1 Générateur de la loi uniforme $[0,1]$ ($\mathcal{U}_{[0,1]}$)

Définition 2.1.1 *Un générateur uniforme pseudo-aléatoire est un algorithme fondé sur une valeur initiale u_0 et une transformation D qui produit une suite $(u_i) = (D^i(u_0))$ à valeurs dans $[0, 1]$ telle que (u_1, \dots, u_n) reproduit, pour tout n , le comportement d'un échantillon uniforme *i.i.d* (v_1, \dots, v_n) , au sens où il fournit les mêmes réponses qu'un échantillon uniforme à une batterie de tests usuels.*

La validation des algorithmes se fait par une batterie de tests pour vérifier que la suite (u_1, \dots, u_n) est acceptable pour des tests statistiques de l'hypothèse u_1, \dots, u_n sont *i.i.d* selon $\mathcal{U}_{[0,1]}$. Ces tests sont généralement les tests de Khi-deux et de Kolmogorov Smirnov et l'utilisation des techniques de séries temporelles pour établir le manque de corrélation entre u_i et $(u_{i-1}, \dots, u_{i-k})$ [3].

Le générateur congruentiel

Le générateur congruentiel sur $\{0, 1, \dots, n\}$ est défini par la fonction :

$$u_i = D(u_{i-1}) = (au_{i-1} + b) \bmod(n + 1)$$

Les performances de ce générateur dépend du choix de (a, b) . Si on veut générer des nombres suivant la distribution $\mathcal{U}_{[0,1]}$ qui est la distribution de base dont sont déduites toutes les autres distributions, il suffit de mettre $u_i = (au_{i-1} + b)/(n + 1)$ [3].

2.1.2 Simulation de lois non uniformes (Méthodes générales)

On présente ici quelques techniques alternatives de simulation de lois non uniformes. Ces méthodes sont plus générales et elles n'utilisent que des propriétés analytiques des densités, comme la méthode d'inversion et celle d'acceptation-rejet.

Méthode d'inversion

Soit F une fonction de répartition inversible, alors si U suit une $\mathcal{U}_{[0,1]}$, la variable aléatoire $X = F^{-1}(U)$ a pour fonction de répartition F . En effet, pour tout réel x ,

$$\begin{aligned}\mathbb{P}(X \leq x) &= \mathbb{P}(F^{-1}(U) \leq x) \\ &= \mathbb{P}(U \leq F(x)) \\ &= F(x).\end{aligned}$$

Où on a appliqué respectivement l'aspect bijectif de F , sa croissance et la forme spécifique de la fonction de répartition de la loi uniforme.

Pour ce qui nous concerne, l'intérêt de ce résultat est manifeste : pour que F soit facilement inversible, alors pour générer une variable aléatoire de loi F , il suffit de générer une variable uniforme U et de lui appliquer F^{-1} . Cette méthode peut-être généralisée grâce à la définition suivant à des fonctions de répartition non bijectives[13].

Définition 2.1.2 (Inverse généralisée)

Soit X une variable aléatoire de fonction de répartition F . On appelle inverse généralisée de F ou fonction pseudo inverse de X , la fonction F^{-1} définie pour tout $u \in [0, 1]$ par

$$F^{-1}(u) = \inf\{x : F(x) \geq u\}.$$

Proposition 2.1.1 (Méthode d'inversion)

Soient U une v.a $\sim \mathcal{U}_{[0,1]}$, F une fonction de répartition et F^{-1} son inverse généralisée.

Alors :

1. La variable aléatoire $X = F^{-1}(U)$ a pour fonction de répartition F .

2. Si X a pour fonction de répartition F et si F est continue, alors la variable aléatoire $F(X)$ est de loi uniforme sur $[0, 1]$.

Algorithme d'inversion

- ◇ Entrées: Le nombre d'itération n .
- ◇ Poser $i = 1$.
- ◇ Pour $i \leq n$ faire
 - Générer un nombre pseudo-aléatoire u_i distribué selon $\mathcal{U}_{[0,1]}$.
 - Poser $x_i = F^{-1}(u_i)$.
 - $i = i + 1$.

Fin de l'algorithme

Exemple 2.1.1 Soit X une variable aléatoire qui suit la loi exponentielle de paramètre λ , sa fonction de répartition est $F(x) = 1 - e^{-\lambda x}$, $\forall x > 0$, et son inverse est

$$F^{-1}(x) = \frac{-\log(1-u)}{\lambda}$$

et pour $u \in [0, 1]$, on a $X = \frac{-\log(1-u)}{\lambda}$

Le programme sous le logiciel **R** qui permet de simuler un échantillon de taille $n = 100$ d'une variable aléatoire qui suit une loi exponentielle est donnée comme suit:

Programme simulation de la loi $\mathcal{E}(\lambda)$

$SimExp = fonction(lambda, n)$

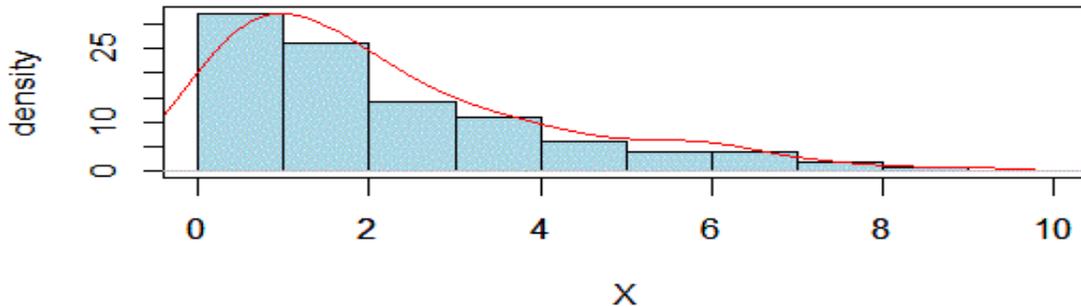
$\{u = runif(n, 0, 1)$

$x = -\log(1 - u)/lambda$

$x\}$

Fin du programme

La figure ci-dessous, représente l'histogramme de la distribution d'un échantillon de taille $n = 100$ généré à partir de la loi exponentielle de paramètre $\lambda = 0.5$.

Figure 2-1-1 : Distribution d'un échantillon gènère à partir de la loi $\mathcal{E}(0.5)$

Limites de la méthode d'inversion

Nous avons vu précédemment que la méthode d'inversion requiert la connaissance et le calcul rapide de $F^{-1}(x)$. Ceci n'est pas toujours envisageable : il suffit de penser à la loi normale centrée réduite, de fonction de densité :

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in \mathbb{R}$$

et dont l'inverse n'admet pas d'expression analytique simple. Dans ce cas la méthode d'acceptation-rejet, détaillée en section suivante, peut représenter une alternative judicieuse.

Méthode d'acceptation-rejet

Si l'on désire générer des valeurs d'une *v.a* avec une densité f , alors que ces propriétés ne sont pas connues, et sa fonction de répartition associée n'est pas inversible d'une manière explicite, la méthode d'acceptation-rejet est utilisée s'il existe une fonction g ayant un domaine identique à celui de f et une constante C positive, qui vérifie:

$$\frac{f(x)}{g(x)} \leq C, \quad \forall x \in \chi.$$

Alors les étapes de simulation d'une réalisation de f sont résumées dans l'algorithme suivant :

Algorithme Acceptation-Rejet

◇ Données : les densités $f(x)$ et $g(x)$.

La constante C .

Le nombre d'itérations T

◇ Poser $t = 1$.

◇ Pour $t \leq T$ faire

1. Générer y à partir de g .

2. Générer u à partir d'une distribution uniforme $\mathcal{U}_{[0,1]}$.

3. **Si** $u \leq \frac{f(y)}{Cg(y)}$

Alors on accepte la valeur générée.

Sinon on rejette u et y , et on reprend de l'étape 1.

Fin de l'algorithme

Détermination de C

Choisir C revient à vérifier l'équation : $\frac{f(x)}{g(x)} \leq C, \forall x \in \chi$.

Cependant, il est nécessaire que C soit la plus petite valeur possible pour avoir un taux d'acceptation moyen élevé. La valeur optimale de C est donc

$$C = \max_{\chi} \left(\frac{f(x)}{g(x)} \right)$$

Le taux d'acceptation de C sera donc défini :

$$\rho = \frac{f(x)}{Cg(x)}.$$

Remarquons que si $g(x)$ est proche de zéro, alors C devient très grande et le taux d'acceptation moyen diminue. L'efficacité de l'algorithme dépend donc de l'adéquation entre f et g . En particulier, pour que C reste borné, il faut que g ait des queues plus lourdes que celle de f , sans toutefois que leur rapport soit trop grand, au risque d'augmenter C [3].

Exemple 2.1.2 Simulation d'une loi normale à partir de la densité de Laplace de paramètre ($\lambda = 1$)

La densité de la loi de Laplace de paramètre ($\lambda = 1$) s'écrit :

$$g(x) = \frac{1}{2}e^{-|x|}, \quad x \in \mathbb{R}.$$

Si on note f la densité de la loi normale $\mathcal{N}(0, 1)$, on peut montrer que

$$f(x) \leq Cg(x) \quad \text{avec } C = \sqrt{\frac{2e}{\pi}}.$$

On peut donc mettre en place l'algorithme suivant :

Algorithme Acceptation-Rejet

1. Simuler une loi $\mathcal{E}(1)$
 - ◇ Simuler deux variables U et V i.i.d selon $\mathcal{U}_{[0,1]}$
 - ◇ Si $U \leq \frac{1}{2}$ alors $X \leftarrow -X$ (X est maintenant distribuée suivant la loi de Laplace)
2. Si $\frac{V}{\sqrt{2\pi}}e^{\frac{1}{2}-|X|} \leq \frac{1}{\sqrt{2\pi}}e^{-\frac{X^2}{2}}$ alors , accepter X , sinon retourner en 1.

Fin de l'algorithme

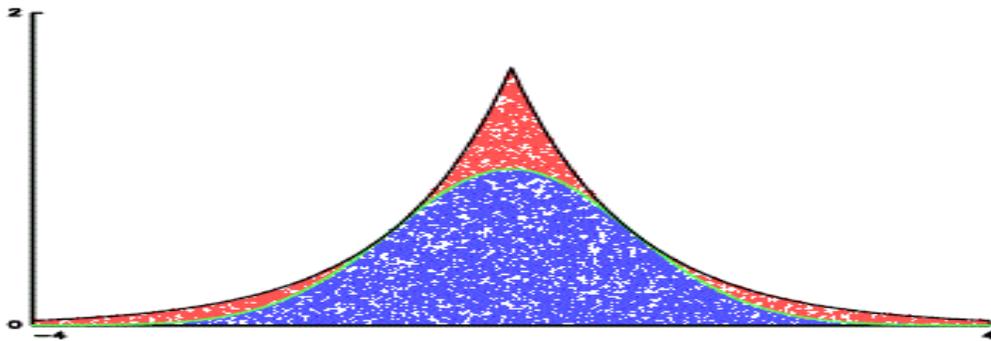


Figure 2-1-2 : Simulation d'une loi normale à partir de la loi de Laplace.

L'abscisse des points acceptés (en bleu) par l'algorithme est distribuée suivant la loi $\mathcal{N}(0, 1)$.

2.2 Méthodes de Monte-Carlo

La section précédente a fourni des outils généraux de production intensive de variables aléatoires distribuées suivant une loi donnée. Nous allons à présent voir comment l'inférence bayésienne peut tirer profit de ces techniques. Nous allons mettre en lumière deux classes principales de problèmes associés à l'inférence bayésienne : les problèmes d'optimisation et ceux d'intégration.

2.2.1 Intégration de Monte-Carlo

Les méthodes de Monte-Carlo sont des méthodes d'intégration numérique qui utilisent l'aléatoire. L'objectif est d'approcher une intégrale I incalculable explicitement, ces derniers sont basées sur la re-écriture de I en une quantité de la forme :

$$I = \mathbb{E}_f[h(X)] = \int_{\mathcal{X}} h(x) f(x) dx. \quad (2.2.1)$$

Le principe de la méthode de Monte-Carlo pour approcher I est de générer un échantillon (X_1, \dots, X_n) *i.i.d* suivant f et de proposer comme approximation la moyenne empirique:

$$\tilde{I} = \frac{1}{n} \sum_{i=1}^n h(X_i)$$

Alors, par la loi des grands nombres, $\mathbb{E}_f[h(X)]$ est approximée par \tilde{I} [20].

L'échantillon généré suivant f peut être calculé par exemple à partir de la méthode acceptation-rejet ou la méthode d'inversion, et la précision de cet estimateur est donnée par le théorème central limite, quand il existe qui vaut :

$$\frac{\sqrt{n}}{\sigma} \left(\frac{1}{n} \sum_{j=1}^n h(x_j) - \mathbb{E}(h(X_1)) \right) \overset{\mathcal{L}}{\rightsquigarrow} \mathcal{N}(0, 1). \quad (2.2.2)$$

Où n est le nombre d'échantillons à générer, σ^2 est la variance.

A partir de l'équation (2.2.2), on peut voir que l'erreur pour approcher une quantité par un estimateur Monte-Carlo est d'ordre $\frac{\sigma}{\sqrt{n}}$. Il est aussi facile de montrer que le nombre d'itérations nécessaires pour approcher une quantité I avec un certain niveau de précision est donné par:

$$n \simeq C\sigma^2. \text{ où } C \text{ est une Constant}$$

Donc afin de réduire le nombre d'itérations nécessaires pour approcher la quantité I , il est en notre intérêt de réduire la variance. Dans la section suivant nous présentons quelques méthodes de réduction de variance.

2.3 Réduction de variance

Nous avons vu de l'équation (2.2.2), que le calcul de la quantité $\mathbb{E}_f[h(X)]$ par la méthode de Monte-Carlo, nous amène à une erreur :

$$\left| \mathbb{E}_f[h(X)] - \frac{h(X_1) + h(X_2) + \dots + h(X_n)}{n} \right|$$

qui est d'ordre $\frac{\sigma}{\sqrt{n}}$, où $\sigma^2 = \text{Var}(h(X))$

Dans cette section, nous allons présenter des méthodes qui permettent d'améliorer la précision grâce à des techniques de réduction de variance. Rappelons ici qu'il existe plusieurs méthodes qui permettent de réduire la variance et donc d'améliorer l'erreur de précision. Mais nous nous contentons par présenter quelques méthodes.

2.3.1 L'échantillonnage préférentiel

Bien qu'il soit parfois difficile de produire un échantillon selon f , cette difficulté peut heureusement être contournée par la méthode dite d'échantillonnage préférentiel (importance sampling, en anglais) permettant de générer l'échantillon $(X_1, \dots, X_i, \dots, X_n)$ non pas de la distribution d'intérêt f , mais plutôt d'une autre densité g telle que:

$$\text{supp } f \subseteq \text{supp } g.$$

L'échantillonnage préférentiel consiste à générer un échantillon distribué selon la fonction g puis de le transformer pour qu'il possède les propriétés de la fonction f . Pour ce faire, une pondération des candidats à l'aide de poids notés ω utilisé dans le but de compenser la disparité entre les échantillons tirés des fonctions f et g . Précisément, pour estimer la fonction f , une fonction instrumentale quelconque g est d'abord définie, puis

un échantillon de référence $(X_1, \dots, X_i, \dots, X_n)$ obtenu par simulation de g , sans aucun rejet, est généré. Ensuite, I est estimé à l'aide de la moyenne pondérée

$$\tilde{I} = \frac{1}{n} \sum_{i=1}^n h(x_i) \omega(x_i) \quad \text{où } \omega(x_i) = \frac{f(x_i)}{g(x_i)} \quad (2.3.1)$$

On trouver une fonction h_1 proche de $|f.g|$ telle que l'on sache simuler suivant la densité suivante :

$$h(x) = \frac{h_1(x)}{\int_{\mathcal{X}} h_1(x)}$$

Cet estimateur, converge vers la véritable valeur de I pour les mêmes raisons que la méthode de Monte-Carlo classique quel que soit le choix de la loi de g parmi les lois ayant un support contenant celui de f [1].

L'effet de la pondération est évident. Supposons que la fonction instrumentale g sur-représente un certain point x , dans l'échantillon utilisé pour estimer I , par rapport à sa représentation dans un échantillon tiré de la loi f . Alors, le poids associé à ce point sera $\omega(x) = \frac{f(x)}{g(x)} < 1$ et sa représentation dans l'estimation sera importante pour compenser sa sur-représentation dans l'échantillon. Dans le cas contraire, $\omega(x) = \frac{f(x)}{g(x)} > 1$ et l'importance du point x augmente dans l'estimation du moment de f [23].

Choix de la distribution instrumentale

Le choix de g n'était pas restreint, cependant, il existe une densité g optimale qui minimise la variance théorique [1].

Certaines précautions sont cependant nécessaires:

- Si $\sup \frac{f}{g} = \infty$, g doit être rejetée.
- Si $\sup \frac{f}{g} = M$, l'échantillonnage pondéré peut être utilisé, mais quelques fois en concurrence avec la méthode d'acceptation-rejet.

Exemple 2.3.1 On cherche à calculer $I = \int_0^1 \cos\left(\frac{\pi x}{2}\right) dx$ par une méthode de Monte-Carlo. On peut approcher I par deux méthodes :

Méthode 1 :

$$I \approx \frac{\cos\left(\frac{\pi U_1}{2}\right) + \cos\left(\frac{\pi U_2}{2}\right) + \dots + \cos\left(\frac{\pi U_n}{2}\right)}{n}$$

avec U_1, U_2, \dots, U_n *i.i.d* de loi $\mathcal{U}_{[0,1]}$ (toujours pour n grand).

Méthode 2 :

$$I \approx \frac{1}{n} \left(\frac{g(Y_1)f(Y_1)}{h(Y_1)} + \dots + \frac{g(Y_n)f(Y_n)}{h(Y_n)} \right)$$

Nous avons $g(x) = \cos\left(\frac{\pi x}{2}\right)$ et $f(x) = \mathbb{I}_{[0,1]}(x)$, on définit une fonction h_1 (grossièrement) proche de $|f - g|$ par $h_1(x) = 1 - x$, donc :

$$h(x) = \frac{h_1(x)}{\int_0^1 h_1(x) dx} dx = 2(1 - x)$$

Si $Y \rightsquigarrow \mathcal{U}_{[0,1]}$ alors $1 - \frac{1}{2}Y$ est de loi de densité h .

Dans le programme suivant, nous estimons les variances des deux méthodes. Nous trouvons $Var_1 = 10^{-1}$ pour la première méthode et $Var_2 = 7.10^{-3}$ pour la méthode d'échantillonnage préférentiel par h . La variance est donc réduite.

2.3.2 Variable de contrôle

Si on sait calculer $\mathbb{E}[f(X)]$ pour une certaine fonction h , alors on peut aussi faire l'approximation :

$$\mathbb{E}[h(X)] \approx \frac{(h(X_1) - f(X_1)) + \dots + (h(X_n) - f(X_n))}{n} + \mathbb{E}[f(X)].$$

Afin de savoir quelle est la méthode la plus avantageuse, il faut comparer les variances des deux méthodes (la méthode de Monte-Carlo classique, variable de contrôle). Dans le cas de la deuxième méthode, l'erreur est :

$$\begin{aligned} & \mathbb{E}[h(X)] - \frac{(h(X_1) - f(X_1)) + \dots + (h(X_n) - f(X_n))}{n} - \mathbb{E}[f(X)] \\ &= (\mathbb{E}[h(X)] - \mathbb{E}[f(X)]) - \frac{(h(X_1) - f(X_1)) + \dots + (h(X_n) - f(X_n))}{n} \end{aligned}$$

La méthode est donc la suivante :

1. Trouver une fonction f proche de h et telle que l'on sache calculer $\mathbb{E}[f(X)]$ (le fait que f soit proche de h devrait faire en sorte que $Var(h(X) - f(X))$ est petite).
2. Estimer les variances $Var(h(X))$ et $Var(h(X) - f(X))$, et les comparer.

2.4 Conclusion

Malheureusement, les méthodes de Monte-Carlo d'estimation sont encore limitées par la modélisation car les lois de probabilité facilement simulables ne sont pas très nombreuses. Néanmoins, ces méthodes présentées jusqu'ici sont basées sur des hypothèses très restrictives et peuvent avoir des rendements plutôt médiocres selon le choix de la fonction instrumentale $g(x)$. D'autres méthodes de simulation comme les méthodes de Monte-Carlo par chaîne de Markov sont utilisées. Ces méthodes sont ainsi présentées dans le chapitre suivant.

Les méthodes de Monte-Carlo par Chaînes de Markov (MCMC)

Introduction

Les méthodes étudiées dans le présent chapitre utilisent les chaînes de Markov. Ce choix n'est pas innocent. En effet, l'utilisation des chaînes de Markov a comme avantage de faciliter l'échantillonnage de la distribution dont la forme explicite est difficile à obtenir ou bien de haute dimension.

On commence ce chapitre par un rappel sur les chaînes de Markov, puis on va présenter la démarche générale derrière les méthodes de Monte-Carlo par chaînes de Markov. Le reste du chapitre est dédié à l'étude des algorithmes de Metropolis-Hasting et de Gibbs, ainsi quelques méthodes de contrôle de convergence de ces algorithmes et on termine par citer quelques difficultés des méthodes MCMC.

3.1 Généralités sur les chaînes de Markov

Soit $(X_n)_{n \in \mathbb{N}}$ une suite de *v.a* à valeurs dans un espace S fini ou dénombrable, appelé espace des états. On dit que $(X_n)_{n \in \mathbb{N}}$ est une chaîne de Markov (CM) si :

$$\begin{aligned} p_{ij} &= \mathbb{P}(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \dots, X_1 = i_1, X_0 = i_0) \\ &= \mathbb{P}(X_{n+1} = j | X_n = i), \end{aligned}$$

pour tout entier $n \in \mathbb{N}$, pour tout état j et pour toute suite d'états i_0, i_1, \dots, i_{n-1} pour lesquels la probabilité conditionnelle a un sens, i.e

$$\mathbb{P}(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \dots, X_1 = i_1, X_0 = i_0) > 0.$$

Si le système transite de l'état i vers l'état j avec une probabilité p_{ij} qui ne dépend que des états i et j , et ne dépend pas de n .

Un tel système est dit sans mémoire ou encore markovien, ce qui signifie que la probabilité p_{ij} ne dépend pas des états antérieurs à i par lesquels le système est passé au cours de son histoire. Ainsi le futur état du système étudié dépend uniquement de son état présent et non de ses états passés.

De plus si $\mathbb{P}(X_{n+1} = j | X_n = i) = \mathbb{P}(X_1 = j | X_0 = i)$ alors la chaîne de Markov $(X_n)_{n \in \mathbb{N}}$ est dite homogène [5].

Dans tout ce qui suit, on ne considère que des chaînes de Markov homogènes dans le temps.

3.1.1 Matrice de transitions

Soit $(X_n)_{n \geq 0}$ une chaîne de Markov à espace d'états S .

On appelle matrice de transition associée à cette CM toute matrice carrée (lorsque S est finie i.e $\text{card}(S) = r$)

$$P = (p_{ij}) = \begin{pmatrix} p_{00} & p_{01} & \cdots & p_{0r} \\ p_{10} & p_{11} & \cdots & p_{1r} \\ \vdots & \vdots & \ddots & \vdots \\ p_{0r} & p_{r1} & \cdots & p_{rr} \end{pmatrix}.$$

Cette matrice est stochastique car le vecteur (stochastique) en ligne i contient les probabilités de toutes les transitions possibles en partant de l'état i sachant que les termes

sont positifs ou nuls et inférieur à un ($0 \leq p_{ij} \leq 1 \forall i, j \in N$) et leur somme est égale à 1 ($\sum_{j=0}^r p_{ij} = 1; \forall i, j = \overline{1, r}$) [22].

3.1.2 Noyau d'une chaîne de Markov

La première notion utile est celle de noyau car une chaîne de Markov homogène est définie par :

- Son noyau.
- La loi initiale de X_0 (ou sa valeur initiale).

Le noyau est une fonction à deux variables qui caractérise les probabilités de transition de X_n à X_{n+1} .

La forme dite intégrale du noyau est :

$$K_n(X, A) = \mathbb{P}(X_{n+1} \in A \mid X_n \in X)$$

A étant un sous-ensemble du domaine X de X_{n+1} . Pour une chaîne homogène, on notera simplement $K(X, A)$ au lieu de $K_n(X, A)$

Ainsi, pour tout couple (X, A) , $K(X, \cdot)$ est une mesure de probabilité sur l'espace X indexée par la valeur de X et $K(\cdot, A)$ est une fonction mesurable comprise entre 0 et 1, indexée par les sous ensemble mesurable de X . Dans la plupart des cas, la mesure de probabilité $K(X, \cdot)$ aura une densité $K(X, X')$ par rapport à une mesure support adéquate (indépendante de n pour une chaîne homogène) est telle que :

$$K(X, X') = \mathbb{P}[X_{n+1} = X' \mid X_n = X]$$

Dans le cas où X est discret, le noyau de transition est une matrice de transition P .

3.1.3 Propriétés fondamentales

I) Loi de probabilités d'états

Introduisons les probabilités d'états

$$\pi_k(n) = \mathbb{P}(X_n = k) \text{ avec } n = 0, 1, 2, \dots \text{ et } k = 1, 2, \dots$$

La distribution de X_n peut être décrite sous forme d'un vecteur ligne.

$$\pi(n) = (\pi_1(n), \pi_2(n), \dots) \text{ avec } \sum_{i \in S} \pi_i(n) = 1.$$

Pour calculer $\pi(n)$ il faut connaître l'état initial du processus, sa distribution initiale $\pi(0)$, et d'après la formule des probabilités totales on a :

$$\pi_k(n) = \sum_{i \in S} \pi_i(0) p_{ik}^n$$

En notation matricielle:

$$\pi(n) = \pi(0)P^n, \quad n = 1, 2, \dots$$

De façon plus analogue:

$$\pi(n+1) = \pi(n)P$$

Remarque 3.1.1 Ces résultats affirment qu'une C.M est complètement définie si l'on connaît sa matrice de probabilité de transition P en une étape, ainsi que sa distribution initiale $\pi(0)$.

II) Comportement asymptotique

a) Régime transitoire et régime permanent

Exprimer les probabilités d'états $\pi_k(n) = \mathbb{P}(X_n = k)$ d'une chaîne de Markov en fonction du nombre de transitions, en terme technique revient à étudier le régime transitoire du phénomène aléatoire en question. Il est évident que la distribution de X_n varie généralement en fonction du temps et qu'il dépend de la distribution initiale π_0 .

D'autre part, on constate souvent que la distribution $\pi(n)$ converge vers une distribution limite si $n \rightarrow \infty$. Si tel est le cas, on dit que cette dernière définit le régime permanent du processus stochastique. Contrairement au régime transitoire, le régime permanent n'est pas influencé par le choix de la distribution initiale. Notons que du point de vue pratique, on admet généralement que le régime permanent d'une chaîne de Markov est atteint à l'issue d'un nombre fini de transitions.

En termes formels, on dit qu'une chaîne de Markov converge vers π ou possède une distribution limite π si :

$$\lim_{n \rightarrow \infty} \pi(n) = \pi$$

indépendamment de la distribution initiale $\pi(0)$ et si π est une distribution de la probabilité. La convergence d'une chaîne de Markov est donc une propriété qui ne dépend que de la matrice de transition P [22].

b) Existence d'une distribution limite

Nous exposons ci-après sans démonstration deux théorèmes permettant d'assurer la convergence d'une chaîne de Markov dont le nombre d'états est fini.

Théorème 3.1.1 [22] *Si la matrice de transition P est telle qu'une au moins de ses puissances n'a que des termes strictement positifs, alors $\pi(n) \rightarrow \pi$ quelle que soit la distribution initiale $\pi(0)$, et $P^n \rightarrow P^*$.*

Lorsque $n \rightarrow \infty$, π est un vecteur de probabilité strictement positif, et P^ une matrice dont tous les linge sont identiques au vecteur limite π . En plus*

$$\pi P^* = \pi.$$

Un deuxième critère utilise les valeurs propres de la matrice P pour établir la convergence de la distribution $\pi(n)$ vers une loi limite.

Théorème 3.1.2 [22] *Si la valeur propre 1 de P est simple (c'est-à-dire de multiplicité 1) et si toute autres valeurs propres de P est de module strictement inférieur à 1, alors les conclusions du théorème 3.1.1 sont valables.*

c) Distribution stationnaire

Une distribution de probabilité π satisfaisant l'équation de balance globale

$$\pi = \pi P$$

est appelée distribution stationnaire de la matrice P .

3.1.4 Classification des chaînes de Markov

irréductibilité

Le noyau P est irréductible si pour tout couple $(i, j) \in S \times S$; il existe un entier k telle que:

$$\mathbb{P}(X_k = j \mid X_0 = i) = P_{ij}^k > 0$$

Récurrence

Soit $i \in S$, on appelle l'instant du premier retour en i , la variable aléatoire τ_i , définie par :

$$\tau_i = \inf \{n \geq 1, X_n = i\}.$$

L'état i est récurrent si $\mathbb{P}_i(\tau_i < +\infty \mid X_0 = i) = 1$.

L'état i est récurrent positif si $\mathbb{E}_i(\tau_i) < +\infty$. Il est récurrent nul si $\mathbb{P}_i(\tau_i < +\infty) = 1$ et $\mathbb{E}_i(\tau_i) = \infty$.

Un état qui n'est pas récurrent est dit transitoire.

Périodicité

La période de l'état i est le plus grand commun diviseur (*PGCD*) de toutes les longueurs de chemins reliant i à lui-même, et parcourus avec une probabilité strictement positive.

Lorsque le *PGCD* obtenu est égale à 1, on dit que l'état i est apériodique.

Ergodicité

Une chaîne de Markov homogène, irréductible, récurrente positive et apériodique est dite ergodique.

Absorption

Un état est absorbant si $p_{ii} = 1$ (on reste dans i une fois qu'on y est).

Réversibilité

On dit qu'une chaîne de Markov de matrice de transition P est réversible par rapport à la probabilité π si on a pour tous $i, j \in S$,

$$\pi(i)p_{ij} = \pi(j)p_{ji}. \quad (3.1.1)$$

En sommant (3.1.1) sur i , on en déduit le lemme suivant:

Lemme 3.1.1 *Si une chaîne de Markov est réversible par rapport à la probabilité π , alors π est une probabilité invariante.*

Proposition 3.1.1 [3] *Le fait qu'une chaîne de Markov homogène soit irréductible récurrente positive, c'est-à-dire : ergodique, assure l'existence d'une unique probabilité invariante (probabilité stationnaire).*

3.2 Méthodes de Monte-Carlo par Chaînes de Markov (MCMC)

Les méthodes de Monte-Carlo par Chaînes de Markov (Monte Carlo Markov Chain en anglais) ou MCMC sont utilisées lorsque la loi d'intérêt (la loi cible) n'est pas simulable directement par les méthodes usuelles (le calcul de la distribution sous forme analytique n'est pas envisageable), ou lorsque sa densité est connue à une constante de normalisation près.

Les méthodes MCMC permettent de simuler une chaîne de Markov $(\theta^t)_{t \geq 0}$ de matrice de transition P convergeant en un certain sens vers la loi d'intérêt π . Un tel algorithme est utilisé lorsque les intégrales sont de la forme :

$$I = \mathbb{E}_\pi[h(\theta)] = \int_{\Theta} h(\theta)\pi(\theta) d\theta \quad (3.2.1)$$

La première question qui se pose est : Pourquoi utiliser les méthodes MCMC et ne pas utiliser simplement les méthodes de Monte-Carlo classiques ?

Puisque la méthode d'échantillonnage pondéré permet de faire appel à une autre distribution que π .

Cette section étudie cette possibilité dans une autre perspective en montrant qu'il est possible d'obtenir un échantillon $(\theta^1, \dots, \theta^n)$ de loi π sans simuler directement suivant π . Le principe de base des méthodes décrites dans cette section est l'utilisation d'une chaîne de Markov ergodique de loi stationnaire π . Partant d'une valeur arbitraire $\theta^{(0)}$, on génère une chaîne $(\theta^{(t)})$ à partir d'un noyau de transition de loi stationnaire π , qui garantit de plus la convergence en lois vers π . Pour T "assez grand", on peut considérer $\theta^{(T)}$ comme distribué suivant π et obtenir ainsi un échantillon $\theta^{(T)}, \theta^{(T+1)}, \dots$, qui est effectivement distribué suivant π , même si les $\theta^{(T+t)}$ ne sont pas indépendants.

Les algorithmes MCMC font appel, comme les méthodes d'échantillonnage préférentiel et d'acceptation-rejet, à la simulation d'une loi instrumentale. Cette loi instrumentale peut être caractérisée par un noyau de transition $q(\cdot|\cdot)$ ou par une loi conditionnelle.

Parmi les algorithmes les plus connus on cite:

- Metropolis-Hastings(1953-1970).
- Échantillonnage de Gibbs (1990).

Définition: on appelle algorithme MCMC (Monte Carlo Markov Chain) toute méthode produisant une chaîne de Markov $(\theta^{(t)})$ ergodique de loi stationnaire la distribution d'intérêt [20].

3.2.1 Algorithme de Metropolis-Hastings

Les algorithmes de Hastings -Metropolis sont introduits par Metropolis, Rosenbluth, Rosenbluth, Teller et Teller (1953) [18] dans un contexte d'optimisation sur un espace d'états discret, ces méthodes ont ensuite été généralisées par Hastings (1970) [14] à la simulation statistique. Malgré, des tentatives ultérieures, ce n'est cependant qu'à partir de développement de l'échantillonnage de Gibbs par Gelfand et Smith (1999) qu'elles sont véritablement imposées à la communauté statistique. Ce décalage de plus de 30 ans est à déplorer seulement partiellement, car l'utilisation intensive des moyens de calcul imposée par ces méthodes n'était pas véritablement accessible avant 1990 [20].

Principe

Nous allons construire une chaîne de Markov $(\theta^t)_{t \geq 0}$ ergodique de loi stationnaire π , à partir d'une distribution conditionnelle $q(x|\theta)$. Pour la mise en œuvre de l'algorithme, on doit pouvoir simuler facilement à partir de la distribution $q(\cdot|\theta)$, et sa densité $q(x|\theta)$ doit être disponible analytiquement (au moins à une constante multiplicative près).

Ou à défaut , le rapport $\frac{f(x)}{q(x|\theta)}$ doit l'être [17].

Algorithme de Metropolis-Hastings[15]

1. Initialiser θ^0 tel que $\pi(\theta^0) \geq 0$
 2. A chaque étape $t \geq 0$: on dispose de θ^t
 - a) Générer une valeur x_t de $X_t : x_t \rightsquigarrow q(\cdot|\theta^t)$
 - b) Simuler une valeur u_t de $U_t \rightsquigarrow \mathcal{U}_{[0,1]}$
 - c) Calcul de $\rho(\theta^t | x_t) = \min\{1, \frac{\pi(x_t) q(\theta^t|x_t)}{\pi(\theta^t) q(x_t|\theta^t)}\}$
 - d) Poser : $\theta^{t+1} = \begin{cases} x_t & \text{si } u_t < \rho(\theta^t|x_t) \\ \theta^t & \text{sinon} \end{cases}$
 - 3) $t = t + 1$ et aller en 2
-

Propriétés

1. Cet algorithme accepte systématiquement les simulations x_t telle que le rapport de vraisemblance $(\frac{\pi(x_t)}{\pi(\theta^t)})$ soit supérieur à la valeur précédente.
2. L'algorithme de Metropolis-Hastings ne fait intervenir que les rapports $\frac{\pi(x_t)}{\pi(\theta^t)}$ et $\frac{q(\theta^t|x_t)}{q(x_t|\theta^t)}$, donc Il est semblable à la méthode d'acceptation-rejet (indépendant de la constante de normalisation).
3. Les derniers points de l'étape 2 de l'algorithme peuvent se traduire:

si $\pi(x_t) q(\theta^t | x_t) \geq \pi(\theta^t)q(x_t|\theta^t)$, on accepte x_t ,

sinon, on l'accepte quand même avec probabilité $\rho(\theta^t|x_t)$
4. La chaîne $(\theta^t)_{t \geq 0}$ peut prendre plusieurs fois la même valeur \implies échantillon non *i.i.d.*

5. La probabilité ρ est obtenue de la manière à ce que la condition de réversibilité soit satisfaite c'est-à-dire [4]

$$\pi(\theta)\mathbb{P}(\theta|x) = \pi(x)\mathbb{P}(x|\theta) \text{ avec } \mathbb{P}(\theta|x) = \rho(\theta, x)q(x|\theta)$$

Exemple 3.2.1 *Simulation de la loi de Poisson*

Nous utilisons l'algorithme *M.H* à marche aléatoire pour simuler la variable aléatoire X de loi de *Poisson* de paramètre λ donnée par:

$$f(x) = \mathbb{P}(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x \in \mathbb{N}.$$

En raisonnant proportionnellement on a:

$$f(x) \propto \frac{\lambda^x}{x!}, \quad x \in \mathbb{N}.$$

La loi de proposition q que nous utilisons est la suivante

$$\tilde{x} = \begin{cases} x^{t-1} + 1 & \text{si } u \geq \frac{1}{2} \\ x^{t-1} - 1 & \text{si } u \leq \frac{1}{2} \end{cases}$$

c'est-à-dire, on passe de $x^{(t-1)}$ à $x^{(t-1)} + 1$ avec la probabilité $\frac{1}{2}$, et de $x^{(t-1)}$ à $x^{(t-1)} - 1$ avec la probabilité $\frac{1}{2}$. L'algorithme *M.H* à marche aléatoire s'écrit alors sous la forme suivante:

1) Sachant que $X^{(t-1)} = x^{(t-1)}$, on passe à $\tilde{x} = x^{(t-1)} + 1$ avec la probabilité $\frac{1}{2}$, ou à $x^{(t-1)} - 1$ avec la probabilité $\frac{1}{2}$.

2) On calcule la probabilité d'acceptation

$$\rho = \min\left(1, \lambda^{\tilde{x}-x^{(t-1)}} \frac{(x^{(t-1)})!}{\tilde{x}!}\right).$$

3) On tire U uniformément sur $[0, 1]$

$$x^{(t+1)} = \begin{cases} \tilde{x} & \text{si } u \leq \rho \\ x^{(t)} & \text{si } u \geq \rho \end{cases}.$$

La figure 4-1-1 permet de visualiser l'évolution de la chaîne de Markov, aussi dans la figure 4-1-2 la vraie distribution de Poisson de paramètre $\lambda = 10$ obtenu à partir d'un

échantillon de taille $n = 1000$ en utilisant l'algorithme $M.H$ à marche aleatoire, pour $X_0 = 3$.

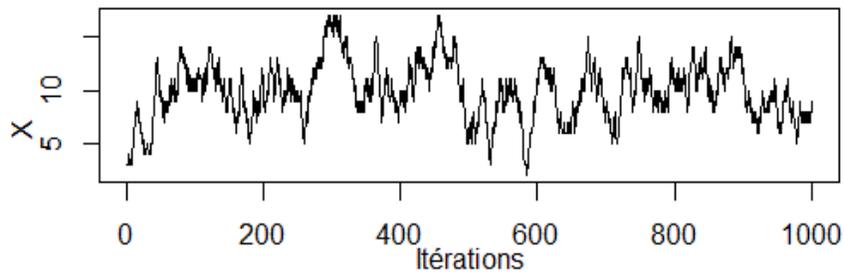


Figure 3-1-1 : L'évolution de la chaîne suivant la loi de $\mathcal{P}(10)$

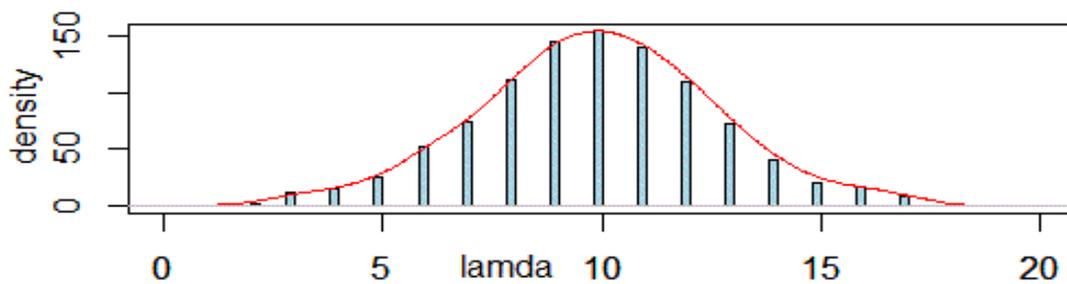


Figure 3-1-2 : Distribution et histogramme de la chaîne simulée suivant $\mathcal{P}(10)$

Théorème 3.2.1 [20] *Pour toute loi conditionnelle q , π est une loi stationnaire de la chaîne $\theta^{(t)}$ produite par l'algorithme de Metropolis-Hastings.*

Proposition 3.2.1 [13] *(Réversibilité de Metropolis-Hastings)*

Si la matrice de transition P est irréductible et apériodique, alors la chaîne (θ^t) produite par l'algorithme de Metropolis est irréductible, apériodique et π est réversible pour cette chaîne.

Metropolis-Hastings - Cas indépendant

L'algorithme de Metropolis-Hastings peut être présenté comme une généralisation de l'algorithme d'acceptation-rejet, tel que la loi instrumentale q est indépendante de θ^t .

Dans ce cas $q(x_t | \theta^t)$ sera noté $g(\theta^t)$, par analogie. Et l'algorithme s'écrit [16]:

—————**Metropolis-Hastings - Cas indépendant**—————

1. Initialiser θ^0 tel que $\pi(\theta^0) \geq 0$
 2. A chaque étape $t \geq 0$: on dispose de θ^t
 - a) Générer une valeur x_t de $X_t : x_t \rightsquigarrow g(\cdot)$
 - b) Simuler une valeur u_t de $U_t \rightsquigarrow \mathcal{U}_{[0,1]}$
 - c) Calcul de $\rho(\theta^t | x_t) = \min \left\{ 1, \frac{\pi(x_t)g(\theta^t)}{\pi(\theta^t)g(x_t)} \right\}$
 - d) Poser : $\theta^{t+1} = \begin{cases} x_t & \text{si } u_t < \rho(\theta^t | x_t) \\ \theta^t & \text{sinon} \end{cases}$
 3. $t = t + 1$ et aller en 2
-

Malgré que les variables x_t soient simulées indépendamment, l'échantillon résultant n'est pas indépendant et identiquement distribué, puisque la probabilité d'acceptation de x_t dépend de θ^t .

Metropolis-Hastings - Marche aléatoire

Une autre version de l'algorithme de Metropolis-Hastings utilise des propositions basées sur des pas de marche aléatoire. En prend en compte la valeur précédemment générée pour simuler la suivante, la loi instrumentale $q(x_t|\theta^t)$ sera ainsi écrite $g(x - \theta)$, c'est-à-dire que x_t peut s'écrire $\theta^t + \epsilon$, tel que ϵ étant une perturbation aléatoire de g , indépendante de θ^t , ainsi la chaîne de Markov générée associée à q est une marche aléatoire.

L'algorithme s'écrit [16]:

—————**Metropolis-Hastings - Marche Aléatoire**—————

1. **Initialiser** θ^0 tel que $\pi(\theta^0) \geq 0$
2. Á chaque étape $t \geq 0$: on dispose de θ^t
 - a) Simuler une valeur e_t de $\epsilon_t \rightsquigarrow g(\cdot)$
 - b) Simuler une valeur u_t de $U_t \rightsquigarrow \mathcal{U}_{[0,1]}$
 - c) Poser $x_t = \theta^t + e_t$
 - d) Calcul de $\rho(\theta^t | x_t) = \min \left\{ 1, \frac{\pi(x_t)g(\theta^t - x_t)}{\pi(\theta^t)g(x_t - \theta^t)} \right\}$

- e) Poser $\theta^{t+1} = \begin{cases} x_t & \text{si } u_t < \rho(\theta^t | x_t) \\ \theta^t & \text{sinon} \end{cases}$
3. $t = t + 1$ et aller en 2
-

Remarque 3.2.1 Si la loi de proposition q est symétrique alors la probabilité d'acceptation devient [20]:

$$\rho(\theta^t | x_t) = \min \left\{ 1, \frac{g(\theta^t)}{g(x_{t-1})} \right\}.$$

Metropolis-Hastings à sauts réversibles

Appelé aussi à une variable à la fois, quand le paramètre à simuler est de dimension grand, on est contraint de trouver une densité instrumentale multidimensionnelle engendrant une chaîne ayant le comportement d'une chaîne de Markov. Pour cela on peut utiliser un algorithme dit à une variable à la fois.

Le principe est de diviser le vecteur $\theta^{(t)}$ en plusieurs composantes $(\theta_1^{(t)}, \theta_2^{(t)}, \dots, \theta_d^{(t)})$ et de les simuler une par une. A chaque itération d'algorithme, on fait évoluer d composantes $\theta_i^{(t)}$ en utilisant d étapes de l'algorithme de Hastings-Metropolis, ce qui signifie que pour obtenir le nouveau vecteur $\theta^{(t+1)}$ il faudra utiliser d densités instrumentale $q_i(\cdot | \theta^{(t)})$, $i = 1, \dots, d$.

Pour simuler une composante $\theta_i^{(t)}$ il faut utiliser la loi instrumentale $q_i(x_i | \theta_{-i}^{(t+1)})$ et la loi cible $\pi_i(\theta_i | \theta_{-i}^{(t+1)})$ où $\theta_{-i}^{(t+1)} = (\theta_1^{(t+1)}, \dots, \theta_{i-1}^{(t+1)}, \theta_{i+1}^{(t+1)}, \dots, \theta_d^{(t+1)})$ (tel que $i = 1, \dots, d$).

On aura ainsi la probabilité d'acceptation de l'étape i [20]:

$$\rho(\theta_i^{(t)}, \theta_{-i}^{(t+1)}, x_i) = \min \left\{ 1, \frac{\pi_i(x_i | \theta_{-i}^{(t+1)}) q_i(\theta_i^{(t)} | x_i, \theta_{-i}^{(t+1)})}{\pi_i(\theta_i^{(t+1)} | \theta_{-i}^{(t+1)}) q_i(x_i | \theta_i^{(t)}, \theta_{-i}^{(t+1)})} \right\}$$

Metropolis-Hastings une variable à la fois

1. MH $(\pi_1(\theta_1 | \theta_{-1}^{(t+1)}), q_1)$
2. MH $(\pi_2(\theta_2 | \theta_{-2}^{(t+1)}), q_2)$
3.
- d. MH $(\pi_d(\theta_d | \theta_{-d}^{(t+1)}), q_d)$

L'algorithme de Metropolis-Hasting pourrait théoriquement être utilisé pour simuler un vecteur aléatoire de dimension p en utilisant une densité instrumentale multidimensionnelle. Mais quand p est grand, ce choix est rarement fait en pratique car la convergence d'un tel algorithme serait extrêmement lent. En effet, plus la dimension de l'espace des paramètres est grande plus la proportion candidate rejetée est importante, on préférera utiliser l'algorithme de Gibbs comme cadre général.

3.2.2 Echantillonnage de Gibbs

La technique de Metropolis-Hastings présentée dans la section précédente est attrayante par son universalité, mais, d'un autre côté, le manque de connexion entre le mécanisme de proposition q et la loi cible π peut être néfaste pour les propriétés de convergence de la méthode et, dans la pratique, peut facilement empêcher la convergence si la probabilité d'atteindre des parties éloignées du support de la loi π est trop petite. L'approche de l'échantillonnage de Gibbs, qui repose sur une perspective différente, est pour sa part fondée sur la loi π . Cette méthode tire son nom des champs aléatoires de Gibbs, où elle a été utilisée pour la première fois par Geman et Geman (1984).

D'un point de vue général, l'échantillonnage de Gibbs tire profit des structures hiérarchiques d'un modèle, par exemple lorsque celui-ci peut s'écrire sous la forme

$$\pi(\theta|x) = \int \pi_1(\theta|x, \lambda)\pi_2(\lambda|x)d\lambda, \quad (3.2.2)$$

l'idée est alors de simuler la loi jointe $\pi_1(\theta|x, \lambda)\pi_2(\lambda|x)$, afin d'obtenir $\pi(\theta|x)$ comme la loi marginale. Bien entendu, lorsque les deux lois $\pi_1(\theta|x, \lambda)$ et $\pi_2(\lambda|x)$ sont connues et peuvent être simulées, la génération de θ de $\pi(\theta|x)$ est équivalente à la génération de λ de $\pi_2(\lambda|x)$, puis de θ de $\pi_1(\theta|x, \lambda)$ [21].

—————**Algorithme d'échantillonnage de Gibbs bivarié**—————

Initialisation: Commencer par une valeur arbitraire $\lambda^{(0)}$.

Itération t: Pour $\lambda^{(t-1)}$ donné, générer

a) $\theta^{(t)}$ selon $\pi_1(\theta|x, \lambda^{(t-1)})$

b) $\lambda^{(t)}$ selon $\pi_2(\lambda|x, \theta^{(t)})$.

Lemme 3.2.1 [21] *Si $\pi_1(\theta|x, \lambda) > 0$ sur Θ ($\pi_2(\lambda|x, \theta) > 0$ sur Λ , respectivement), les deux suites $(\theta^{(m)})$ et $(\lambda^{(m)})$ sont des chaînes de Markov ergodiques de lois invariantes $\pi(\theta|x)$ et $\pi(\lambda|x)$, respectivement.*

Généralisation [21]

Une généralisation de l'Algorithme de Gibbs bivarié, consiste à considérer plusieurs groupes de paramètres, $\theta, \lambda_1, \dots, \lambda_p$, tels que

$$\pi(\theta|x) = \int \dots \int \pi(\theta, \lambda_1, \dots, \lambda_p|x) d\lambda_1 \dots d\lambda_p. \quad (3.2.3)$$

Cette généralisation correspond par exemple à l'introduction de niveaux additionnels dans le modèle hiérarchique (3.4.1), pour des raisons de modélisation ou de simulation, ou de décomposition de l'hyperparamètre λ ou du paramètre θ en des composantes de plus petites dimensions.

Comme expliqué avant, dans la Section 3.4 à propos du procédé de Gibbs bivarié, l'échantillonnage de Gibbs fournit des simulations de la loi jointe $\pi(\theta, \lambda_1, \dots, \lambda_p|x)$, lorsque certaines des lois conditionnelles associées à π sont disponibles. Bien entendu, lorsque $\pi(\theta|x)$ se décompose elle-même en lois conditionnelles, il n'y a pas besoin d'introduire des paramètres additionnels λ_i ($1 \leq i \leq p$).

La formulation générale de l'algorithme d'échantillonnage de Gibbs pour une loi jointe $\pi(\theta_1, \dots, \theta_p)$, de lois conditionnelles complètes π_1, \dots, π_p est exposée comme suit à l'algorithme d'échantillonnage de Gibbs [21].

————— **Algorithme d'échantillonnage de Gibbs** —————

1. **Initialisation** $(\theta_1^{(0)}, \dots, \theta_p^{(0)})$
2. Pour $(\theta_1^{(t)}, \dots, \theta_p^{(t)})$ donnés, simuler
 - a) $\theta_1^{(t+1)} \sim \pi_1(\theta_1|\theta_2^{(t)}, \dots, \theta_p^{(t)})$
 - b) $\theta_2^{(t+1)} \sim \pi_2(\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)}, \dots, \theta_p^{(t)})$
 -

$$p) \theta_p^{(t+1)} \sim \pi_p(\theta_p | \theta_1^{(t+1)}, \dots, \theta_{p-1}^{(t+1)}).$$

3. $t = t + 1$ et aller en 2
-

Les densités π_1, \dots, π_p sont les lois conditionnelles complètes, et un trait particulier de l'échantillonneur de Gibbs est que ce sont les seules densités qu'on utilise pour la simulation. Ainsi, même dans un problème en haute dimension, toutes les simulations peuvent être faites en dimension un, ce qui est généralement un avantage.

Propriétés

1. Le taux d'acceptation de l'échantillonnage de Gibbs est uniformément égale à 1. Toutes les valeurs simulées sont acceptées.
2. L'utilisation de l'échantillonnage de Gibbs entraîne des limitations fortes sur le choix des paramètres des lois instrumentales et suppose une connaissance préalable de certaines propriétés probabilistes ou analytiques de π .
3. L'échantillonnage de Gibbs est nécessairement multidimensionnel, même si certaines composantes du vecteur simulé sont artificielles du point de vue du problème inférentiel correspondant.
4. L'échantillonnage de Gibbs ne fonctionne pas lorsque le nombre de paramètre est une variable.

Remarques

1. Le schéma de simulation est d'autant plus efficace lorsque les distributions conditionnelles sont simulables rapidement $\pi_i(\theta_i | \theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_p)$.
2. Un des points remarquables de l'algorithme de Gibbs est qu'il rend possible la simulation de la distribution jointe $\pi(\theta) = \pi(\theta_1, \dots, \theta_p)$ à partir des distributions conditionnelles $\pi_i(\theta_i | \theta_{j \neq i})$.

3. Notons que l'échantillonnage de Gibbs s'applique particulièrement bien aux modèles hiérarchiques, c'est-à-dire à un modèle bayésien $(\pi(x|\theta), \pi(\theta))$, où la loi a priori $\pi(\theta)$ est décomposée en distributions conditionnelles.
4. Dans l'algorithme présenté, le balayage des éléments des vecteur est déterministe (ici, les simulations sont effectuées dans l'ordre, du premier au $P^{\text{ème}}$ élément de θ) : c'est un exemple de balayage systématique. Mais il est possible d'effectuer un balayage aléatoire où l'ordre de simulation est choisi aléatoirement à chaque itération.
5. L'intérêt du balayage aléatoire est que la chaîne de Markov produite est réversible et de distribution stationnaire la loi cible.

3.3 Contrôle de convergence des méthodes MCMC

Sous des conditions relativement générales, les chaînes produites par les algorithmes MCMC sont ergodiques, ces résultats sont cependant insuffisants lorsqu'on considère la mise en œuvre des méthodes MCMC, puisqu'ils ne fournissent pas directement des moyens de contrôle de la chaîne produite (dans le sens d'une règle d'arrêt permettant d'assurer que le nombre d'itérations est suffisant)

L'objectif de cette section est de proposer des outils qui permettent de répondre à :

- A partir de quel moment pouvons-nous affirmer que l'échantillon $(\theta^{(t)})$ sont distribuées selon la loi cible?

- Combien d'échantillons sont nécessaires pour obtenir une estimation précise de $(\theta^{(t)})$?

Pour cela plusieurs méthodes sont élaborées, et parmi ces méthodes on cite :

3.3.1 Méthodes de comparaison inter-intra chaînes

Cette méthode permet de contrôler la convergence des problèmes de calcul des intégrales sous cette forme :

$$I = \int_{\Theta} h(\theta) f(\theta) d\theta.$$

Notons par

NC : Longueur de la période de chauffage de la chaîne de Markov.

NI : Nombre d'itérations de la chaîne de Markov construite pour réaliser l'estimation.

Cette intégrale peut être estimée par la moyenne empirique :

$$\frac{1}{NI} \sum_{t=1}^{NI} h(\theta^{NC+t}). \quad (3.3.1)$$

Avant d'estimer l'intégrale, il faudra déterminer la période de chauffage et le nombre d'itérations d'intérêt [6].

Choix de période de chauffage

Le choix de la période de chauffage a été élaborée par Gelman et Rubin (1992) [7], il utilisent les critères de variances inter-intra chaîne pour assurer la convergence de la chaîne construite. Le principe est de construire plusieurs chaînes de Markov en parallèle de même période d'intérêt mais avec des valeurs initiales différentes.

Etant donnée M le nombre de chaînes de Markov construites.

Les variances inter-chaîne B et intra-chaîne W pour ces M chaînes de Markov sont définies respectivement par [19]:

$$B = \frac{NI}{M-1} \sum_{m=1}^M (\bar{\theta}_m - \bar{\theta})^2$$

$$W = \frac{1}{M} \sum_{m=1}^M S_m^2$$

avec

$$S_m^2 = \frac{1}{NI-1} \sum_{t=1}^{NI} (\theta_m^{(t)} - \bar{\theta}_m)^2$$

et

$$\begin{cases} \bar{\theta}_m = \frac{1}{NI} \sum_{t=1}^{NI} \theta_m^{(t)} \\ \bar{\theta} = \frac{1}{M} \sum_{m=1}^M \bar{\theta}_m \end{cases} .$$

Où $\bar{\theta}$ est le paramètre d'intérêt et $\theta_m^{(t)}$ est le $t^{\text{ème}}$ échantillon de la $m^{\text{ème}}$ chaîne.

La convergence des chaînes est alors mesurée par le critère $\hat{\rho}$, que nous appelons facteur d'échelle qui est défini par :

$$\hat{\rho} = \frac{1}{W} \left(\frac{NI-1}{NI} W + \frac{1}{NI} B \right) \quad (3.3.2)$$

Une valeur de $\sqrt{\hat{\rho}}$ proche de 1 indique une bonne convergence de l'échantillonneur, c'est-à-dire qu'un nombre d'itérations de chauffage égale à NC est suffisant pour obtenir des échantillons $(\theta^{NC+t}), t = 1, \dots, NI - NC$, distribués suivant la loi cible [7].

Choix du nombre d'itérations d'intérêt

Une fois que le nombre d'itérations de chauffage a été choisi, il est nécessaire de déterminer le nombre d'itérations NI requis pour obtenir un estimateur correct de I .

On va définir tout d'abord, un estimateur de référence noté \hat{E} est calculé pour un grand nombre d'itérations afin d'assurer à la fois la convergence de la chaîne et la précision de l'estimation. L'erreur quadratique moyenne (EQM) entre cet estimateur de référence \hat{E} et l'estimateur obtenu après $NI = p$ itérations est ensuite calculé grâce à :

$$e^2(p) = \left\| \hat{E} - \frac{1}{P} \sum_{t=1}^P h(x^{NC+t}) \right\| \quad (3.3.3)$$

3.3.2 Méthodes non paramétriques

Des tests non paramétriques standards peuvent être appliqués aux échantillons produits par les algorithmes MCMC à condition de prendre en compte la structure Markovienne. Robert, Rydén et Titterington [2000] proposent ainsi d'utiliser les tests de spearman et de Kolmogorov-Smirnov pour déduire les règles d'arrêt sur [7] :

- La dépendance à la condition initiale;
- La stationnarité de la chaîne.

Le test de kolmogorov-Smirnov est fondé sur une séparation de l'échantillon MCMC en deux parties et sur un sous-échantillonnage des deux moitiés, produisant deux échantillons quasi-indépendants et de même loi sous hypothèse de stationnarité, $(\theta_1^{(1)}, \theta_1^{(G+1)}, \dots)$ et $(\theta_2^{(1)}, \theta_2^{(G+1)}, \dots)$.

Philippe [1997], utilise une approche non paramétrique alternative en évaluant par Rao-Blackwellisation la densité marginale d'une composante de $\theta^{(t)}$, par la décomposition générale :

$$\begin{aligned} \int h(\theta)\pi(\theta|x)d\theta &= \int h(\theta)\pi(\theta_l|\theta_{-l},s)\pi(\theta_{-l}|x)d\theta_l d\theta_{-l} \\ &= \int \left(\int h(\theta)\pi(\theta_l|\theta_{-l},s)d\theta_l \right) \pi(\theta_{-l}|x)d\theta_{-l} \end{aligned}$$

Pour $l = 1, \dots, p$, où $\theta_{-l} = (\theta_r, r \neq l)$. Pour $h(\theta)$ elle en déduit une règle d'arrêt en suivant la convergence vers 1 des quantités suivantes ($l = 1, \dots, p$):

$$\delta_l^h = T^{-1} \sum_{t=1}^{T-1} (\theta_l^{[t+1]} - \theta_l^{[t]}) \sum_{b=1}^T \pi(\theta_l^{[t]}|\theta_{-l}^{(b)}, x)$$

où les $\theta_l^{[t]}$ sont les statistiques d'ordre associées avec la l -ième composante de l'échantillon MCMC $(\theta^{(1)}, \dots, \theta^{(T)})$.

Robert et al.[2000] développent une approche reposant sur le théorème central limite en montrant que, pour une chaîne $(\theta^{(t)})$ sous-échantionnée à des instants t_k générés par

$$t_{k+1} - t_k - 1 \rightsquigarrow \mathcal{P}(vk^d) \quad (v \geq 1, d > 0)$$

il est possible de recouvrir le théorème central limite usuel. En fait, si:

$$\hat{\mu}_T = T^{-1} \sum_{t=1}^T h(\theta^{(t)}), \quad \hat{V}_T = T^{-1} \sum_{t=1}^T h^2(\theta^{(t)}) - (\hat{\mu}_T)^2,$$

et si N_T est le nombre d'instant d'échantillonnage avant T , il vient que:

$$S_T = \frac{1}{\sqrt{N_T \hat{V}_T}} \sum_{k=1}^{N_T} (h(x^{(t_k)}) - \hat{\mu}_T)$$

est asymptotiquement $\mathcal{N}(0, 1)$.

3.3.3 Méthodes graphiques.

Une approche naturelle de contrôle de convergence est de suivre et d'examiner graphiquement le comportement de la chaîne simulée (**pour plus de détails voir [20] ou [7]**).

3.4 Difficultés des méthodes MCMC

Les méthodes MCMC présentent des difficultés importantes à savoir :

- Le choix de la loi instrumentale.
- La convergence : c'est-à-dire l'atteinte de l'équilibre de la chaîne de Markov peut être très lente, surtout lorsque $\pi(\theta|y)$ est difficile à approximer par une densité de transition q facilement simulable.
- Par ailleurs, même si il y a la convergence, détecter "le temps de chauffe" est loin d'être évident.
- MCMC est couteuse en temps de calcul.

3.5 Conclusion

Dans ce chapitre nous avons montré ce que les méthodes de Monte-Carlo par chaînes de Markov, nous a apporté de plus par rapport aux méthodes de Monte-Carlo classiques, en présentant les différents algorithmes MCMC, et la raison pour laquelle on les utilise avec la manière de leur déroulement en vérifiant les conditions de leurs utilisations. Ainsi quelques aspects qui permettent d'assurer le contrôle de leur convergence vers la distribution limite.

Application des méthodes MCMC à l'estimation bayésienne

Dans ce chapitre, nous présentons des exemples en appliquant les algorithmes de Metropolis-Hastings (*M.H*) et d'échantillonneur de Gibbs avec différents paramètres.

Dans le premier exemple nous utilisons l'algorithme *M.H* pour connaître la probabilité p de tomber sur *face* pour un lancer d'une pièce donnée, et nous utilisons l'échantillonneur de *Gibbs* pour une application sur des données réelles.

Pour ce faire, nous avons programmé les algorithmes à l'aide du logiciel statistique **R** (programmes en annexe).

4.1 Estimation de la probabilité de tomber sur *face*

On veut connaître la probabilité p de tomber sur *face* pour une pièce donnée. On procède à n lancers, on obtient l'échantillon (X_1, \dots, X_n) , soit x «*face*» pour $n - x$ «*pile*». La probabilité d'obtenir ce résultat s'exprime en fonction de p par :

$$f(x|p) = \mathbb{P}(X = x) = C_n^x p^x (1 - p)^{n-x}.$$

Il s'agit d'une loi binomiale $\mathcal{B}(n, p)$

Pour $n = 10$, on obtient l'échantillon $(1, 1, 1, 0, 1, 1, 0, 0, 1, 1)$ c'est-à-dire $x = 7$. Donc, on a :

$$f(x|p) = \mathbb{P}(X = 7) = C_{10}^7 p^7 (1-p)^3.$$

On vérifie que la vraisemblance est maximale pour $p = 0.7$.

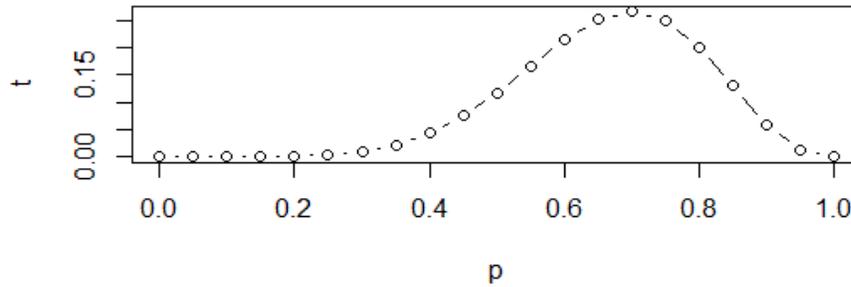


Figure 4-1-1 : Distribution de la loi de vraisemblance

$$\mathbb{P}(x = 7|p)$$

Puisque on ne dispose d'aucune connaissance de la loi a priori, on fait appelle à la loi de Jeffreys.

$$\begin{aligned} \frac{\partial^2 \ln f(x|p)}{\partial^2 p} &= -\frac{x}{p^2} - \frac{n-x}{(1-p)^2} \\ I(p) &= -\mathbb{E}\left(\frac{\partial^2 \ln f(x|p)}{\partial^2 p}\right) \\ &= n\left[\frac{1}{p} + \frac{1}{p-1}\right]. \end{aligned}$$

Donc la loi de Jeffreys pour ce modèle est

$$\pi(p) \propto [p(1-p)]^{-\frac{1}{2}}.$$

Il s'agit d'une $\mathcal{B}e\left(\frac{1}{2}, \frac{1}{2}\right)$.

D'après la formule de Bayes la loi a posteriori $\pi(p | x)$, est proportionnelle au produit de la loi a priori et la vraisemblance .

$$\pi(p | x) \propto p^{x+\frac{1}{2}-1} (1-p)^{n-x+\frac{1}{2}-1}.$$

D'où

$$p \mid x \rightsquigarrow \mathcal{Be}(x + \frac{1}{2}, n - x + \frac{1}{2}).$$

Nous allons utiliser la méthode MCMC pour estimer le paramètre p .

4.1.1 Algorithme

L'algorithme utilisé est celui de Metropolis-Hastings.

Pour $n = 10$, on obtient l'échantillon $(1, 1, 1, 0, 1, 1, 0, 0, 1, 1)$ c'est-à-dire $x = 7$

$$\pi(p|x) \propto p^{6.5}(1-p)^{2.5}.$$

-La loi de proposition est une loi *uniforme*

-À l'étape t généré $p' \rightsquigarrow \mathcal{U}_{[0,1]}$

-Calculer la probabilité d'acceptation $\rho(p^t|p') = \min\{1, \frac{\pi(p')U(p^t)}{\pi(p^t)U(p')}\}$

- généré $u \rightsquigarrow \mathcal{U}_{[0,1]}$ et prendre:

$$p^{t+1} = \begin{cases} p' & \text{si } u < \rho(p^t|p') \\ p^t & \text{si } u > \rho(p^t|p') \end{cases}$$

4.1.2 Résultats

L'estimateur du paramètre p obtenu pour 10000 itérations, partant de $p^0 = 1$ est:

$$p = 0.6893849.$$

La représentation graphique de l'évolution de p est donnée dans la figure 4-1-2

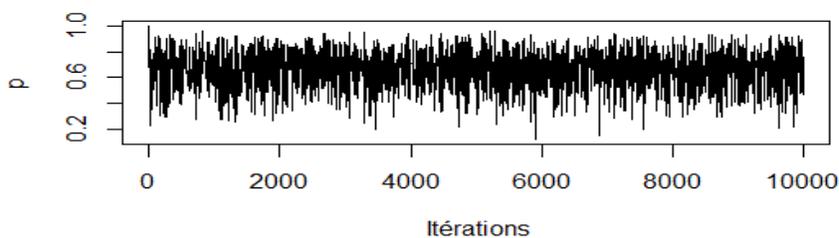


Figure 1-2-2 : L'évolution de la chaîne du paramètre p

La distribution et l'histogramme de la chaîne de p est donnée dans la figure 4-1-3:

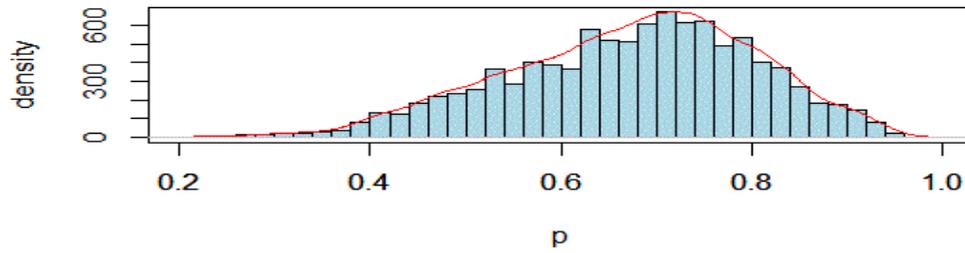


Figure 4-1-3: Distribution et histogramme de la chaîne du paramètre p

4.1.3 Etude de convergence

Pour vérifier la convergence, nous avons calculé le facteur d'échelle en utilisant $M = 3$ et pour le nombre d'itérations (1000, 5000, 10000). Les résultats sont donnés dans le tableau suivant:

Nombre de chaînes	Nombre d'itération	Facteur d'échelle
3	1000	1.0295238
3	5000	1.000067
3	10000	0.9999933

Tableau 4.1.1 : Facteur d'échelle pour la probabilité p

La représentation graphique correspondante à la convergence de l'estimateur de la probabilité p est dans la figure suivant:

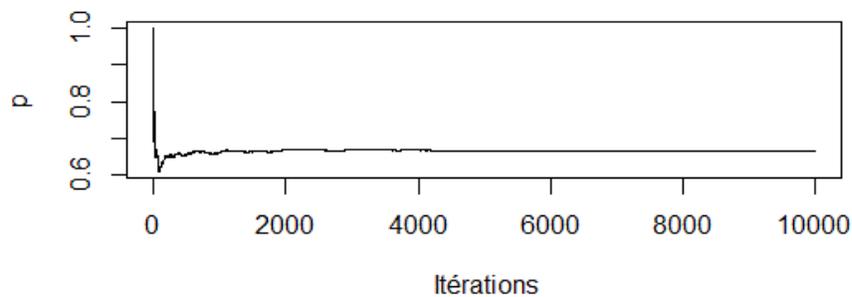


Figure 4-1-3: Convergence de l'estimateur de la probabilité p

4.1.4 Discussion des résultats

Les résultats montrent que pour la méthode MCMC, la chaîne de Markov obtenue converge vers la valeur du paramètre p à partir d'un nombre d'itérations supérieur à 1000.

4.2 Estimation de la taille d'un garçon âgé de 10 ans

Le tableau suivant présente la taille de 15 garçons âgés de 10 *ans* fréquentant les écoles primaires de la province de Luxembourg (source : Observatoire de la sante de la province du Luxembourg (Belgique)) [2].

133.2	137.1	138.1	131.1	152.6
135.6	145.7	146.5	131.8	131.2
141.7	154.1	140.4	146.1	124.4

Tableau 4-2-1 : Taille (cm) de 15 garçons âgés de 10 ans.

4.2.1 Test d'ajustement

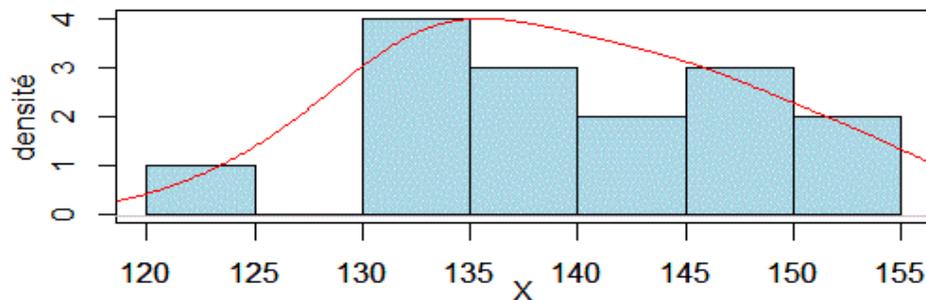


Figure 4-2-1 : Distribution de la taille d'un garçon âgés de 10 ans

La représentation graphique de l'histogramme et la distribution de l'échantillon (sous forme d'une cloche) nous induit à proposer la loi normale pour ajuster la distribution de cet échantillon. pour ce faire, on va utiliser le test de Kolmogorov-Smirnov.

La commande utilisée pour tester la normalité sous **R** est **:ks.test (voir plus en annexe)**

La statistique calculée de Kolmogorov et la p-value sont données par :

$$D = 0.1075,$$

$$p - value = 0.9874$$

Interprétation :

L'ajustement de la taille des garçons par la loi normale donne une p-valeur égale à 0.9874. Cette valeur est supérieure au seuil de signification $\alpha = 0.05$.

De plus, on trouve une statistique de Kolmogorov Smirnov égale à 0.1075. Cette valeur est inférieure à la valeur lue dans la table de Kolmogorov Smirnov à savoir 0.3376 au niveau de signification $\alpha = 0.05$.

Par conséquent on peut dire que les données suivent la loi normale.

$$y_i \rightsquigarrow \mathcal{N}(m, \sigma^2), \text{ avec } m = 139.3067, \text{ et } \sigma^2 = 72.15923.$$

4.2.2 Approche bayésienne

On suppose que la loi a priori de m est *gaussienne* de moyenne μ_0 et de variance σ_0^2 , et la loi a priori de σ^2 est une *inverse gamma* de paramètres α et β , c'est-à-dire $m \rightsquigarrow \mathcal{N}(\mu_0, \sigma_0^2)$ et $\sigma^2 \rightsquigarrow \mathcal{IG}(\alpha, \beta)$.

On peut montrer en utilisant la formule de Bayes que la loi a posteriori de m est une loi *normale* et celle de σ^2 est une *inverse gamma* et leurs paramètres sont donnés respectivement par :

La moyenne

$$m \mid \sigma^2, x \rightsquigarrow \mathcal{N}(M, sig^2)$$

avec $M = \frac{\sigma^2}{\sigma^2+n\sigma_0^2}\mu_0 + \frac{n\sigma_0^2}{\sigma^2+n\sigma_0^2}\bar{x}$

où:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

et

$$sig^2 = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}$$

La variance

$$\sigma^2 \mid m, x \rightsquigarrow \mathcal{IG}\left(\frac{n}{2} + \alpha, \frac{1}{2} \sum_{i=1}^n (x_i - m)^2 + \beta\right)$$

4.2.3 Estimation des paramètres

Pour notre application on suppose les valeurs des paramètres des lois a priori :

$$\mu_0 = 140, \sigma_0^2 = 0.25, \alpha = 0.002 \text{ et } \beta = 0.04.$$

Méthode de Monte-Carlo

- Pour estimer la moyenne m , on génère un échantillon i.i.d selon la loi *normale* $\mathcal{N}(M, sig^2)$, puis on calcule la moyenne empirique de l'échantillon généré.

On trouve

$$\hat{m}_{MC} = 139.9898$$

- Pour estimer la variance σ^2 , on génère aussi un échantillon i.i.d selon *inverse gamma* $\mathcal{IG}(\frac{n}{2} + \alpha, \frac{1}{2} \sum_{i=1}^n (x_i - m)^2 + \beta)$ et on calcule la moyenne empirique de l'échantillon généré.

On trouve

$$\hat{\sigma}_{MC}^2 = 17.10175$$

Méthode MCMC

- Pour estimer la moyenne m , on génère une chaîne de Markov de loi stationnaire la distribution *normale* $\mathcal{N}(M, sig^2)$, puis on calcule la moyenne empirique de l'échantillon généré.
- Pour estimer la variance σ^2 , on génère aussi une chaîne de Markov de loi stationnaire la distribution *inverse gamma* $\mathcal{IG}(\frac{n}{2} + \alpha, \frac{1}{2} \sum_{i=1}^n (x_i - m)^2 + \beta)$ et on calcule la moyenne empirique de l'échantillon généré.

L'algorithme utilisé est celui de l'échantillonneur de *Gibbs*, il est appliqué sur un échantillon X généré selon la loi *normale*.

Algorithme Générer $X \rightsquigarrow \mathcal{N}(m, \sigma^2)$

Initialiser les paramètres: $\mu_0, \sigma_0^2, \alpha, \beta$

À une iteration t :

Générer $m^t \rightsquigarrow \mathcal{N}(M, sig^2)$

Générer $\sigma^{2t} \rightsquigarrow \mathcal{IG}(\frac{n}{2} + \alpha, \frac{1}{2} \sum_{i=1}^n (x_i - m^t)^2 + \beta)$

(Voir l'algorithme sous **R** en annexe)

Résultat Les estimateurs de la moyenne et de la variance obtenus en générant 1000 variables aléatoires $X \rightsquigarrow \mathcal{N}(139.3067, 72.15923)$ sont:

m	σ^2
139.3136	70.26361

Les représentations graphiques de ces résultats sont donnés dans les figures (4-2-2), (4-2-3), (4-2-4) et (4-2-5)

La moyenne

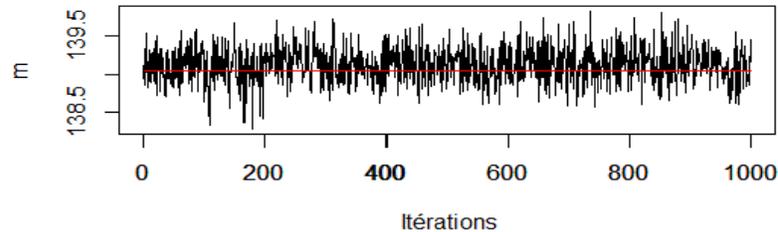


Figure 4-2-2 : Evolution de la moyenne m .

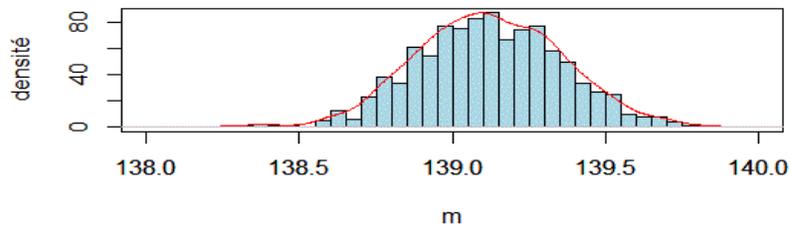


Figure4-2-3: Densité et histogramme de la moyenne m

La variance:

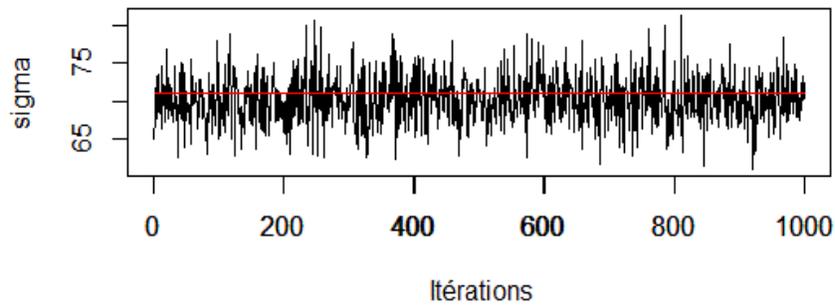


Figure4-2-4: Evolution de la chaîne de la variance σ^2

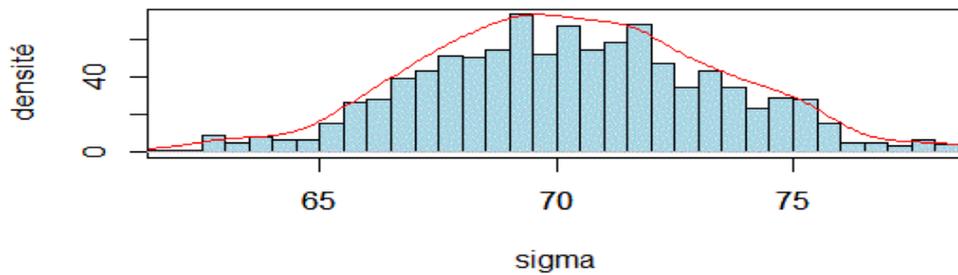


Figure4-2-5 : Densité et histogramme de la chaîne de la variance σ^2

Etude de convergence Pour vérifier la convergence de la méthode MCMC, nous avons utilisé le critère inter-intra chaîne, dont le principe est donnée dans le troisième chapitre , nous avons calculé le facteur d'échelle en utilisant $M = 5$ et pour le nombre d'itérations (1000, 5000, 10000). Les résultats obtenus sont donnés dans le tableau suivant:

la moyenne:

Nombre de chaînes	Nombre d'itération	Facteur d'échelle
5	1000	1.103379
5	5000	1.042843
5	10000	1.006545

Tableau 4-2-2 : Facteur d'échelle pour la moyenne m

La représentation graphique correspondante à la convergence de l'estimateur de la moyenne m est dans la figure suivante :

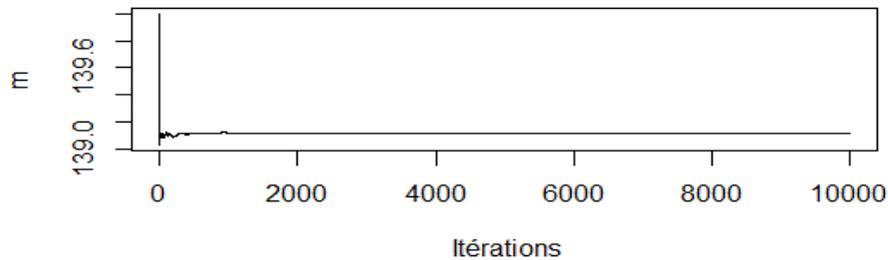


Figure 4-2-6 : Convergence de l'estimateur de la moyenne m

La variance

Nombre de chaînes	Nombre d'itérations	Facteur d'échelle
5	1000	1.028103
5	5000	1.017446
5	10000	1.008126

Tableau 4-2-3 : Facteur d'échelle pour la variance σ^2

La représentation graphique correspondante à la convergence de l'estimateur de la variance σ^2 est dans la figure suivante :

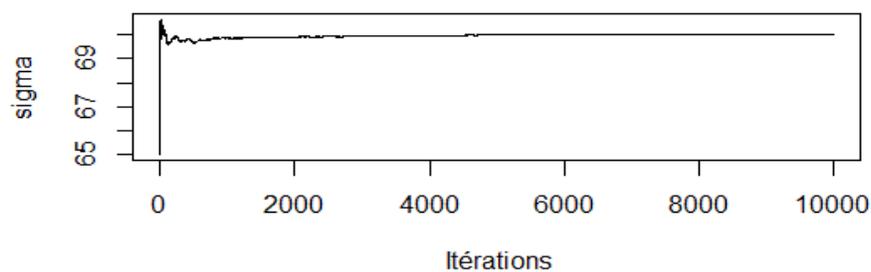


Figure 4-2-7 : Convergence de l'estimateur de la variance σ^2

4.2.4 Discussion des résultats

En comparant les résultats obtenus par la méthode de Monte-Carlo, avec ceux obtenus par la méthode MCMC, on constate que les estimateurs de m et σ^2 obtenus par la méthode MCMC sont plus proches des valeurs exactes que ceux obtenus en utilisant la méthode de Monte-Carlo.

l'erreur	\hat{m}	$\hat{\sigma}^2$
MC	0.6831	55.05748
MCMC	0.0069	1.895622

On peut ajouter que le nombre d'itérations nécessaire pour la convergence de la chaîne de Markov obtenue pour la moyenne est de 1000 itérations, et de 3000 itérations pour la variance.

Conclusion générale

L'objectif de ce travail est de présenter le principe et les algorithmes de base des méthodes de Monte-Carlo par chaînes de Markov (MCMC) qui sont des outils de simulation stochastique très puissants permettant de résoudre une grande partie des problèmes d'inférence statistique.

En dépit du fait que les méthodes MCMC sont basées sur des propriétés asymptotiques de convergence et donc qu'elles soient très lourdes en calcul, elles sont très intéressantes car elles proposent une structure universelle de simulation, et ce, quel que soit la complexité et la dimension de la densité à simuler.

On a appliqué les méthodes MCMC pour l'estimation bayésienne pour estimer le paramètre p d'un modèle *binomiale* qui représente la probabilité de tomber sur *face* pour une pièce donnée. Une application sur des données réelles représentant la taille de 15 garçons âgés 10 ans, la distribution des données est ajustée par loi *Normale*. L'ajustement est validé par le test de *Kolmogorov Smirnov*. On a utilisé l'échantillonneur de Gibbs pour estimer les paramètres de ce modèle *gaussien*, on a constaté que la convergence de cet algorithme est très rapide puisque les lois conditionnelles des paramètres permettent d'accepter toutes les valeurs générées des paramètres.

Les estimateurs obtenus sont meilleurs que ceux obtenus par Monte-Carlo classique.

Les résultats obtenus dans le cadre de ce mémoire ouvrent de nombreuses perspectives, telles que :

- Application de ces méthodes pour d'autres problèmes d'inférence statistique.
- Utiliser d'autre approche de contrôle de convergence comme les tests non paramétriques ou les méthodes graphiques.

- Estimer les paramètres avec d'autres méthodes de Monte-Carlo telle que la population de Monte-Carlo (PMC).

Annexe

```
*****Simulation de la loi de Poisson*****
fact = fonction(n){gamma(n+1)}

PoissonHastingMetropolis = fonction(n,lambda, x0)
{
X = x0
xn = x0
for (i in 1 :n)
{
u=runif(1)
xnCandidat = (xn+1)*(u<1/2)+(xn-1)*(u>=1/2)
critere = min(1,lambda^(xnCandidat-xn)* fact (xn) /fact (xnCandidat))
if (runif(1) < critere){ xn = xnCandidat }
else{ xn = xn}
X = c(X,xn)}
X
}
*****
.
```

```

3 4 3 4 5 6 7 8 7 8 9 8 8 7 7 6 6 5 6 5 4 4 4 4 5
5 4 4 4 4 5 5 6 7 8 9 10 11 12 13 12 11 11 10 9 10 9 8 7 8
9 8 9 8 9 8 9 10 11 10 9 10 10 11 10 9 10 9 10 11 12 13 14 14 13
14 13 13 12 12 13 12 11 10 11 12 11 10 11 10 11 11 10 11 10 11 10 11 12 11
12 11 11 11 10 9 10 11 12 11 10 11 12 13 14 14 14 14 13 12 12 12 11 12 11
10 11 12 13 12 11 10 11 10 9 10 11 10 9 8 9 8 9 10 11 10 11 10 9 8
9 8 7 7 6 7 8 7 8 9 10 11 12 11 10 9 8 9 8 7 8 7 6 5 5
6 6 7 8 9 10 9 8 9 8 7 8 9 8 9 10 11 12 11 10 9 8 9 9 10
9 10 11 12 13 12 12 12 11 10 9 10 11 12 12 13 12 11 10 11 10 9 8 9 10
9 10 9 10 11 12 11 10 11 10 9 10 11 10 9 10 9 10 9 9 10 9 9 10 9
8 7 6 5 6 6 7 8 7 8 9 10 9 10 11 10 11 12 11 12 11 12 12 13 12
12 13 12 13 12 13 14 15 15 16 15 15 16 17 16 16 16 17 16 16 15 16 17 17 16
17 17 16 15 16 17 16 15 14 14 13 13 14 15 14 15 14 13 12 13 13 13 13 12 11
11 10 11 10 9 8 9 8 9 8 9 8 9 8 9 8 9 10 11 10 11 11 10 9 10
11 10 9 10 11 12 13 14 15 15 14 13 12 11 10 9 8 7 8 9 10 11 10 11 12
11 10 10 9 9 10 11 10 9 10 9 8 9 8 9 8 7 8 9 10 9 10 9 10 9
8 9 10 9 10 11 12 13 12 11 12 11 10 9 10 11 12 13 14 13 12 13 14 14 13
12 12 13 12 13 12 11 12 11 12 11 10 11 12 11 12 13 14 13 14 15 14 15 16 17
16 16 17 16 15 15 15 14 13 14 13 12 12 13 12 13 13 13 13 12 13 14 15 14 15
14 13 13 12 11 10 9 10 9 10 11 10 9 8 7 6 5 6 5 6 7 6 5 6 7
8 7 6 5 6 7 6 7 7 8 9 9 10 11 10 9 10 9 8 7 7 6 5 5 4
3 4 5 6 6 7 6 7 8 9 8 9 9 10 11 10 11 10 9 8 9 8 9 10 11
10 9 10 11 10 11 10 9 10 9 10 11 10 9 10 9 8 9 8 7 8 7 6 5 4
3 3 3 2 2 3 4 5 5 6 6 6 7 8 9 8 9 10 11 10 11 12 12 13 12
13 13 12 12 13 12 11 12 12 11 12 11 10 11 12 11 10 11 12 11 10 9 8 7 8
7 6 7 8 7 6 6 6 7 6 6 6 6 7 6 6 7 7 6 7 8 9 8 7 6
7 8 9 8 8 9 10 9 10 11 11 10 11 10 11 12 13 14 15 14 13 12 11 10 11
12 12 13 12 11 10 9 10 10 11 12 11 10 11 10 9 8 8 7 8 9 8 8 7 8
7 6 6 5 5 6 7 8 7 6 5 5 6 6 7 8 9 10 11 12 11 12 12 13 13
12 12 11 11 12 11 10 9 8 9 10 9 10 11 12 13 14 15 14 13 12 11 11 10 11
10 9 10 11 10 11 10 11 12 11 11 10 9 8 9 10 9 10 9 8 9 10 9 10 9
8 9 8 9 8 9 10 9 9 8 9 10 11 12 13 12 11 12 11 10 11 11 10 11 12
11 10 11 10 10 11 10 11 12 11 11 10 9 8 9 10 11 12 12 13 14 13 14 13 12
11 10 10 11 12 11 12 13 12 11 12 11 12 13 14 13 12 11 10 9 10 11 10 9 10
9 8 9 8 9 10 9 10 11 12 13 12 13 12 11 12 11 10 11 12 12 12 11 12 12
13 14 15 14 13 13 12 13 14 13 14 13 14 13 13 14 14 13 12 11 12 11 10 9 8
9 8 9 10 9 8 9 8 7 8 7 8 7 8 7 6 7 6 7 8 7 8 9 10 9
8 9 8 8 8 9 8 8 9 10 9 10 9 10 9 10 11 10 9 8 7 6 7 7 6
7 7 8 9 9 10 10 9 10 11 10 9 8 9 8 7 8 9 8 7 7 6 5 5 6
7 8 9 8 8 7 8 7 8 8 7 7 8 7 8 7 7 7 8 8 9 8 7 7 8

```

L'évolution de la chaîne de Markov produit par une loi $\mathcal{P}(10)$ pendant 1000 étapes.

```
*****Probabilité de tomber sur face*****  
f=function(t){ dbeta(t,7.5,3.5)}  
beta=function(T){p=array(dim=c(T))  
p[1]=0.5  
for(i in 1:T)  
{  
yi=runif(1,0,(p[i]+1))  
d=f(yi)*dunif(p[i],0,(yi+1))  
e=f(p[i])*dunif(yi,0,(p[i]+1))  
k=d/e  
prob=min(1,k)  
u=runif(1)  
if(u<prob){p[i+1]=yi}  
else{p[i+1]=p[i]}  
}  
return(mean(p))  
}  
*****
```

```

***** Teste de la loi normal *****
y=c(133.2,137.1,138.1,131.1,152.6,135.6,145.7,
146.5,131.8,131.2,141.7,154.1,140.4,146.1,124.4)
ks.test(y,pnorm,mean(y),sd(y))
*****

```

Tests de Kolmogorov–Smirnov

La statistique de Kolmogorov–Smirnov apparaît lors d’un test d’adéquation d’une loi observée avec une loi de probabilité sur \mathbb{R} sans partie discrète, c’est-à-dire de fonction de répartition $F : \mathbb{R} \rightarrow [0, 1]$ continue. Elle est égale à :

$$k = \sup_{x \in \mathbb{R}} |F(x) - F_n(x)| = \max_{i=1}^n (x_{(i)}(F(x_{(i)}) - (i-1)/n) \vee (i/n - F(x_{(i)}))$$

Où $(x_{(i)})_{i=1}^n$ est l’échantillon ordonné, et $a \vee b = \max(a, b)$. Au seuil α donné, on accepte l’hypothèse d’égalité des lois si $k \leq k_{n,1-\alpha}$, cette dernière valeur étant donnée par la table qui suit :

$\alpha \backslash n$	0	1	2	3	4	5	6	7	8	9
0,01	1,0000	0,9950	0,9293	0,8290	0,7342	0,6685	0,6166	0,5758	0,5418	0,5133
0,05	1,0000	0,9750	0,8419	0,7076	0,6239	0,5633	0,5193	0,4834	0,4543	0,4300
0,10	1,0000	0,9500	0,7764	0,6360	0,5652	0,5094	0,4680	0,4361	0,4096	0,3875
0,15	1,0000	0,9250	0,7261	0,5958	0,5248	0,4744	0,4353	0,4050	0,3806	0,3601
0,20	1,0000	0,9000	0,6838	0,5648	0,4927	0,4470	0,4104	0,3815	0,3583	0,3391

$\alpha \backslash n$	10	11	12	13	14	15	16	17	18	19
0,01	0,4889	0,4677	0,4490	0,4325	0,4176	0,4042	0,3920	0,3809	0,3706	0,3612
0,05	0,4092	0,3912	0,3754	0,3614	0,3489	0,3376	0,3273	0,3180	0,3094	0,3014
0,10	0,3687	0,3524	0,3381	0,3255	0,3142	0,3040	0,2947	0,2863	0,2785	0,2714
0,15	0,3425	0,3273	0,3141	0,3023	0,2918	0,2823	0,2737	0,2659	0,2587	0,2520
0,20	0,3226	0,3083	0,2957	0,2847	0,2748	0,2658	0,2577	0,2503	0,2436	0,2373

$\alpha \backslash n$	20	21	22	23	24	25	26	27	28	29
0,01	0,3524	0,3443	0,3367	0,3295	0,3229	0,3166	0,3106	0,3050	0,2997	0,2947
0,05	0,2941	0,2872	0,2809	0,2749	0,2693	0,2640	0,2591	0,2544	0,2499	0,2457
0,10	0,2647	0,2586	0,2528	0,2475	0,2424	0,2377	0,2332	0,2290	0,2250	0,2212
0,15	0,2459	0,2402	0,2348	0,2298	0,2251	0,2207	0,2166	0,2127	0,2089	0,2054
0,20	0,2315	0,2261	0,2211	0,2164	0,2120	0,2079	0,2040	0,2003	0,1968	0,1934

$\alpha \backslash n$	30	31	32	33	34	35	36	37	38	39
0,01	0,2899	0,2853	0,2809	0,2768	0,2728	0,2690	0,2653	0,2618	0,2584	0,2552
0,05	0,2417	0,2379	0,2342	0,2308	0,2274	0,2242	0,2212	0,2183	0,2154	0,2127
0,10	0,2176	0,2141	0,2108	0,2077	0,2047	0,2018	0,1991	0,1965	0,1939	0,1915
0,15	0,2021	0,1989	0,1958	0,1929	0,1901	0,1875	0,1849	0,1825	0,1801	0,1779
0,20	0,1903	0,1873	0,1844	0,1817	0,1791	0,1766	0,1742	0,1718	0,1696	0,1675

$\alpha \backslash n$	40	42	44	46	48	50	52	54	56	58
0,01	0,2521	0,2461	0,2406	0,2354	0,2306	0,2260	0,2217	0,2177	0,2138	0,2102
0,05	0,2101	0,2052	0,2006	0,1963	0,1922	0,1884	0,1848	0,1814	0,1782	0,1752
0,10	0,1891	0,1847	0,1805	0,1766	0,1730	0,1696	0,1664	0,1633	0,1604	0,1577
0,15	0,1757	0,1715	0,1677	0,1641	0,1607	0,1575	0,1545	0,1517	0,1490	0,1465
0,20	0,1654	0,1616	0,1579	0,1545	0,1514	0,1484	0,1456	0,1429	0,1404	0,1380

$\alpha \backslash n$	60	65	70	75	80	85	90	95	100	105
0,01	0,2067	0,1988	0,1917	0,1853	0,1795	0,1742	0,1694	0,1649	0,1608	0,1570
0,05	0,1723	0,1657	0,1597	0,1544	0,1496	0,1452	0,1412	0,1375	0,1340	0,1308
0,10	0,1551	0,1491	0,1438	0,1390	0,1347	0,1307	0,1271	0,1238	0,1207	0,1178
0,15	0,1441	0,1385	0,1336	0,1291	0,1251	0,1214	0,1181	0,1150	0,1121	0,1094
0,20	0,1357	0,1305	0,1258	0,1216	0,1178	0,1144	0,1112	0,1083	0,1056	0,1031

$\alpha \backslash n$	110	120	130	140	150	160	170	180	190	200
0,01	0,1534	0,1470	0,1413	0,1362	0,1316	0,1275	0,1237	0,1203	0,1171	0,1142
0,05	0,1279	0,1225	0,1178	0,1135	0,1097	0,1063	0,1031	0,1003	0,0976	0,0952
0,10	0,1151	0,1103	0,1060	0,1022	0,0988	0,0957	0,0929	0,0903	0,0879	0,0857
0,15	0,1070	0,1025	0,0985	0,0950	0,0918	0,0889	0,0863	0,0839	0,0817	0,0796
0,20	0,1008	0,0965	0,0928	0,0895	0,0865	0,0838	0,0813	0,0790	0,0769	0,0750

Dudley (1964) a montré que pour tout $\mu < 0$,

$$\lim_{n \rightarrow \infty} \mathbb{P} \{k_n \leq \mu/\sqrt{n}\} = 1 + 2 \sum_{k=1}^{\infty} (-1)^k e^{-2k^2\mu^3},$$

formule qui permet d'approcher les p -valeurs $1 - F_{k_n}(k) = 1 - P\{k_n \leq k\}$ du test de Kolmogorov–Smirnov pour n assez grand.

```
*****Modèle gaussien*****
y=c(133.2,137.1,138.1,131.1,152.6,135.6,145.7,
146.5,131.8,131.2,141.7,154.1,140.4,146.1,124.4)
m=mean(y); SD=sd(y)
library(MCMCpack)
gibbs=function(T)
{x=rnorm(1000,m,SD)
m0=140
sig0=0.25
a0= 0.002
b0=0.04
tau= sd(y)
m= mean(y)
tau[1]=sd(y)
n=length(x)
for(i in 1:T)
{
sig=(tau[i]*sig0)/(tau[i]+n*sig0)
mea=(tau[i]/(tau[i]+n*sig0))*m0+(n*sig0/(tau[i]+n*sig0))*mean(x)
m[i]=rnorm(1,mea,sig)
a=a0+(n/2)
b=b0+0.5*sum((x-m[i])^2)
tau[i+1]=rinvgamma(1,a,b)
}
print(mean(m))
print(mean(tau))
}
gibbs(1000)
*****
```

Bibliographie

- [1] BELAID. N et DJERROUD. L, "Les méthodes de Monte-Carlo: (MCMC et PMC). Applications", mémoire master Recherche Opérationnelle, Bejaia, Juin 2013.
- [2] BOREUX. J. J, PARENT. E et BERNIER. J, "Pratique du calcul bayésien", Edition Springer, Paris, 2010.
- [3] BOUNHAR. K et SMAIL. N, "Méthode de Monte-Carlo par chaînes de Markov et leur application", mémoire master Recherche Opérationnelle, Bejaia. 2010/2011.
- [4] CHIB. S and GREENBERG. E, "Understanding the Metropolis-Hastings algorithm", The American Statistician, 49(4): 327-335, Novembre1998.
- [5] COUPIER. D, "Processus stochastiques", Polytech Lille, 2002
- [6] DOBIGEON. N, "Modèle bayésien hiérarchique pour le traitement multi-capteur", Thèse de doctorat, Institut National Polytechnique de Toulouse, Octobre 2007.
- [7] DROESBEKE. J. J, FINE. J et SAPORTA. G, "Méthodes bayésiennes en statistique", Editions TECHNIP, France, 2002.
- [8] DUHEILLE. F, "Processus stochastiques", Cours master 2, Université Claude Bernard Lyon 1, 2006/2007.
- [9] FORT. G, "Méthodes de Monte Carlo et Chaînes de Markov pour la simulation", Mémoire présenté pour l'obtention de l'Habilitation à Diriger les Recherches, université J. Fourier, France, 2009.

-
- [10] GELFAND. A. et SMITH. A, " Sampling based approaches to calculating marginal densities", J. American Statist. Assoc, 85, 398–409, 1990.
- [11] Genz, A, "An adaptive multidimensional quadrature procedure", Computer Physics Communications, 4, 11–15, 1972.
- [12] GUILLIN. A, MARIN. J. M, and ROBERT. C. P, "Estimation bayésienne approximative par échantillonnage préférentiel", Rapport de recherche, Université Paris IX Dauphine, 2005.
- [13] GUYADER. A, "Simulations Monte-Carlo", Université Pierre et Marie Curie, 2015/2016.
- [14] Hastings. W. K, "Monte Carlo sampling methods using Markov chains and their application", Biometrika, 57, 97–109, 1970.
- [15] KHRIBI. L, "L'échantillonnage de Gibbs pour l'estimation bayésienne dans l'analyse de survie", Université du Québec à Montréal, Février 2007.
- [16] FROMONT. M , " ISAO - Méthodes MCMC", ENSAI deuxième année, 2007/2008.
- [17] MAZET. V, "Introduction aux Méthodes de Monte Carlo par chaînes de Markov", Séminaire MTDE, Université Henri Poincaré, 2003.
- [18] METROPOLIS. N, ROSENBLUTH. A. W, ROSENBLUTH. M. N, TELLER. A. H. et TELLER. E, "Equations of state calculations by fast computing machines", Journal of Chemical Physics, 21, 1087–1092, 1953.
- [19] RIOU FRANÇA. L, "Statistique Bayésienne éléments de culture générale",INSERM U669, 2009.
- [20] ROBERT. C, "Méthodes de Monte Carlo par chaînes de Markov", Economia, 1996.
- [21] ROBERT. C. P, "Le choix bayésien. Principes et pratique", INSEE, Paris, 2006.
- [22] RUEGG. A, "Processus Stochastiques", Presse Polytechniques Romandes, 1989.

- [23] TOUZIN. G, "Etude des méthodes de Monte Carlo et de leurs efficacités relatives",
Université du Québec, 2013.
- [24] ZOUGAB. Nabil, "Cours Simulation de variables aleatoires", Université M. Mammeri
de Tizi-Ouzou, 2014-2015.