Université de Montréal

Étude de la performance d'un algorithme Metropolis-Hastings avec ajustement directionnel

par
Matei Mireuta

Département de mathématiques et de statistique
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de Maître ès sciences (M.Sc.)
en mathématiques

août, 2011

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé:

Étude de la performance d'un algorithme Metropolis-Hastings avec ajustement directionnel

présenté par:

Matei Mireuta

a été évalué par un jury composé des personnes suivantes:

Pierre Lafaye de Micheaux,	président-rapporteur
Mylène Bédard,	directrice de recherche
Jean-François Angers,	membre du jury

Mémoire accepté le: . . . . . .

RÉSUMÉ

Les méthodes de Monte Carlo par chaîne de Markov (MCMC) sont des outils très populaires pour l'échantillonnage de lois de probabilité complexes et/ou en grandes dimensions. Étant donné leur facilité d'application, ces méthodes sont largement répandues dans plusieurs communautés scientifiques et bien certainement en statistique, particulièrement en analyse bayésienne. Depuis l'apparition de la première méthode MCMC en 1953, le nombre de ces algorithmes a considérablement augmenté et ce sujet continue d'être une aire de recherche active.

Un nouvel algorithme MCMC avec ajustement directionnel a été récemment développé par Bédard *et al.* (IJSS, 9 :2008) et certaines de ses propriétés restent partiellement méconnues. L'objectif de ce mémoire est de tenter d'établir l'impact d'un paramètre clé de cette méthode sur la performance globale de l'approche. Un second objectif est de comparer cet algorithme à d'autres méthodes MCMC plus versatiles afin de juger de sa performance de façon relative.

Mots clés: échantillonneur indépendant, algorithme Metropolis-Hastings de type marche aléatoire, taux de convergence, algorithme Metropolis adaptatif, candidats multiples, diagnostics de convergence.

ABSTRACT

Markov Chain Monte Carlo algorithms (MCMC) have become popular tools for sampling from complex and/or high dimensional probability distributions. Given their relative ease of implementation, these methods are frequently used in various scientific areas, particularly in Statistics and Bayesian analysis. The volume of such methods has risen considerably since the first MCMC algorithm described in 1953 and this area of research remains extremely active.

A new MCMC algorithm using a directional adjustment has recently been described by Bédard *et al.* (IJSS, 9:2008) and some of its properties remain unknown. The objective of this thesis is to attempt determining the impact of a key parameter on the global performance of the algorithm. Moreover, another aim is to compare this new method to existing MCMC algorithms in order to evaluate its performance in a relative fashion.

Keywords: independent sampler, random walk Metropolis-Hastings algorithms, convergence rate, adaptive Metropolis algorithm, multiple proposals, convergence diagnostics.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ANNEXES

À Marina, Ioana, Neculai, Hunter et Askim

REMERCIEMENTS

Je tiens avant tout à remercier ma directrice de recherche, Mme Mylène Bédard, qui a été d'un immense soutien pendant toute la durée de ce projet. Grâce à sa perspicacité et à son expertise remarquables, elle a été une grande source d'inspiration et de conseils. Elle a toujours su se montrer disponible, à l'écoute et a su partager ses connaissances de manière exceptionnelle. J'aimerais aussi remercier les membres du jury pour la correction de ce mémoire et pour leurs commentaires qui sauront améliorer la qualité de ce travail.

INTRODUCTION

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. Parfois, il est nécessaire de calculer une valeur-p, une valeur-s ou tout simplement d'acquérir un échantillon d'une distribution donnée. 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). 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 appelées méthodes 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 capacité de facilement générer des valeurs de la distribution en question et la capacité de produire un gros 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.

Les méthodes de Monte Carlo par chaîne de Markov (MCMC) constituent une des approches les plus utilisées dans la communauté statistique. Ces méthodes 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) et généralisé par Hastings (1970), 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 obtenable pour une fonction de densité à postériori. 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.

En première partie, ce mémoire fera une description générale de l'approche MCMC. Il sera question de la construction de tels algorithmes, de leur convergence ainsi que du théorème central limite s'appliquant aux méthodes MCMC. Certains algorithmes de base seront aussi décrits en détails et finalement, plusieurs mesures empiriques de convergence seront présentées.

La deuxième partie traitera d'un algorithme Metropolis avec ajustement directionnel développé récemment par Bédard et Fraser (2008). Cette méthode sera d'abord décrite d'un point de vue théorique ainsi que d'un point de vue pratique. Étant donné la nouveauté de cette approche, plusieurs de ses propriétés restent partiellement mécon-

nues. Notre travail tentera donc d'élucider quelques aspects de la performance et de la convergence de cette méthode. Pour ce faire, nous utiliserons, en un premier lieu, deux exemples réalistes afin de démontrer l'importance du choix d'un certain paramètre λ intrinsèque à cette approche. Ensuite, nous construirons un exemple unidimensionnel afin de montrer que cette méthode peut parfois présenter des problèmes de convergence importants.

Enfin, le dernier chapitre portera sur une comparaison de la performance de l'algorithme Metropolis avec ajustement directionnel et d'autres approches MCMC. Notre comparaison sera faite surtout avec des méthodes locales, plus précisément les algorithmes Metropolis adaptatifs, avec essais multiples et avec rejet différé. En un même temps, nous complèterons également l'analyse d'un exemple décrit dans Bédard et Fraser (2008).

CHAPITRE 1

INTRODUCTION AUX ALGORITHMES MCMC

Les algorithmes MCMC représentent une méthode alternative pour l'échantillonnage de variables aléatoires ayant des densités complexes et/ou en grandes dimensions. Ce chapitre servira d'introduction aux approches MCMC et détaillera certaines des principales propriétés théoriques. En un premier lieu, il sera question de la construction de ces algorithmes et de quelques applications de base. Ensuite, les sections suivantes porteront sur leur convergence ainsi que sur certaines propriétés asymptotiques. Finalement, la dernière section sera consacrée à des mesures de convergence empiriques.

1.1 Construction des algorithmes MCMC

Un espace X et une fonction de densité $\pi_u(x)$, possiblement non normalisée, mais satisfaisant $0 < \int_{\mathsf{X}} \pi_u(x) < \infty$ sont donnés. Bien que d'autres situations soit possibles, le présent mémoire considérera seulement le cas le plus courant, c'est-à-dire où X est un sous-ensemble ouvert de $\mathbb{R}^n$ et la mesure sous-jacente est Lebesgue. Dans ce cas, pour tout $A \subseteq \mathsf{X}$ mesurable, la mesure de probabilité $\pi(\cdot)$ sur l'espace X est définie par

$$\pi(A) = \frac{\int_A \pi_u(x)dx}{\int_{\mathsf{X}} \pi_u(x)dx}.$$

Afin de produire un échantillon de la distribution $\pi(\cdot)$, l'approche MCMC consiste à générer une chaîne de Markov qui aura comme loi stationnaire $\pi(\cdot)$. En d'autres mots, il s'agit de trouver des probabilités de transition $P(x,dy) = P(X_{j+1} \in dy | X_j = x)$ (dy est un voisinage infinitésimal de y) pour la chaîne telles que

$$\int_{x \in \mathsf{X}} \pi(dx)P(x,dy) = \pi(dy) \qquad \forall x, y \in \mathsf{X}. \tag{1.1}$$

Conséquemment, si la chaîne est simulée assez longtemps, la distribution du j-ième état X_j est approximativement stationnaire $\mathscr{L}(X_j) \approx \pi(\cdot)$ (sujet à certaines conditions

mentionnées à la section 1.3) et une première observation aléatoire $Z_1 = X_j$ est alors obtenue. Il est ensuite possible d'utiliser une seconde valeur de départ (possiblement la même) et de simuler de nouveau la chaîne de manière identique afin d'obtenir Z_2. Le processus peut être répété en fonction de la taille échantillonnale souhaitée.

Bien que cette approche paraisse attrayante, construire une chaîne de Markov avec une telle propriété peut sembler une tâche ardue à priori. En fait, en utilisant certains concepts clé, il est étonnamment simple de bâtir de telles chaînes. Une idée importante est celle de la réversibilité.

Définition 1. *Une chaîne de Markov avec probabilités de transition $P(x,dy)$ sur un espace X est dite réversible par rapport à une distribution $\pi(\cdot)$ si*

$$\pi(dx)P(x,dy) = \pi(dy)P(y,dx) \qquad \forall x,y \in \mathsf{X}. \tag{1.2}$$

Ensuite, il est facile de voir que lorsqu'une chaîne de Markov est réversible par rapport à $\pi(\cdot)$, la distribution $\pi(\cdot)$ est stationnaire pour cette chaîne :

$$\begin{aligned} \int_{x\in\mathsf{X}} \pi(dx)P(x,dy) &= \int_{x\in\mathsf{X}} \pi(dy)P(y,dx) \\ &= \pi(dy)\int_{x\in\mathsf{X}} P(y,dx) = \pi(dy)\ . \end{aligned} \tag{1.3}$$

Donc, en théorie, il suffit de construire une chaîne de Markov réversible par rapport à $\pi(\cdot)$ et de simuler suffisamment d'états afin d'obtenir une observation aléatoire $Z_1 \sim \pi(\cdot)$. Une manière simple d'assurer la réversibilité de la chaîne est d'utiliser l'algorithme de Metropolis-Hastings (Metropolis *et al.* (1953); Hastings (1970)).

Cette méthode se sert d'une chaîne de Markov auxiliaire, simple à implémenter, disons $Q(x,\cdot)$, dont la densité de transition n'est pas nécessairement normalisée, $Q(x,dy) \propto q(x,y)dy$. En premier, $X_0 = x_0$ est choisi comme état initial. Ensuite, étant donné un état $X_j = x$, un nouvel état y est proposé selon la loi $Q(x,\cdot)$ et une variable indépendante B est générée, où $B \sim Bernoulli(\alpha)$ avec probabilité de succès définie par :

$$\alpha(x,y) = \min\left(1, \frac{\pi_u(y)q(y,x)}{\pi_u(x)q(x,y)}\right). \tag{1.4}$$

Si $B = 1$, l'état y est accepté et devient le nouvel état de la chaîne ($X_{j+1} = y$). Si $B = 0$, l'état y est rejeté et x est posé comme nouvel état ($X_{j+1} = x$).

De façon pratique, ce processus crée une nouvelle chaîne de Markov $\{X_j\}$ qui a la propriété additionelle d'être réversible par rapport à $\pi(\cdot)$. En fait, la réversibilité est assurée par la variable aléatoire B à travers la forme de la probabilité de succès (1.4).

Pour vérifier la propriété (1.2), supposons que $x \neq y$ (le cas $x = y$ est trivial) et dénotons $c = \int_{\mathsf{X}} \pi_u(x)dx$. Il s'en suit alors que :

$$\begin{aligned}
\pi(dx)P(x,dy) &= c^{-1}\pi_u(x)dx \;\; q(x,y)\alpha(x,y)dy \\
&= c^{-1}\pi_u(x)q(x,y)\min\left(1, \frac{\pi_u(y)q(y,x)}{\pi_u(x)q(x,y)}\right)dxdy \\
&= c^{-1}\min\left(\pi_u(x)q(x,y), \pi_u(y)q(y,x)\right)dxdy,
\end{aligned}$$

qui est symétrique en x et y (voir Metropolis *et al.* (1953)).

Donc, la construction d'une chaîne de Markov avec distribution stationnaire $\pi(\cdot)$ est bel et bien faisable. Par conséquent, il suffit de choisir un état initial et ensuite de simuler une chaîne de Markov selon l'algorithme Metropolis-Hastings plusieurs fois afin de générer un échantillon aléatoire provenant de $\pi(\cdot)$. En fait, bien qu'il soit possible de générer un échantillon aléatoire en réinitialisant la chaîne pour chaque observation souhaitée, il est souvent préférable de garder la queue d'une seule chaîne. En pratique, une première tranche d'observations, appelée période de chauffe (*Burn-in*), est laissée de côté parce que l'on considère que l'algorithme n'a pas encore convergé. Au-delà de la période de chauffe, les valeurs sont considérées comme un échantillon provenant de la distribution $\pi(\cdot)$. En vérité, les valeurs ainsi générées sont dépendantes, mais cela n'affecte généralement pas de façon considérable les quantités à calculer (souvent des espérances) et le processus de simulation est beaucoup plus efficace, exigeant moins de ressources. La distribution $\pi(\cdot)$ et la densité $c\pi_u(x)$ sont appelées distribution et densité cible tandis que $Q(x,\cdot)$ et $cq(x,\cdot)$ sont appelées distribution et densité instrumentales (c étant la constante de normalisation). Afin de ne pas introduire une notation superflue, $\pi(\cdot)$ se référera à la densité ainsi qu'à la distribution cible et $q(x,\cdot)$ se référera à la densité

ainsi qu'à la distribution instrumentale au-delà de ce chapitre introductoire.

Une autre caractéristique importante de l'approche Metropolis-Hastings est le fait que la constante de normalisation de la densité cible ne doit pas nécessairement être connue puisque cette densité n'apparaît que sous forme de ratio dans l'expression (1.4). Cela est très important dans le contexte bayésien, où cette constante est souvent incalculable. En effet, un des objectifs de l'approche bayésienne est d'inférer sur la distribution à postériori d'un paramètre θ. En utilisant une densité à priori $p(\theta)$, la vraisemblance des réalisations observées $L(\mathbf{x}|\theta)$ et en appliquant le théorème de Bayes, la forme de la densité à postériori de θ est donnée par :

$$\pi_u(\theta) \propto L(\mathbf{x}|\theta)p(\theta) \ .$$

La constante de normalisation est $\int_\theta L(\mathbf{x}|\theta)p(\theta)d\theta$. Dans la plupart des algorithmes de simulation Monte Carlo, le calcul de cette constante est requis pour générer des observations aléatoires de la distribution à postériori. Puisque cette étape n'est pas nécessaire dans l'approche MCMC, l'utilisation de ce type d'algorithme est extrêmement répandue dans la communauté bayésienne.

1.2 Algorithmes MCMC élémentaires

Il existe une multitude de façons de construire la chaîne de Markov instrumentale, mais cette section ne discutera que des plus usitées. La plupart des méthodes suivantes sont présentées pour le cas unidimensionnel, mais la généralisation en n dimensions est immédiate.

(a) Algorithme de type Metropolis-Hastings avec marche aléatoire (random walk Metropolis-Hastings (RWMH))

La méthode RWMH comporte une chaîne de Markov instrumentale dont la densité de transition dépend de l'état présent, plus précisément elle satisfait $q(x,y) = q(y-x)$. Par exemple, on peut choisir une densité instrumentale $N(x,\sigma^2)$. Donc, étant donné un

état présent x, l'algorithme génère un état potentiel y provenant d'une distribution normale centrée à x et avec variance σ^2. Si le nouvel état y est accepté, le prochain état potentiel sera généré selon une distribution $N(y,\sigma^2)$. Sinon, x sera posé comme état actuel et un autre état y éventuel sera proposé selon une loi $N(x,\sigma^2)$. Ce type d'algorithme est très versatile puisque son application nécessite peu d'information au sujet de la distribution cible. L'exemple ci-haut peut théoriquement s'appliquer à n'importe quelle distribution cible donnée, bien qu'il se comportera mieux dans certaines conditions que dans d'autres, et son efficacité peut généralement être améliorée en ajustant la variance σ^2. Une autre forme usitée pour la distribution instrumentale est la distribution uniforme, par exemple $Uniforme(x-1,x+1)$, mais d'autres choix sont aussi possibles.

(b) Algorithme symétrique de type Metropolis-Hastings

Certains algorithmes, comme la méthode RWMH avec loi instrumentale $N(x,\sigma^2)$ et bien d'autres, ont une propriété additionnelle conférée par la symétrie de la densité instrumentale, notamment le fait que $q(x,y) = q(y,x)$. Cette caractéristique permet de simplifier le calcul de α :

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

(c) Échantillonneur Metropolis-Hastings indépendant (Independent Sampler (IS))

La méthode IS est une approche où la distribution instrumentale est indépendante de l'état présent, c'est-à-dire que $q(x,y) = q(y)$. On peut choisir, par exemple, une distribution instrumentale $Student_f$, où f représente les degrés de liberté. Étant donné un état actuel x, un nouvel état y est proposé selon une loi $Student_f(y)$ indépendamment de x. Cette nouvelle valeur est ensuite acceptée avec probabilité (1.4). Cette méthode requiert en général une connaissance plus approfondie de la distribution cible pour une convergence optimale. Par exemple, il est préférable que les régions de forte densité de la distribution instrumentale coïncident avec les régions de forte densité de la distribution cible. D'un autre côté, il est souhaitable que l'épaisseur des queues des deux densités soit relativement similaire afin de ne pas sous-estimer certaines régions, particulièrement

dans le calcul de valeurs-p.

(d) Échantillonneur de Gibbs (Gibbs Sampler (GS))

Dans certaines situations, il peut être plus simple de générer des observations à partir de densités conditionnelles plutôt que de densités conjointes. L'approche GS se base sur cette idée et peut être un algorithme très puissant pourvu que l'on puisse aisément exprimer les densités conditionnelles.

On suppose que $\pi_u(\cdot)$ est une densité n-dimensionnelle sur un sous-espace X de $\mathbb{R}^n$ et on exprime les points de la façon $\mathbf{x} = (x_1, \ldots, x_n)$. Pour un état présent $\mathbf{x}^{(0)}$, la méthode GS génère une observation $x_1^{(1)}$ pour la première composante selon $\pi(x_1|x_2^{(0)}, x_3^{(0)}, x_4^{(0)}, \ldots, x_n^{(0)})$, ensuite une observation $x_2^{(1)}$ selon $\pi(x_2|x_1^{(1)}, x_3^{(0)}, x_4^{(0)}, \ldots, x_n^{(0)})$ et ainsi de suite jusqu'à l'observation $x_n^{(1)}$. Cette méthode génère une chaîne de Markov (voir par exemple Gelman *et al.* (1995)) qui aura comme loi stationnaire $\pi(\cdot)$. En effet, les probabilités de transition satisfont (1.1) :

$$
\begin{aligned}
&\int_{\mathbf{x}\in\mathbb{R}^n} \pi(d\mathbf{x}) P(\mathbf{x}, d\mathbf{y}) \\
&= \int_{x_n}\int_{x_{n-1}} \ldots \int_{x_1} \pi_u(x_1, x_2, x_3, \ldots, x_n) * \pi_u(y_1|x_2, x_3, \ldots, x_n) * \pi_u(y_2|y_1, x_3, \ldots, x_n) * \ldots * \\
&\quad \pi_u(y_i|y_1, y_2, \ldots, y_{i-1}, x_{i+1}, \ldots, x_n) * \ldots * \pi_u(y_n|y_1, y_2, \ldots, y_{n-1}) dx_1 dx_2 \ldots dx_n d\mathbf{y} \\
&= \int_{x_n}\int_{x_{n-1}} \ldots \int_{x_2} \pi_u(x_2, x_3, \ldots, x_n) * \pi_u(y_1|x_2, x_3, \ldots, x_n) * \pi_u(y_2|y_1, x_3, \ldots, x_n) * \ldots * \\
&\quad \pi_u(y_i|y_1, y_2, \ldots, y_{i-1}, x_{i+1}, \ldots, x_n) * \ldots * \pi_u(y_n|y_1, y_2, \ldots, y_{n-1}) dx_2 dx_3 \ldots dx_n d\mathbf{y} \\
&= \int_{x_n}\int_{x_{n-1}} \ldots \int_{x_2} \pi_u(y_1, x_2, x_3, \ldots, x_n) * \pi_u(y_2|y_1, x_3, \ldots, x_n) * \ldots * \\
&\quad \pi_u(y_i|y_1, y_2, \ldots, y_{i-1}, x_{i+1}, \ldots, x_n) * \ldots * \pi_u(y_n|y_1, y_2, \ldots, y_{n-1}) dx_2 dx_3 \ldots dx_n d\mathbf{y}
\end{aligned}
$$

$$
\begin{aligned}
&= \int_{x_n}\int_{x_{n-1}}\dots\int_{x_3} \pi_u(y_1,x_3,\dots,x_n) * \pi_u(y_2|y_1,x_3,\dots,x_n) * \dots * \\
&\quad \pi_u(y_i|y_1,y_2,\dots,y_{i-1},x_{i+1},\dots,x_n) * \dots * \pi_u(y_n|y_1,y_2,\dots,y_{n-1}) dx_3 dx_4 \dots dx_n d\mathbf{y} \\
&= \int_{x_n}\int_{x_{n-1}}\dots\int_{x_3} \pi_u(y_1,y_2,x_3,\dots,x_n) * \pi_u(y_3|y_1,y_2,x_4\dots,x_n) * \dots * \\
&\quad \pi_u(y_i|y_1,y_2,\dots,y_{i-1},x_{i+1},\dots,x_n) * \dots * \pi_u(y_n|y_1,y_2,\dots,y_{n-1}) dx_3 dx_4 \dots dx_n d\mathbf{y} \\
&= \dots \\
&= \pi_u(y_n|y_1,y_2,\dots,y_{n-1}) \int_{x_n} \pi_u(y_1,y_2,\dots,y_{n-1},x_n) dx_n d\mathbf{y} \\
&= \pi(y_1,\dots,y_n) d\mathbf{y} = \pi(\mathbf{y}) d\mathbf{y}.
\end{aligned}
$$

L'algorithme GS peut s'implémenter de façon déterministe comme l'exemple ci-haut ou bien de façon aléatoire, c'est-à-dire que la composante mise à jour à chaque itération est choisie au hasard parmi les n composantes. En général, l'échantillonneur de Gibbs aléatoire est réversible, mais celui déterministe ne l'est pas (Liu *et al.* (1995)).

1.3 Convergence des algorithmes MCMC

Il est maintenant clair qu'il existe plusieurs façons simples de construire des chaînes de Markov avec distribution stationnaire $\pi(\cdot)$. Toutefois, même si la distribution stationnaire est connue, elle n'est pas nécessairement unique et la chaîne n'y converge pas dans tous les cas. À titre illustratif, deux exemples tirés de Roberts et Rosenthal (2004) démontrent la nécessité de l'irréductibilité et de l'apériodicité de la chaîne.

Exemple : Supposons que $\mathsf{X} = \{1,2,3\}$ et que $\pi\{1\} = \pi\{2\} = \pi\{3\} = 1/3$. Aussi, $P(1,\{1\}) = P(1,\{2\}) = P(2,\{1\}) = P(2,\{2\}) = 1/2$ et $P(3,\{3\}) = 1$. Ici, $P(x,\{y\}) = P(X_{j+1} = y|X_j = x)$. La distribution stationnaire pour cette chaîne est $\pi(\cdot)$, mais si X_0 est 1, alors $X_j \in \{1,2\}$ pour tout j et donc $P(X_j = 3)$ ne converge pas vers $\pi(\{3\})$.

Cet exemple illustre une chaîne de Markov dite réductible, puisque l'état 3 ne peut jamais être atteint à partir de l'état 1 ou 2. Dans ce cas, la distribution stationnaire n'est pas unique et dépend du point de départ de la chaîne. Lorsque X est dénombrable, la

caractéristique d'irréductibilité signifie que chaque état a une probabilité non nulle d'être éventuellement atteint à partir de n'importe quel autre état. Dans un contexte où X est non dénombrable, une condition similaire, quoique moins robuste, peut être définie.

Définition 2. *Une chaîne de Markov est ϕ-irréductible s'il existe une mesure ϕ, non-nulle et σ-finie, sur l'espace X telle que pour tout $A \subseteq \mathsf{X}$ avec $\phi(A) > 0$ et pour tout $x \in \mathsf{X}$, il existe $j \in \mathbb{N}$ tel que $P^j(x,A) > 0$, où $P^j(x,A) = P(X_j \in A | X_0 = x)$.*

Exemple : Supposons encore que $\mathsf{X} = \{1,2,3\}$ et que $\pi\{1\} = \pi\{2\} = \pi\{3\} = 1/3$. Aussi, $P(1,\{2\}) = P(2,\{3\}) = P(3,\{1\}) = 1$. La distribution stationnaire est encore π et la chaîne est irréductible. Toutefois, si $X_0 = 1$ alors $P(X_j = 1)$ oscillera entre 0 et 1 à tous les multiples de 3, et donc ne convergera pas vers 1/3.

Ce deuxième exemple démontre le concept de périodicité, c'est-à-dire qu'un état n'est atteignable que dans un nombre précis de pas ou tout multiple de ce nombre. La condition d'irréductibilité (ou de ϕ-irréductibilité) n'est donc pas suffisante et une deuxième condition, celle de l'apériodicité, est requise.

Définition 3. *Une chaîne de Markov est apériodique s'il n'existe pas de sous-ensembles disjoints $\mathsf{X}_1, \mathsf{X}_2, \ldots, \mathsf{X}_d \subseteq \mathsf{X}$ avec $d > 1$ et $\pi(\mathsf{X}_i) > 0$ pour tout i, tel que $P(x, \mathsf{X}_{i+1}) = 1$ pour tout $x \in \mathsf{X}_i$ $(1 \leq i \leq d-1)$ et $P(x, \mathsf{X}_1) = 1$ pour tout $x \in \mathsf{X}_d$.*

Ensemble, ces deux conditions garantissent la convergence d'une chaîne de Markov vers la distribution stationnaire (unique) $\pi(\cdot)$.

Théorème 1. *Une chaîne de Markov ϕ-irréductible et apériodique avec distribution stationnaire π sur un espace X avec σ-algèbre dénombrablement générée, converge asymptotiquement vers π :*

$$\lim_{j \to \infty} ||P^j(x,\cdot) - \pi(\cdot)|| = 0,$$

pour π-presque partout $x \in \mathsf{X}$. De plus, $\lim_{j\to\infty} P^j(x,A) = \pi(A)$ pour tout A mesurable.

Ici, $||\cdot||$ *dénote la distance de variation totale entre deux mesures quelconques* $\mu_1(\cdot), \mu_2(\cdot)$ *:*

$$||\mu_1(\cdot) - \mu_2(\cdot)|| = sup_{A \subseteq \mathsf{X}} |\mu_1(A) - \mu_2(A)|.$$

Théorème 2. *Sous les mêmes conditions que le théorème précédent, pour toute fonction* $h : \mathsf{X} \to \mathbb{R}$ *une forme de loi forte des grands nombres existe*

$$\lim_{N \to \infty} \frac{1}{N} \sum_{j=1}^{N} h(X_j) = E_\pi(h(x)) \equiv \pi(h)$$

presque toujours si $\pi(|h|) < \infty$*.*

La démonstration de ces théorèmes est assez longue et sera omise dans ce mémoire, mais elle se retrouve dans son entièreté dans Roberts et Rosenthal (2004). Les conditions de ces théorèmes sont souvent satisfaites dans le cas des algorithmes MCMC. Par construction, la réversibilité des chaînes MCMC garantit l'existence d'une distribution stationnaire π. Les conditions de ϕ-irréductibilité et d'apériodicité doivent être vérifiées, mais généralement elles seront satisfaites dans presque tous les cas pratiques.

1.4 Taux de convergence des algorithmes MCMC

Bien que le théorème 1 garantisse la convergence asymptotique des chaînes MCMC, il n'y a aucune indication quant à la vitesse de cette convergence. Lorsque cela est possible, il est souhaitable d'obtenir des bornes analytiques pour la vitesse de convergence d'un algorithme. Il existe certaines propriétés qui peuvent s'avérer utiles dans un nombre limité de cas.

Définition 4. *Une chaîne de Markov avec distribution stationnaire* $\pi(\cdot)$ *est dite uniformément ergodique s'il existent* $0 < \rho < 1$ *et* $M < \infty$ *tels que*

$$||P^j(x, \cdot) - \pi(\cdot)|| \leq M\rho^j, \quad j = 1, 2, 3, \ldots \quad \forall x \in \mathsf{X}.$$

Définition 5. *Un sous-ensemble $C \subseteq \mathsf{X}$ est dit petit s'il existe un entier positif $n_0 \in \mathbb{N}$, un réel $\varepsilon > 0$ et une mesure de probabilité $\nu(\cdot)$ sur X tels que*

$$P^{n_0}(x,\cdot) \geq \varepsilon \nu(\cdot) \quad x \in C$$

et donc, $P^{n_0}(x,A) \geq \varepsilon \nu(A)$ pour tout $x \in C$.

De façon intuitive, cette définition implique que les probabilités de transition en n_0 pas à partir de C ont une composante ε en commun. Un exemple tiré de Roberts et Rosenthal (2004) peut permettre de clarifier cette définition.

On suppose que la densité de transition $q(x,y) \propto e^{\frac{-(y-x)^2}{2}}$ et que

$$\pi_u(x) = \begin{cases} |x|^{-\frac{1}{2}}, & |x| < 1 \\ 0, & sinon \end{cases} .$$

Les probabilités de transition satisfont

$$P(x,dy) = q(x,y)\, dy\, min\{1, \frac{\pi_u(y)}{\pi_u(x)}\}$$

Il est facile de voir qu'un voisinage de 0 n'est pas un ensemble petit, puisque $\pi_u(x)$ n'est pas bornée et la probabilité de transition peut être arbitrairement près de 0. D'un autre côté, un ensemble compact C où $\pi_u(x) < k < \infty$ pour tout $x \in C$ est petit. En effet, si D est un autre ensemble compact avec mesure de Lebesgue et π positives et que $\inf_{x \in C, y \in D} q(x,y) = \varepsilon > 0$ alors

$$P(x,dy) \geq \varepsilon\, dy\, min\{1, \frac{\pi_u(y)}{\pi_u(x)}\} \geq \varepsilon\, dy\, min\{1, \frac{\pi_u(y)}{k}\}.$$

Cette dernière est une mesure positive indépendante de x et donc C est bel et bien petit.

Théorème 3. *Si une chaîne de Markov possède une distribution stationnaire $\pi(\cdot)$ et l'ensemble d'états X est un ensemble petit pour $n_0 \in \mathbb{N}$, $\varepsilon > 0$ et $\nu(\cdot)$, alors la chaîne est uniformément ergodique et $||P^j(x,\cdot) - \pi(\cdot)|| \leq (1-\varepsilon)^{\lfloor j/n_0 \rfloor}$ pour tout $x \in \mathsf{X}$ (ici, $\lfloor \cdot \rfloor$ représente la fonction partie entière).*

Démonstration. Voir Roberts et Rosenthal (2004). □

Ce dernier théorème est important puisqu'il permet d'obtenir des bornes analytiques pour la distance de variation totale entre $P^j(x,\cdot)$ et $\pi(\cdot)$ et donc pour le taux de convergence de l'algorithme. En sachant n_0 et ε, on peut trouver j tel que, par exemple, $||P^j(x,\cdot) - \pi(\cdot)|| \leq 0,01$. Dans ce cas, on sera certain que la différence entre la distribution stationnaire et la distribution des états à la j-ième itération sera d'au plus 0,01 (ou tout autre nombre choisi par l'utilisateur). Ce théorème est applicable dans certaines situations particulières, comme par exemple au chapitre 2 du présent mémoire.

Il existe une deuxième propriété relative à la vitesse de convergence qui peut aussi s'avérer utile.

Définition 6. *Une chaîne de Markov avec distribution stationnaire $\pi(\cdot)$ est dite géométriquement ergodique s'il existent $0 < \rho < 1$ et $M(x) < \infty$ pour π-presque partout $x \in \mathsf{X}$ tels que*

$$||P^j(x,\cdot) - \pi(\cdot)|| \leq M(x)\rho^j, \quad j = 1,2,3,\ldots$$

Dans ce mémoire, la propriété de l'ergodicité géométrique restera qualitative, mais il est possible d'arriver à des bornes quantitatives qui peuvent être utiles dans certains contextes, quoiqu'assez difficilement applicables (voir Roberts et Rosenthal (2004)).

1.5 Théorème central limite des algorithmes MCMC

Le théorème central limite pour les méthodes MCMC est vastement applicable et peut permettre de comprendre et de quantifier les erreurs Monte Carlo. En général, l'intérêt d'une méthode MCMC est non seulement d'obtenir un échantillon de la distribution cible mais aussi d'en estimer certains paramètres comme par exemple : la moyenne, la variance, les quantiles et ainsi de suite. Donc, on construira une fonction des valeurs x_j générées et on sera intéressé à la comparer avec sa valeur théorique. Plus précisément, si une chaîne de Markov possède une loi stationnaire $\pi(\cdot)$ et $h : \mathsf{X} \to \mathbb{R}$ est une fonction avec

$\pi(h) = \int_{x \in \mathsf{X}} h(x)\pi(dx) < \infty$, on s'intéressera à l'estimateur naturel $\overline{h} = \frac{1}{N}\sum_{j=1}^{N} h(X_j)$. Il faut rappeler que cet estimateur est justifiable par la loi forte des grands nombres pour les chaînes de Markov (théorème 2).

Définition 7. *Étant donnée une fonction $h : \mathsf{X} \to \mathbb{R}$ et une chaîne de Markov avec distribution stationnaire $\pi(\cdot)$, la fonction h satisfait un théorème central limite si*

$$\lim_{N\to\infty} \frac{1}{\sqrt{N}} \sum_{j=1}^{N} [h(X_i) - \pi(h)] \sim N(0, \sigma^2) \qquad \textit{et } \sigma^2 < \infty.$$

Il est immédiatement clair qu'une telle propriété est intéressante dans un contexte MCMC, car elle permet de construire des intervalles de confiance arbitrairement précis pour un certain paramètre. De plus, elle consent une certaine quantification objective des erreurs Monte Carlo permettant au lecteur indépendant de tirer des conclusions autonomes.

La seule difficulté apparaît lorsque l'on se rend compte de la complexité de la variance de la distribution limite qui est due à la corrélation inhérente de la chaîne de Markov. En général, $\sigma^2 \neq Var_\pi(h)$ et donc la variance empirique de $h(X_j)$, qui est un estimateur naturel, n'est pas applicable. Toutefois, il existe plusieurs autres façons d'estimer σ^2, mais pour le moment la forme de cette variance est présentée.

Théorème 4. *Pour une chaîne de Markov avec distribution stationnaire $\pi(\cdot)$, si une fonction $h : \mathsf{X} \to \mathbb{R}$ satisfait un théorème central limite, alors la variance σ^2 de la distribution limite normale est donnée par :*

$$\sigma^2 = \lim_{N\to\infty} \frac{1}{N} E\left[\left(\sum_{j=1}^{N} [h(X_j) - \pi(h)]\right)^2\right]$$

ou de façon équivalente par $\sigma^2 = \tau Var_\pi(h)$, où $\tau = \sum_{k\in\mathbb{Z}} Corr(X_0, X_k)$.

Une démonstration de ce théorème peut être trouvée dans Chan et Geyer (1994). La quantité τ, aussi appelée le temps d'autocorrélation intégrée, est une mesure de la corrélation de la chaîne. De par la deuxième expression pour σ^2, il est clair que la condition $\sigma^2 < \infty$ de la définition 7 est satisfaite lorsque non-seulement $Var_\pi(h) < \infty$ (ou

$\pi(h^2) < \infty$), mais aussi $\tau < \infty$. En effet, même si $Var_\pi(h) < \infty$, il peut arriver qu'une chaîne MCMC est tellement sous-optimale qu'elle reste confinée à une seule région et $\tau = \infty$ (voir par exemple Roberts (1999)).

L'utilité du théorème central limite pour les chaînes MCMC est maintenant claire, mais les conditions garantissant cette propriété n'ont pas encore été mentionnées. Les théorèmes suivants présentent des conditions suffisantes pour assurer un théorème central limite pour une fonction h et les démonstrations sont disponibles dans Kipnis et Varadhan (1986); Cogburn (1972); Roberts et Rosenthal (1997).

Théorème 5. *Pour une chaîne de Markov réversible, ϕ-irréductible et apériodique, un théorème limite central existe pour une fonction h si $\sigma^2 < \infty$.*

Les conditions de ce théorème sont similaires à celles qui garantissent la convergence d'un algorithme MCMC (théorème 1), donc elles seront satisfaites la plupart du temps. Cependant, il existe des algorithmes MCMC, comme la version déterministe de l'échantillonneur de Gibbs, qui sont convergents même si non réversibles. En plus, l'existence d'un théorème central limite repose sur le fait que $\sigma^2 < \infty$, ce qui nécessite la vérification de $\pi(h^2) < \infty$ et $\tau < \infty$. Toutefois, si le taux de convergence de l'algorithme est connu, il est possible de restreindre les conditions sur $\pi(h^2)$ seulement.

Théorème 6. *Pour une chaîne de Markov avec distribution stationnaire $\pi(\cdot)$ et uniformément ergodique, un théorème central limite existe pour une fonction h si $\pi(h^2) < \infty$.*

Théorème 7. *Pour une chaîne de Markov réversible et géométriquement ergodique, un théorème central limite existe pour une fonction h si $\pi(h^2) < \infty$.*

Il semble donc que le théorème central limite existe pour une quantité considérable de méthodes MCMC. Cependant, même si cette convergence est garantie, il reste à trouver un estimateur pour la variance de la distribution limite. Comme mentionné, l'estimateur usuel par la variance empirique est inadéquat à cause de la corrélation entre les états d'une chaîne de Markov. Il existe plusieurs techniques pour estimer σ^2, mais dans ce

mémoire il sera question d'une des méthodes les plus populaires, soit celle de la moyenne par séries (*Batch Means*). Cette technique est à la fois simple et versatile. Supposons que nous disposons d'un échantillon de taille N généré par un algorithme MCMC. L'idée est de séparer cet échantillon en a séries de taille b, tel que $N = ab$. On définit ensuite

$$\overline{Y}_l = \frac{1}{b} \sum_{i=(l-1)b}^{lb-1} h(x_i)$$

pour $l = 1, \ldots, a$ et l'estimateur de la variance σ^2 par

$$\hat{\sigma}^2_{MS} = N\frac{1}{a}\left(\frac{1}{a-1}\sum_{l=1}^{a}(\overline{Y}_l - \overline{h}_N)^2\right) = \frac{b}{a-1}\sum_{l=1}^{a}(\overline{Y}_l - \overline{h}_N)^2$$

où $\overline{h}_N$ représente $\frac{1}{a}\sum_{l=1}^{a}(Y_l) = \frac{1}{N}\sum_{j=1}^{N}(h(x_j))$.

Cette méthode ne produit pas un estimateur cohérent en général. Cependant, sous certaines conditions relativement souples établies par Damerdji (1994), cet estimateur peut converger même presque toujours. Une fois $\hat{\sigma}^2_{MS}$ calculé, un intervalle de confiance peut être généré de la manière habituelle avec demi-largeur $t_{a-1}\hat{\sigma}_{MS}/\sqrt{N}$, où t_{a-1} est le quantile associé au niveau de confiance désiré d'une distribution Student avec $a-1$ degrés de liberté. Il est à noter qu'il existe différentes versions de la méthode par moyenne de séries, notamment celle par moyenne de séries chevauchées (*Overlapping Batch Means*), avec des propriétés similaires.

Ces méthodes supposent que les moyennes de chaque série sont approximativement indépendantes et identiquement distribuées et donc l'estimateur empirique de la variance de ces moyennes est justifié. Une autre approche est la méthode de Geyer dont l'idée est d'estimer τ et d'utiliser cette estimation en conjonction avec la variance empirique de $h(x_j)$ pour obtenir $\hat{\sigma}^2$ selon le théorème 4. Pour cette technique et d'autres méthodes spectrales, le lecteur est invité à consulter les références suivantes Flegal et Jones (2010); Geyer (1992).

1.6 Diagnostics de convergence

En pratique, il est difficile de prouver l'ergodicité uniforme ou géométrique d'un algorithme MCMC. En effet, dans la plupart des contextes réalistes il est impossible ou très compliqué de connaître la vitesse de convergence de la méthode utilisée afin de savoir quand terminer la simulation. Pour cette raison, plusieurs techniques d'analyse de sorties MCMC ont été développées afin de juger de la performance d'un algorithme donné. Ces méthodes s'appellent collectivement « diagnostics de convergence » et il en existe un nombre considérable. Il est à noter que leur application nécessite une certaine expérience de l'utilisateur. En plus, elles ne donnent aucune garantie rigoureuse quant à la convergence et peuvent parfois induire l'utilisateur en erreur (voir Matthews (1993)). Néanmoins, elles constituent une partie importante de l'implémentation de l'approche MCMC et sont souvent la première façon de valider un algorithme donné. En effet, tel que mentionné à la section 1.2, il existe plusieurs méthodes MCMC et elles auront une performance différente d'un contexte à l'autre. L'utilisation de diagnostics de convergence peut éliminer les algorithmes inefficaces et orienter l'utilisateur vers les méthodes appropriées. Il existe différentes mesures ou diagnostics et ce mémoire en présentera quelques-uns des plus couramment utilisés.

1.6.1 Graphique de la trace d'un paramètre d'intérêt

Une façon simple de vérifier la convergence d'un algorithme MCMC est de produire le graphique de la trace d'un paramètre d'intérêt, c'est-à-dire sa valeur par rapport à l'itération. Si la méthode a convergé vers la distribution stationnaire, le paramètre devra varier aléatoirement dans le temps et le graphique devra ressembler à du bruit de fond. Si le graphique présente des tendances quelconques ou une stagnation pendant de longues périodes, cela est une forte indication que la distribution stationnaire n'est pas atteinte. Par exemple, à la figure 1.1, le graphique de la trace de gauche ne montre à priori aucun problème de convergence, tandis que celui de droite indique un algorithme sous-optimal.

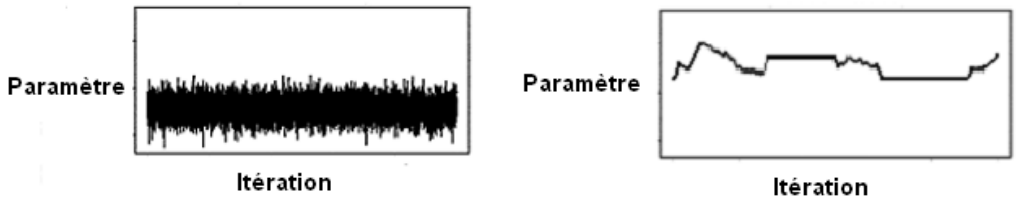

Figure 1.1 – Graphique de la trace de paramètres en fonction des itérations. À gauche un paramètre qui semble fluctuer aléatoirement. À droite un paramètre qui stagne.

1.6.2 Densité (histogramme) de valeurs générées et chaînes parallèles

L'histogramme des valeurs générées par un algorithme MCMC peut donner une idée de la convergence de l'algorithme. Généralement, une densité empirique avec plusieurs modes indique un possible problème de convergence. Aussi, il est possible de comparer des chaînes parallèles de mêmes dimensions mais avec valeurs initiales différentes. Si l'algorithme est optimal, la densité empirique (ou bien un diagramme en boîte) ne devrait pas dépendre de la valeur initiale. La figure 1.2, présente à gauche deux chaînes parallèles qui convergent vers la même distribution et à droite un algorithme sous-optimal.

1.6.3 Taux d'acceptation

La nature de l'approche MCMC requiert une étape d'acceptation ou de refus d'une valeur générée de la distribution instrumentale afin de garantir la réversibilité du processus. Le taux d'acceptation est une mesure facilement extraite de l'implémentation informatique d'un algorithme donné. En général, un taux d'acceptation trop élevé ($> 85\%$)

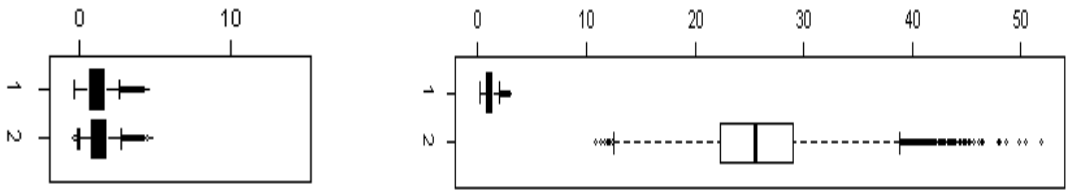

Figure 1.2 – Graphique en boîte de valeurs générées par deux chaînes parallèles avec valeur initiale différente. À gauche les deux chaînes semblent converger vers la même distribution. À droite elles convergent vers deux distributions distinctes.

indique une acceptation de sauts qui sont souvent petits et donc une mauvaise exploration générale de l'espace. D'un autre côté, un taux d'acceptation trop petit ($< 15\%$) signifie le rejet de la plupart des sauts et donc encore une exploration de l'espace inadéquate. Le chapitre 3 abordera ces concepts et quelques lignes directrices générales quant au taux d'acceptation optimal en plus de détails. Évidemment, ce taux dépend souvent du contexte, mais généralement des valeurs extrêmes peuvent être indicatrices d'un problème de convergence.

1.6.4 Autocorrélation et distance de saut carrée moyenne

Tel que mentionné plus tôt, les valeurs générées par un algorithme MCMC possèdent une corrélation inhérente puisqu'elles sont dépendantes au moins par paires. Une manière de quantifier cette corrélation est d'utiliser un estimateur empirique de l'autocorrélation de la chaîne.

Définition 8. *L'autocorrélation entre deux temps j et l d'un processus stochastique chronologique est définie par*

$$Acorr(j,l) = \frac{E\left[(X_j - \mu_j)(X_l - \mu_l)\right]}{\sigma_j \sigma_l},$$

où μ_j, μ_l, σ_j, σ_l sont les moyennes et écarts-types aux temps j et l.

Dans un contexte MCMC, les moyennes et variances sont invariables dans le temps lorsque la chaîne a convergé vers la distribution stationnaire. Donc, l'autocorrélation pour un processus MCMC peut s'exprimer à travers un laps de temps (ou intervalle) k qui sépare deux temps donnés :

$$Acorr(k) = \frac{E\left[(X_j - \mu)(X_{j+k} - \mu)\right]}{\sigma^2} .$$

Un estimateur empirique naturel de l'autocorrélation d'une chaîne MCMC peut donc être défini par

$$\hat{\rho}_k = \widehat{\mathrm{Corr}}(X_j, X_{j+k}) = \frac{\sum_{j=1}^{N-k}(x_j - \bar{x})(x_{j+k} - \bar{x})}{\sum_{j=1}^{N}(x_j - \bar{x})^2} .$$

Même s'il y a convergence, il est normal que l'autocorrélation soit forte pour un intervalle k petit, mais elle devrait diminuer fortement au fur et à mesure que k augmente. Une autocorrélation qui ne s'estompe pas signifie généralement une convergence sous-optimale. À titre d'exemple, à la figure 1.3, le corrélogramme de gauche n'indique pas à priori un problème de convergence, mais celui de droite indique une situation problématique.

Figure 1.3 – Graphique de l'autocorrélation de valeurs générées. À gauche l'autocorrélation diminue de façon satisfaisante. À droite elle demeure excessive même pour un grand intervalle.

La distance de saut carrée moyenne (DSCM) est une autre mesure de la performance d'un algorithme MCMC. Comme le nom l'indique, elle mesure la moyenne de la distance carrée entre deux états consécutifs. Elle est souvent utilisée comme mesure de

performance lorsque plusieurs méthodes s'offrent à l'expérimentateur ou lorsqu'un paramètre de la distribution instrumentale est ajusté. Intuitivement, une grande DSCM indique un algorithme qui explore bien l'espace et est préférable à une DSCM inférieure, toutes autres mesures égales par ailleurs.

Définition 9. *Étant une chaîne de Markov avec distribution stationnaire $\pi(\cdot)$. L'espérance de la distance de saut carrée (EDSC) entre deux états consécutifs j et $j+1$ est définie par*

$$EDSC = E\left[(X_j - X_{j+1})^2\right] \ . \tag{1.5}$$

La distance de saut carrée moyenne (DSCM) est l'estimateur naturel empirique de (1.5). Pour un algorithme MCMC avec N itérations, la DSCM est définie par

$$DSCM = \frac{1}{N}\sum_{j=0}^{N-1}(x_j - x_{j+1})^2.$$

Enfin, il est à noter qu'il existe une relation entre l'autocorrélation et l'espérance de la distance de saut carrée pour un paramètre donné lorsque $X_j, X_{j+1} \sim \pi(\cdot)$.

$$\begin{aligned} E\left[(X_j - X_{j+1})^2\right] &= E\left[X_j^2\right] - 2E\left[X_j X_{j+1}\right] + E\left[X_{j+1}^2\right] \\ &= \sigma^2 + \mu^2 - 2E\left[X_j X_{j+1}\right] + \sigma^2 + \mu^2 \\ &= 2\sigma^2 - 2\left(E\left[X_j X_{j+1}\right] - \mu^2\right) \\ &= 2\sigma^2 - 2Cov(X_j, X_{j+1}) \\ &= 2\sigma^2(1 - Acorr(1)). \end{aligned} \tag{1.6}$$

Donc la DSCM est intimement reliée à l'autocorrélation d'intervalle 1, justifiant son utilité et son mérite en pratique. Dans le contexte multidimensionnel, il est possible de définir la DSCM comme la distance euclidienne carrée entre deux états consécutifs et dans ce cas, elle est équivalente à la somme des DSCM de chaque composante.

1.6.5 Statistiques de convergence

Plusieurs tests statistiques de sorties MCMC ont été développés afin d'obtenir un niveau de confiance plus objectif pour la convergence. Les plus courants sont les tests de Gelman-Rubin (Gelman et Rubin (1992)), de Geweke (Geweke (1992)) et de Raftery-Lewis (Raftery et Lewis (1992)). Le test de Gelman-Rubin est analogue à une analyse de variance, c'est-à-dire qu'il examine l'écart entre la variance dans les chaînes et la variance entre les chaînes ayant des valeurs de départ différentes. Le test de Geweke se base sur l'écart entre la moyenne des premières N_a observations et la moyenne des N_b observations subséquentes d'une seule chaîne telle que $N = N_a + N_b$. Finalement, le test de Raftery-Lewis indique le nombre d'itérations requis pour estimer les quantiles d'une fonction des valeurs générées ainsi que la dépendance entre ces valeurs. La théorie motivant ces tests ne fait pas l'objet de ce mémoire, mais le lecteur est invité à consulter les références mentionnées pour de plus amples détails.

Les diagnostics de convergence présentés s'utilisent en pratique de façon complémentaire. Cela dit, il existe de nombreux exemples où un problème de convergence ne peut être détecté par ces méthodes Cowles et Carlin (1996). Néanmoins, les diagnostics de convergence restent la première tentative à postériori afin de juger de la convergence d'un algorithme. Ces méthodes peuvent être incorporées manuellement dans l'implémentation d'une méthode donnée ou bien utilisées directement sur les sorties MCMC à l'aide de certains logiciels (ex : le greffon CODA pour le logiciel R).

1.7 Variations d'algorithmes MCMC

Il existe plusieurs autres approches MCMC qui sont une variation des algorithmes élémentaires décrits plus tôt. L'objectif de ces méthodes est principalement d'améliorer la convergence des algorithmes existants. Ces techniques sont souvent plus complexes et nécessitent de plus amples efforts pour leur implémentation. Cependant, elles offrent une alternative intéressante dans plusieurs contextes où les algorithmes traditionnels se

comportent de façon médiocre. Le chapitre 2 de ce mémoire présentera un algorithme avec ajustement directionnel récemment développé dans Bédard et Fraser (2008) qui est une variante de l'échantillonneur indépendant. Ensuite, le chapitre 3 présentera un algorithme adaptatif, développé par Haario *et al.* (2001), qui met à jour certains paramètres de la distribution instrumentale basé sur les valeurs générées antérieurement. En plus, ce chapitre étudiera d'autres variations comme l'algorithme à essais multiples (*Multiple-Try MCMC*) et l'algorithme avec rejet différé (*Delayed Rejection MCMC*). Les propriétés théoriques ainsi que les motivations derrière ces méthodes seront abordées dans ces sections.

CHAPITRE 2

ALGORITHME DE TYPE METROPOLIS-HASTINGS AVEC AJUSTEMENT DIRECTIONNEL

L'algorithme de type Metropolis-Hastings avec marche aléatoire (RWMH) est une des méthodes les plus versatiles et les plus couramment utilisées. La fonction de densité instrumentale est de façon répandue une fonction de densité normale ou uniforme, ce qui requiert seulement une optimisation de la variance ou du support respectivement. Cependant, l'application de la méthode RWMH dans presque n'importe quel contexte avec sensiblement les mêmes stratégies d'optimisation résulte en une convergence qui peut être extrêmement lente selon la situation.

L'échantillonneur indépendant (IS) est une méthode dont la densité instrumentale est indépendante de l'état présent. L'application de cette approche nécessite une connaissance plus approfondie de la distribution cible afin de bénéficier d'une convergence raisonnable. En général, plus la densité instrumentale se rapproche de la densité cible plus l'algorithme converge rapidement, mais moins il est facile de générer des valeurs proposées. Cependant, avec un choix adéquat de la densité instrumentale, nécessitant souvent considérablement plus d'efforts, la convergence de l'algorithme IS est souvent supérieure à celle de l'algorithme RWMH.

L'échantillonneur indépendant avec ajustement directionnel (DA)(Bédard et Fraser (2008)) est une nouvelle approche qui vise à combiner la versatilité de l'algorithme RWMH avec la haute performance de l'algorithme IS pour des densités lisses et unimodales. Cette méthode utilise une densité instrumentale Student centrée au mode de la densité cible et dont les queues sont ajustées de façon directionnelle afin d'approximer le plus possible la densité cible. En un premier lieu, ce chapitre présentera une introduction détaillée de la méthode DA. Ensuite, la deuxième section sera consacrée à l'ajustement d'un paramètre clé de l'algorithme et finalement, la dernière partie portera sur la conver-

gence de cette approche.

2.1 L'algorithme DA

Étant donnée une densité cible $\pi(\cdot)$ n-dimensionnelle, unimodale et lisse et en supposant qu'il est possible d'en trouver le mode, il peut être avantageux de tenir compte de cette information dans le choix d'une densité instrumentale. Plus précisément, il serait préférable pour la densité instrumentale en question d'être centrée au mode $\hat{\mathbf{x}} = \arg\sup_{\mathbf{x}} \pi(\mathbf{x})$ et d'avoir la même courbure que la densité cible au mode. La matrice hessienne évalue la courbure de π à un point particulier et est donnée par

$$\begin{bmatrix} \frac{\partial^2 \pi}{\partial x_1^2} & \frac{\partial^2 \pi}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 \pi}{\partial x_1 \partial x_n} \\ \frac{\partial^2 \pi}{\partial x_2 \partial x_1} & \frac{\partial^2 \pi}{\partial x_2^2} & \cdots & \frac{\partial^2 \pi}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \pi}{\partial x_n \partial x_1} & \frac{\partial^2 \pi}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 \pi}{\partial x_n^2} \end{bmatrix}.$$

Cette matrice ne sera typiquement pas définie positive au mode. Par conséquent, il est préférable d'utiliser une autre forme familière, $\hat{H} = -\partial^2 \log \pi(\mathbf{x}) / \partial \mathbf{x} \partial \mathbf{x}' |_{\mathbf{x}=\hat{\mathbf{x}}}$, c'est-à-dire le négatif de la matrice hessienne du logarithme de la densité évaluée au mode.

À ce point, il serait possible d'implémenter un algorithme IS avec une densité instrumentale normale ajustée pour coïncider avec $\hat{\mathbf{x}}$ et $\hat{H}^{-1}$. Il est à noter que $\hat{H} = \Sigma^{-1}$ dans le cas d'une densité normale multivariée. En effet, si $y \sim N(\mu, \Sigma)$, alors

$$\begin{aligned} \hat{H} &= \frac{\partial^2}{\partial \mathbf{y} \partial \mathbf{y}'} - \log\left((2\pi)^{(-n/2)} |\Sigma|^{\frac{1}{2}} e^{\left(-\frac{1}{2}(\mathbf{y}-\mu)'\Sigma^{-1}(\mathbf{y}-\mu)\right)} \right) \\ &= \frac{\partial^2}{\partial \mathbf{y} \partial \mathbf{y}'} \left(\frac{1}{2}(\mathbf{y}-\mu)'\Sigma^{-1}(\mathbf{y}-\mu) \right) \\ &= \frac{1}{2} \frac{\partial}{\partial \mathbf{y}'} (\Sigma^{-1}(\mathbf{y}-\mu) + \Sigma^{-1\prime}(\mathbf{y}-\mu)) \\ &= \frac{1}{2} (\Sigma^{-1\prime} + \Sigma^{-1})' \\ &= \frac{1}{2}\, 2\Sigma^{-1} = \Sigma^{-1}. \end{aligned} \tag{2.1}$$

Cette approche, par la nature des queues courtes de la densité normale, peut sous-estimer les régions se trouvant dans les queues de la densité cible, si ces dernières sont plus épaisses. À l'inverse, l'utilisation d'une distribution instrumentale Cauchy, par exemple, peut résulter inutilement en une trop grande proportion de sauts rejetés si les queues de la densité cible sont plus minces. Afin de remédier à ces problèmes potentiels, l'approche DA vise à construire une densité instrumentale qui n'est pas rigide et qui est adaptée à la densité cible.

Pour ce faire, la densité instrumentale $Student_f(0, I_n)$ est utilisée, dont la forme canonique est :

$$\begin{aligned} q_f(\mathbf{t}) &= \frac{\Gamma\left(\frac{f+n}{2}\right)}{\pi^{n/2}\Gamma\left(\frac{f}{2}\right)}\left(1+t_1^2+\ldots+t_n^2\right)^{-\frac{f+n}{2}} \\ &= \frac{\Gamma\left(\frac{f+n}{2}\right)}{\pi^{n/2}\Gamma\left(\frac{f}{2}\right)}\left(1+\mathbf{t}'\mathbf{t}\right)^{-\frac{f+n}{2}}, \end{aligned} \tag{2.2}$$

où $\mathbf{t}' = (t_1, \ldots, t_n)$.

Aussi, il est à noter qu'il est facile de générer une variable aléatoire $\mathbf{t}'$ à partir de variables indépendantes $z_1, \ldots, z_n$ distribuées selon une loi normale standard et une variable aléatoire χ_f^2 distribuée selon une loi chi-carrée avec f degrés de liberté et indépendante des z_i :

$$\mathbf{t}' = \left(\frac{z_1}{\chi_f}, \ldots, \frac{z_n}{\chi_f}\right). \tag{2.3}$$

La matrice hessienne du négatif du log de la densité Student canonique à $t = 0$ est égale à $(f+n)I_n$ et en général, par un exercice similaire à (2.1), $\hat{H} = (f+n)\Sigma^{-1}$ au mode d'une densité $Student_f(\mu, \Sigma)$. Afin de faire correspondre le mode et la matrice hessienne à ceux de la densité cible, une possibilité est d'utiliser une densité instrumentale $Student_f\left(\hat{\mathbf{x}}, (f+n)\hat{H}^{-1}\right)$ avec valeurs proposées ($\mathbf{y}$) générées de la façon suivante :

$$\mathbf{y} = \hat{\mathbf{x}} + (f+n)^{1/2}\hat{H}^{-1/2}\mathbf{t},$$

où $\hat{H}^{1/2}$ est la matrice racine carrée de droite telle que $\hat{H} = (\hat{H}^{1/2})'(\hat{H}^{1/2})$ et $\mathbf{t}$ est donné par (2.3).

Dans ce cas, la forme de la densité des valeurs proposées $\mathbf{y}$ est alors donnée par

$$q_f(\mathbf{y}) = \frac{\Gamma\left(\frac{f+n}{2}\right)}{\pi^{n/2}\Gamma\left(\frac{f}{2}\right)} \left(1 + \frac{(\mathbf{y}-\hat{\mathbf{x}})'\hat{H}(\mathbf{y}-\hat{\mathbf{x}})}{f+n}\right)^{-\frac{f+n}{2}} \frac{\left|\hat{H}^{1/2}\right|}{(f+n)^{n/2}}.$$

Cependant, d'une façon équivalente et plus simple, il est aussi possible de reparamétriser les densités cible et instrumentale en posant $\mathbf{x}^* = \hat{H}^{1/2}(\mathbf{x}-\hat{\mathbf{x}})$ et $\mathbf{T} = (f+n)^{1/2}\,\mathbf{t}$. Cela implique que $\hat{\mathbf{x}}^* = \mathbf{0}$, $\hat{H}^* = I_n$ et $\hat{\mathbf{T}} = \mathbf{0}$, $\hat{H} = I_n$. De plus, la densité cible reparamétrisée est proportionnelle à l'originale, ce qui signifie que le ratio d'acceptation (1.4) et l'approximation directionnelle (2.6) qui sera détaillée sous peu seront calculables à partir de $\pi(\cdot)$:

$$\pi^*(\mathbf{x}^*) = \pi\left(\hat{\mathbf{x}} + \hat{H}^{-1/2}\mathbf{x}^*\right)\left|\hat{H}^{-1/2}\right| \propto \pi\left(\hat{\mathbf{x}} + \hat{H}^{-1/2}\mathbf{x}^*\right) = \pi(\mathbf{x}).$$

En outre, il est possible à tout moment d'accéder à une variable $\mathbf{x}$ à partir de $\mathbf{x}^*$ par la transformation inverse : $\mathbf{x} = \hat{\mathbf{x}} + \hat{H}^{-1/2}\mathbf{x}^*$.

Cette transformation implique que la nouvelle densité instrumentale est donnée par

$$q_f(\mathbf{T}) = \frac{\Gamma\left(\frac{f+n}{2}\right)}{\pi^{n/2}\Gamma\left(\frac{f}{2}\right)} \left(1 + \frac{\mathbf{T}'\mathbf{T}}{f+n}\right)^{-\frac{f+n}{2}} (f+n)^{-n/2}. \tag{2.4}$$

Donc, plusieurs choix existent quant à la reparamétrisation des densités cible et instrumentale. Celle qui a été décrite à l'instant sera utilisée dans la description ainsi que dans l'implémentation de l'algorithme DA tout au long de ce mémoire.

Il serait maintenant possible de spécifier un degré de liberté global pour la densité instrumentale et d'utiliser un algorithme IS. Cependant, l'approche DA vise justement à en faire davantage en optimisant la densité instrumentale dans chaque direction. Pour ce faire, il est intéressant d'exprimer les états de la chaîne de Markov en termes de deux composantes : la direction et la distance par rapport au mode. Ainsi, la direction d'un

état présent $\mathbf{x}_j$ est donnée par $\mathbf{u}_j = \mathbf{x}_j / \left|\mathbf{x}_j\right|$ où $|\mathbf{x}| = \sqrt{x_1^2 + \ldots + x_n^2}$ et sa distance radiale par $s_j = \left|\mathbf{x}_j\right|$. Donc, en posant $\mathbf{x}_j = \mathbf{u}_j \cdot s_j$, il est possible d'exprimer la densité cible en terme de la densité de chaque composante, c'est-à-dire $\tilde{\pi}\left(\mathbf{u}_j, s_j\right) = \tilde{\pi}\left(s_j \left| \mathbf{u}_j \right.\right) \tilde{\pi}\left(\mathbf{u}_j\right)$. De façon similaire, une valeur proposée peut également être exprimée en termes de direction et distance, $\mathbf{y}_{j+1} = \mathbf{u}_{j+1}^{prop} \cdot s_{j+1}^{prop}$ et la densité instrumentale correspondante devient

$$\tilde{q}\left(\mathbf{u}_{j+1}^{prop}, s_{j+1}^{prop}\right) = \tilde{q}\left(s_{j+1}^{prop} \left| \mathbf{u}_{j+1}^{prop} \right.\right) \tilde{q}\left(\mathbf{u}_{j+1}^{prop}\right).$$

Le but de cette technique est de diviser l'étape de proposition d'un nouvel état en deux parties afin de permettre un ajustement directionnel. Donc, au lieu de générer de façon conventionnelle $\mathbf{y}_{j+1}$ selon $q(\mathbf{y}_{j+1})$, cette méthode permet de générer en premier une direction $\mathbf{u}_{j+1}^{prop}$ selon $\tilde{q}\left(\mathbf{u}_{j+1}^{prop}\right)$ et ensuite une distance radiale s_{j+1}^{prop} selon $\tilde{q}\left(s_{j+1}^{prop} \left| \mathbf{u}_{j+1}^{prop} \right.\right)$. À ce point, il reste seulement à poser la forme des deux densités instrumentales. Il est bon de rappeler que l'objectif est de choisir ces densités de façon à ce qu'elles approchent le mieux possible les densités cibles correspondantes. Généralement, une bonne approximation de la distribution directionnelle cible est une distribution uniforme sur une sphère de rayon 1 dans $\mathbb{R}^n$. En d'autres mots, on considère que $\tilde{\pi}(\mathbf{u}) \approx \tilde{q}(\mathbf{u}) = 1/A_n$, où $A_n = 2\pi^{n/2}/\Gamma(n/2)$ représente l'aire de la sphère. Ensuite, dans un contexte général, la densité $\tilde{q}(s|\mathbf{u})$ est choisie comme la meilleure approximation disponible de la densité conditionnelle $\tilde{\pi}(s|\mathbf{u})$.

Dans le contexte présent, la densité instrumentale globale $q\left(\mathbf{y}_{j+1}\right)$ est voulue de la forme $Student_f$. Par conséquent, en utilisant l'approximation par une distribution uniforme sur la sphère de rayon 1, une direction proposée est générée par :

$$(\mathbf{u}_{j+1}^{prop})' = \frac{(z_1, \ldots, z_n)}{|\mathbf{z}|}, \text{ où } \mathbf{z} = (z_1, \ldots, z_n) \text{ et } z_i \sim N(0,1). \tag{2.5}$$

Ensuite, une distance radiale proposée est générée par $s_{j+1}^{prop} = (f+n)^{1/2}|\mathbf{z}|/\chi_f$ où $|\mathbf{z}|$ est le même vecteur ayant servi à obtenir la direction. En utilisant cette méthode, il

est clair, de (2.3), que la distribution instrumentale globale $q\left(\mathbf{y}_{j+1}\right)$ est une loi $Student_f$ avec densité donnée par (2.4).

Enfin, l'ajustement directionnel est effectué par la sélection du degré de liberté f. Le choix optimal est différent d'une direction à l'autre afin d'approximer le mieux possible la densité conditionnelle $\tilde{\pi}\left(s_{j+1}^{prop}\left|\mathbf{u}_{j+1}^{prop}\right.\right)$. En fait, f est choisi tel que les décroissances des densités cible et instrumentale à un point s^* relativement au mode soient les mêmes. En d'autres mots, le ratio de la densité cible à une distance donnée s^* et au mode est égal à celui de la densité instrumentale

$$\begin{aligned}\frac{\pi^*\left(\mathbf{u}_{j+1}^{prop}\cdot s^*\right)}{\pi^*\left(\mathbf{0}\right)} &= \frac{\pi\left(\hat{\mathbf{x}}+\hat{H}^{-1/2}\left(\mathbf{u}_{j+1}^{prop}\cdot s^*\right)\right)}{\pi\left(\hat{\mathbf{x}}\right)} = \frac{q_{f_{j+1}^{prop}}\left(\mathbf{u}_{j+1}^{prop}\cdot s^*\right)}{q_{f_{j+1}^{prop}}\left(\mathbf{0}\right)} \\ &= \left(1+\frac{\left(\mathbf{u}_{j+1}^{prop}\cdot s^*\right)'\left(\mathbf{u}_{j+1}^{prop}\cdot s^*\right)}{f_{j+1}^{prop}+n}\right)^{-\frac{f_{j+1}^{prop}+n}{2}} \qquad (2.6)\end{aligned}$$

où $q_{f_{j+1}^{prop}}$ est donnée par (2.4).

Une solution analytique de f_{j+1}^{prop} est ardue, mais en pratique il est suffisant de choisir la valeur la plus appropriée parmi les entiers entre 1 (distribution de Cauchy) et 50 (distribution presque normale).

En posant $r^2 = 2\log\left\{\pi\left(\hat{\mathbf{x}}+\hat{H}^{-1/2}\left(\mathbf{u}_{j+1}^{prop}\cdot s^*\right)\right)/\pi\left(\hat{\mathbf{x}}\right)\right\}$, $Q^2 = \left(\mathbf{u}_{j+1}^{prop}\cdot s^*\right)'\left(\mathbf{u}_{j+1}^{prop}\cdot s^*\right)$ et $f_{j+1}^{prop}+n=\overline{f}$, il suffit de trouver la meilleure solution $f_{j+1}^{prop}\in\{1,2,\ldots,50\}$ de l'équation

$$\overline{f}\log\left(1+\frac{Q^2}{\overline{f}}\right)=r^2. \qquad (2.7)$$

Donc, la méthode DA utilise une densité instrumentale globale qui peut être vue comme un amalgame de densités $Student_f$, chacune ayant comme support une direction donnée à partir du mode et un degré de liberté f spécifié par cette direction. L'approche

générale peut être illustrée de façon simpliste pour le cas unidimensionnel par la figure 2.1, reproduite de Bédard et Fraser (2008).

Le seul paramètre libre restant est s^*, c'est-à-dire la distance où le ratio (2.6) des densités cible et instrumentale est évalué. En un premier lieu, on considère le cas où π est une densité normale multivariée avec composantes indépendantes distribuées selon une loi normale standard. Dans ce cas, les degrés de liberté proposés par la méthode DA seront fixes à 50 dans chaque direction et la distribution instrumentale sera approximativement π. Donc, dans ce contexte, u_{j+1}^{prop} peut être généré selon (2.5), et la distance radiale proposée peut être approximée par $(s_{j+1}^{prop}) = |\mathbf{z}|$. La distribution de (s_{j+1}^{prop}) sera donc χ avec n degrés de liberté et une espérance approximative de $\sqrt{n}$. Pour cette raison, les auteurs de la méthode DA suggèrent d'exprimer la valeur s^* en termes de cette espérance et d'un paramètre de rodage λ, c'est-à-dire $s^* = \lambda\sqrt{n}$. Dans Bédard et Fraser (2008), il est aussi suggéré d'utiliser une valeur de λ autour de 2 ou 3.

En fait, le choix de ce paramètre peut avoir un impact considérable sur la performance de la méthode DA et une grande partie de ce chapitre sera consacrée à ce sujet. Cependant, une brève description de l'algorithme est tout d'abord nécessaire afin de clarifier et de résumer les étapes de cette approche.

2.2 Étapes de l'algorithme DA

En un premier lieu, il faut déterminer les valeurs $\hat{\mathbf{x}}$, $\hat{H}^{1/2}$ et $\hat{H}^{-1/2}$. Dans le logiciel R, ces quantités peuvent être évaluées à l'aide des fonctions nlm, fdHess, pdMat et pdFactor (greffon nlme) .

Ensuite, l'algorithme procède comme suit :

1 Déterminer la direction de la valeur initiale reparamétrisée $\mathbf{x}_0^* = \hat{H}^{1/2}(\mathbf{x}_0 - \hat{\mathbf{x}})$, c'est-à-dire $\frac{\mathbf{x}_0^*}{|\mathbf{x}_0^*|}$ et trouver $f_0 \in \{1, 2, \cdots, 50\}$, la meilleure solution de (2.7).

2 Pour $j \geq 0$

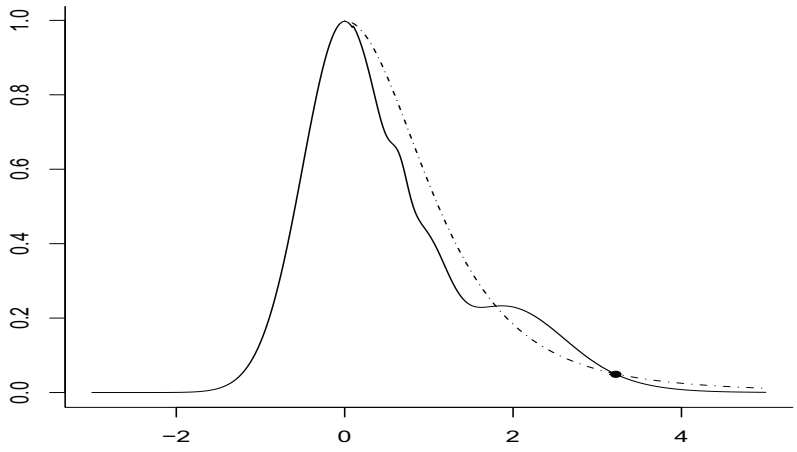

Figure 2.1 – Graphique de la densité cible (ligne pleine) et instrumentale (ligne pointillée) lorsque la direction choisie est vers la droite (+1). Le point d'intersection est $\mathbf{u}_{j+1}^{prop} \cdot s^* = s^*$.

a) Générer $\mathbf{u}_{j+1}^{prop}$ à partir de la relation (2.5) et stocker la valeur $|\mathbf{z}_{j+1}|$.

b) Trouver $f_{j+1}^{prop} \in \{1, 2, \cdots, 50\}$, la meilleure solution de (2.7).

c) Obtenir une distance radiale $s_{j+1}^{prop} = (f_{j+1}^{prop} + n)^{1/2} |\mathbf{z}_{j+1}| / \chi_{f_{j+1}^{prop}}$, où $\chi_{f_{j+1}^{prop}}$ est une variable indépendante distribuée selon une loi Chi et $|\mathbf{z}_{j+1}|$ est la valeur trouvée en (a).

d) Poser $\mathbf{y}_{j+1}^* = \mathbf{u}_{j+1}^{prop} \cdot s_{j+1}^{prop}$.

e) Calculer la probabilité d'acceptation

$$\alpha\left(\mathbf{x}_j^*, \mathbf{y}_{j+1}^*\right) = 1 \wedge \frac{\pi\left(\hat{\mathbf{x}} + \hat{H}^{-1/2}\mathbf{y}_{j+1}^*\right) q_{f_j}\left(\mathbf{x}_j^*\right)}{\pi\left(\hat{\mathbf{x}} + \hat{H}^{-1/2}\mathbf{x}_j^*\right) q_{f_{j+1}^{prop}}\left(\mathbf{y}_{j+1}^*\right)}, \tag{2.8}$$

où $q_f(\mathbf{x})$ est donnée par (2.4).

f) Générer une valeur r_{j+1} d'une loi $U[0,1]$.

g) Si $r_{j+1} \leq \alpha\left(\mathbf{x}_j^*, \mathbf{y}_{j+1}^*\right)$, accepter la valeur proposée et poser $\mathbf{x}_{j+1}^* = \mathbf{y}_{j+1}^*$, $f_{j+1} = f_{j+1}^{prop}$. Sinon, rejeter la valeur proposée et poser $\mathbf{x}_{j+1}^* = \mathbf{x}_j^*$, $f_{j+1} = f_j$.

h) Obtenir $\mathbf{x}_{j+1} = \hat{\mathbf{x}} + \hat{H}^{-1/2}\mathbf{x}_{j+1}^*$.

i) Poser $j = j+1$ et revenir à l'étape (a), pour N itérations.

2.3 Impact de la valeur de λ sur la performance de l'algorithme DA

Il a été brièvement mentionné que le paramètre λ pouvait avoir un impact sur la performance de la méthode DA. Le reste de ce chapitre se penchera sur ce sujet. En premier, deux exemples tirés de l'article Bédard et Fraser (2008) seront analysés en fonction du paramètre λ et enfin un exemple unidimensionnel sera construit afin de démontrer les problèmes de convergence potentiels de l'algorithme.

2.3.1 Exemple 1

Dans un premier problème de l'article Bédard et Fraser (2008), on cherche à obtenir un échantillon d'une distribution de paramètres de régression. À l'origine, cet exemple a été étudié dans Bédard *et al.* (2007) et se base sur un modèle de régression $y_i = \alpha + \beta x_i + \sigma z_i$, où y est la variable réponse, x la variable explicative et z représente l'erreur distribuée selon une loi $Student_7$. Les données suivantes ont été en fait générées avec $\alpha = 0, \beta = 1, \sigma = 1$:

x_i	-3	-2	-1	0	1	2	3
y_i	-2,68	-4,02	-2,91	0,22	0,38	-0,28	0,03

.

On suppose qu'il y a un intérêt particulier pour l'hypothèse $\beta = 1$ et que, d'un point de vue bayésien, la distribution à postériori des trois paramètres $(\alpha, \beta, \tau = \log\sigma)$ est donnée par

$$\pi_1\left(\alpha, \beta, \tau \,|\mathbf{y}^0\right) d\alpha\, d\beta\, d\tau = c\, e^{-7\tau} \prod_{i=1}^{7} \left\{1 + \frac{(y_i^0 - \alpha - \beta x_i)^2}{7e^{2\tau}}\right\}^{-4} d\alpha\, d\beta\, d\tau. \quad (2.9)$$

Dans ce contexte, π_1 sera la densité cible et l'objectif est une estimation de la valeur-s pour l'hypothèse $\beta = 1$, c'est-à-dire $s(\beta = 1) = P(\beta \geq 1)$. Étant donné un échantillon de la distribution cible, une approximation de cette quantité est donnée par la valeur-s empirique :

$$s(\beta) = \frac{1}{N}\sum_{j=1}^{N} I\left(\beta_j \geq \beta\right) = \frac{1}{N}\sum_{j=1}^{N} I\left(\beta_j \geq 1\right),$$

où N est la taille de l'échantillon généré. Pour cet exemple, il a été démontré dans Bédard et Fraser (2008) que la méthode DA offre une meilleure performance comparativement à d'autres méthodes RWMH et IS sélectionnées par les auteurs, résultant en une variance empirique de $s(\beta)$ inférieure.

Afin de mesurer l'impact du paramètre $s^* = \lambda\sqrt{n}$ sur la performance de l'algorithme DA, trois mesures empiriques ont été calculées pour différentes valeurs de λ. En premier, l'écart-type de la valeur-s pour $\beta = 1$ a été obtenue. Dans ce contexte, une légère variante de la mesure de variance par moyenne de séries présentée au chapitre 1 a été utilisée. Une période de chauffe initiale de 10 000 itérations a été écartée et ensuite, 4 000 000 valeurs générées ont été divisées en 4000 séries de 1000 observations. Dans chacune de ces séries, les cinquante premières valeurs ont été rejetées afin de minimiser encore davatage la corrélation entre les séries. Pour des raisons de cohérence, ces paramètres seront fixes sauf avis contraire pour tous les algorithmes simulés dans le cadre de ce mémoire. Donc, en gardant la même notation qu'au chapitre 1, $a = 4000$, $b = 950$ et la variance par moyenne de séries est donnée par :

$$\hat{\sigma}^2_\lambda = \frac{950}{3999}\sum_{l=1}^{4000}(s_l - \bar{s}_N)^2$$

où s_l est la valeur-s empirique de la l-ième série et $\bar{s}_N = \frac{1}{4000}\sum_{l=1}^{4000} s_l$ est la valeur-s moyenne globale. La figure 2.2 présente l'écart-type empirique de $s(\beta)$ et démontre qu'une valeur de $\lambda \in [2;4]$ semble minimiser cette mesure.

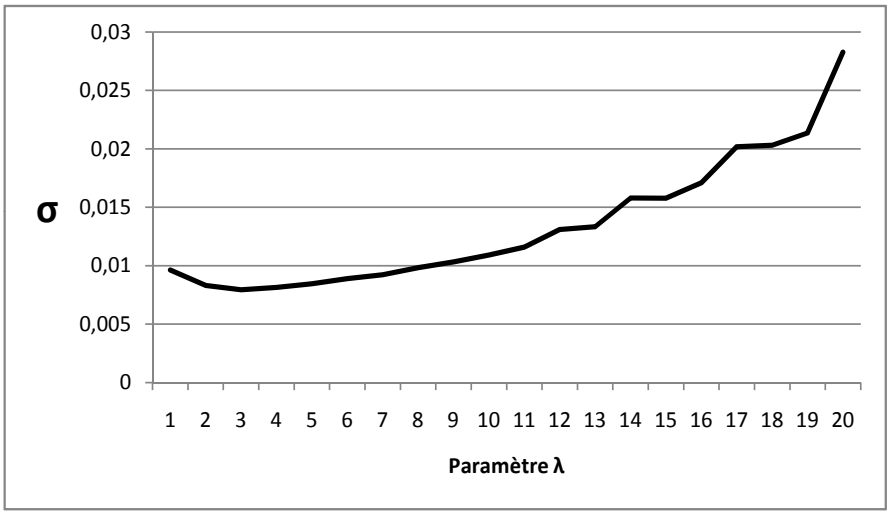

Figure 2.2 – Graphique de l'écart-type σ de $s(\beta = 1)$ en fonction de λ (algorithme DA, exemple 1).

Ensuite, le taux d'acceptation de l'algorithme a été utilisé à titre de deuxième mesure empirique. La figure 2.3 indique que le taux d'acceptation semble être maximal autour des mêmes valeurs λ.

Un grand taux d'acceptation est généralement insuffisant pour garantir une bonne performance d'un algorithme MCMC. Cependant, dans le cas présent, si l'algorithme explore bien l'espace et ne sous-estime pas les queues tel qu'argumenté dans Bédard et Fraser (2008), on peut présumer qu'un meilleur taux d'acceptation est préférable.

Selon la figure 2.4, qui présente la distance de saut carrée moyenne, une valeur de λ de 3 ou 4 semble encore être idéale. Comme mentionné à l'équation (1.6), la DSCM est reliée à l'autocorrélation de la chaîne. On peut donc aussi s'attendre à une autocorrélation totale d'intervalle 1 plus faible lorsque l'algorithme emploie une valeur de λ entre 2 et 4.

D'un autre côté, la densité instrumentale de l'algorithme DA est ajustée selon la

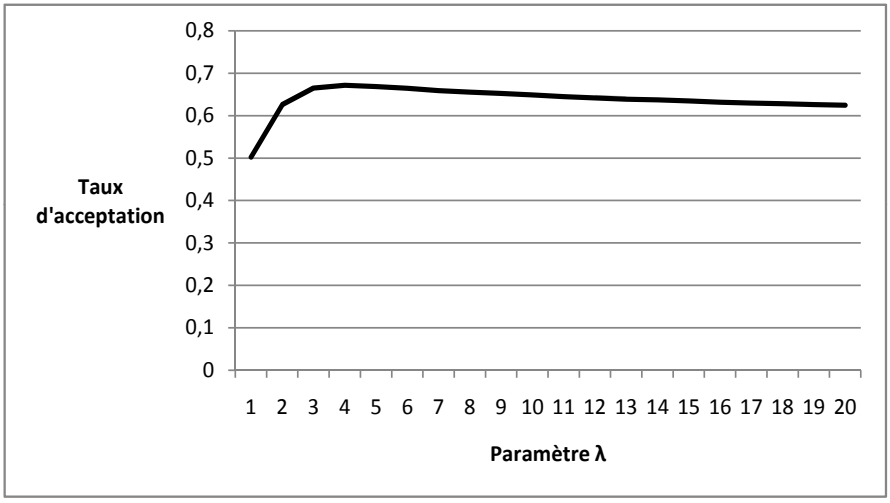

Figure 2.3 – Graphique du taux d'acceptation en fonction de λ (algorithme DA, exemple 1).

direction choisie et une mesure potentiellement intéressante est la moyenne des degrés de liberté de la densité instrumentale en fonction de la valeur de λ. Pour cet exemple, il a été montré dans Bédard et Fraser (2008) que la moyenne des degrés de liberté de la densité instrumentale pour une valeur de $\lambda = 2$, se situe à 28,57. Cependant, lorsque l'on suit son évolution, on peut discerner une tendance de cette moyenne à augmenter en fonction du paramètre λ. La figure 2.5 présente ces résultats. Ce graphique indique que la vitesse moyenne de décroissance des queues par rapport au mode n'est pas constante. En d'autres mots, la densité cible décroît de façon relativement lente au début, s'apparentant à une densité $Student_{27}$, mais au fur et à mesure que la distance par rapport au mode augmente, la vitesse de décroissance des queues augmente et à $\lambda = 20$, la densité cible s'apparente plutôt à une densité $Student_{33}$. Cette idée est importante et sera explorée tout au long de ce chapitre.

Tel que mentionné auparavant, l'algorithme DA a été comparé et jugé supérieur à des

Figure 2.4 – Graphique de la DSCM en fonction de λ (algorithme DA, exemple 1).

algorithmes RWMH et IS spécifiques implémentés dans Bédard et Fraser (2008). Dans un même ordre d'idées, il peut être intéressant de comparer la méthode DA avec des algorithmes IS se situant aux extrémités de la famille des distributions Student, par exemple $q_1(\mathbf{x}) = Student_1(\hat{\mathbf{x}}, (f+n)\hat{H}^{-1})$(Cauchy) et $q_{50}(\mathbf{x}) = Student_{50}(\hat{\mathbf{x}}, (f+n)\hat{H}^{-1})$ (qui sera, pour des raisons de simplicité, appelée normale).

La supériorité de l'approche DA devient évidente lorsque les trois algorithmes sont comparés selon les diagnostics de convergence présentés précédemment. La figure 2.6 démontre que la distance de saut carrée moyenne est plus élevée pour n'importe quelle valeur de λ pour la méthode DA.

Ensuite, la figure 2.7 indique que le taux d'acceptation est plus élevé pour la méthode DA que pour l'échantillonneur indépendant avec distribution instrumentale Cauchy. Étant donné que ces deux algorithmes ne sous-estiment pas les queues de la densité cible, nous déduisons que la méthode avec le taux d'acceptation le plus élevé devrait être préférée. Notons qu'ici, le taux d'acceptation doit être interprété avec prudence dans le

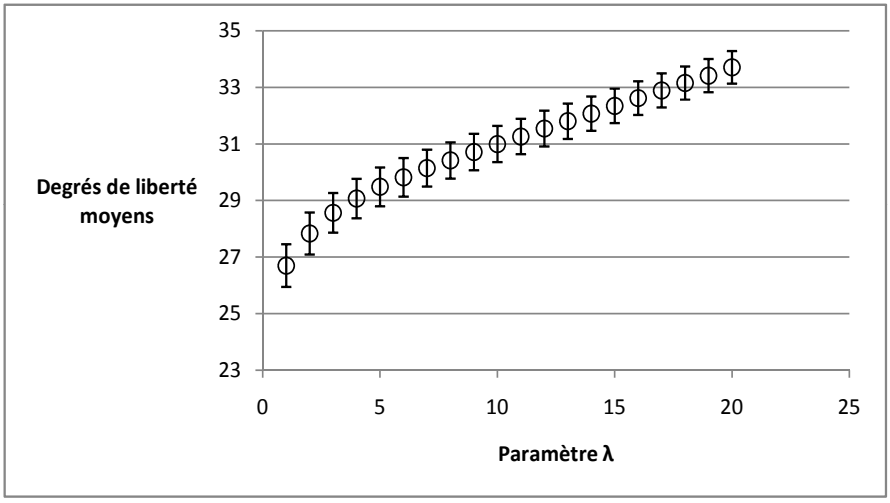

Figure 2.5 – Graphique des degrés de liberté moyens (avec un écart-type) en fonction de λ (algorithme DA, exemple 1).

cas de l'échantillonneur indépendant avec distribution instrumentale normale. En effet, cet algorithme aura tendance à sous-estimer les queues de l'exemple considéré et le taux d'acceptation s'en trouvera (artificiellement) gonflé.

Finalement, la figure 2.8 révèle que l'écart-type de $s(\beta = 1)$ obtenu par la méthode DA est inférieur à celui obtenu par l'algorithme IS avec distribution Cauchy pour une valeur $\lambda \leq 10$ et supérieur pour $\lambda > 10$. Cependant, l'algorithme IS avec densité instrumentale normale résulte en un écart-type 6 fois plus élevé que les deux autres approches. Ce dernier résultat indique un problème potentiel de convergence dans le cas de l'algorithme IS avec densité instrumentale normale. En fait, pour cet exemple, cette méthode converge beaucoup plus lentement que les deux autres approches pour des raisons qui seront abordées plus loin dans ce chapitre.

Donc, ces résultats réitèrent que l'approche DA, malgré un effort computationnel accru, peut mener à des résultats supérieurs à d'autres approches IS lorsque peu d'in-

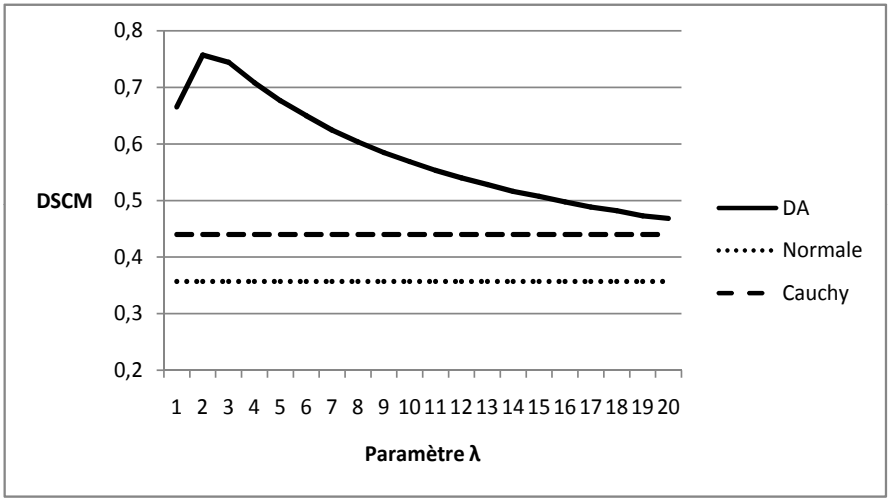

Figure 2.6 – Graphique de la DSCM des algorithmes IS et DA (exemple 1).

formations au sujet de la densité cible sont disponibles. De plus, il semble qu'un choix optimal existe pour le paramètre λ et puisse avoir un impact sur la performance de la méthode. L'écart-type de la valeur-s a aussi été calculé pour d'autres valeurs de β avec des résultats identiques.

2.3.2 Exemple 2

Le second exemple analysé est aussi tiré de l'article Bédard et Fraser (2008), à l'origine étudié dans Cox et Snell (1981). Dans ce contexte, on cherche encore à obtenir des mesures relativement à des paramètres de régression. Les données concernent 32 réacteurs à eau légère situés aux États-Unis. La variable réponse est le coût d'investissement d'un nouveau réacteur et il y 10 variables explicatives dont 7 sont considérées significatives. Le modèle de régression est posé par $\mathbf{y} = X\beta + \sigma\mathbf{z}$, où X est une matrice 32x7 contenant les variables explicatives et $z \sim Student_4$ représente l'erreur. La distribution

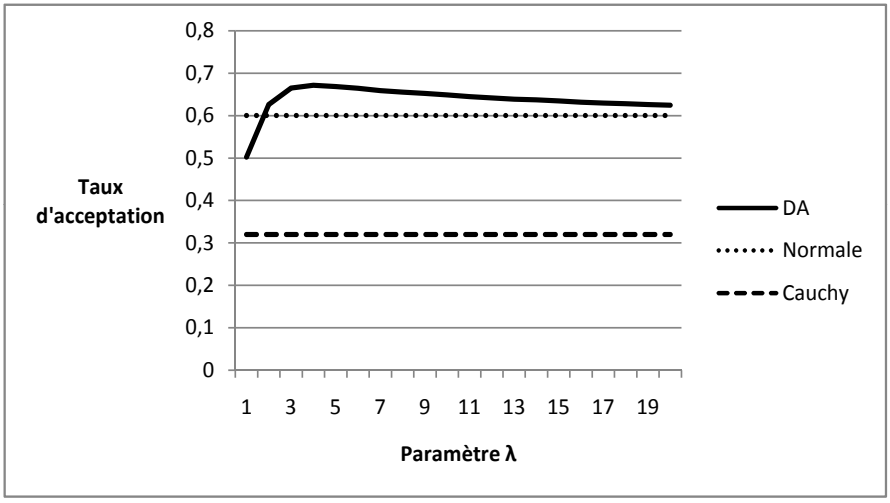

Figure 2.7 – Graphique du taux d'acceptation des algorithmes IS et DA (exemple 1).

de la variable réponse est

$$f(\mathbf{y};\alpha,\beta,\sigma)\,d\mathbf{y} = \sigma^{-7}\prod_{i=1}^{7} h\left(\frac{y_i-\alpha-x_i\beta}{\sigma}\right)dy_i,$$

où $h(\cdot)$ est la densité $Student_4$

D'un point de vue classique, après une reparamétrisation dont les détails seront omis (voir Bédard et Fraser (2008)), la distribution cible est donnée par

$$\pi_2\left(b,a\,\middle|\,\mathbf{d}^0\right)db\,da = c\prod_{i=1}^{n} h\left(e^a d_i^0 + X_i b\right)e^{a(n-r)}db\,da\,. \tag{2.10}$$

La densité π_2 est en huit dimensions, b représente le vecteur des sept coefficients de régression et $s = exp(a)$ représente l'écart-type de l'erreur de régression. Le terme d_i^0 représente le i-ème résidu observé standardisé ; ces résidus satisfont $\mathbf{d}^0 = \left(\mathbf{y}^0 - Xb^0\right)/s^0$ où b^0 dénote les estimateurs par moindres carrés et $(s^0)^2 = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2/(n-r)$, avec $n = 32$ et $r = 7$. Le terme X_i dénote la i-ème rangée de la matrice des variables explicatives. Le lecteur est invité à consulter Bédard et Fraser (2008) pour de plus amples

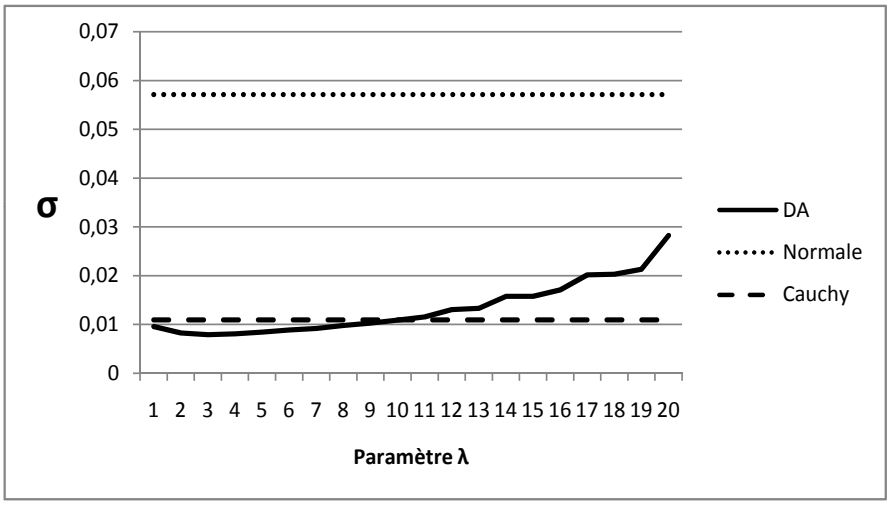

Figure 2.8 – Graphique de l'écart-type σ de $s(\beta = 1)$ des algorithmes IS et DA (exemple 1).

détails ainsi que le tableau complet des valeurs observées (y_i) et explicatives (x_i). Pour l'objectif de ce chapitre, la forme générale (2.10) de la distribution cible est suffisante.

Dans ce contexte, il y a un intérêt particulier pour le sixième coefficient de régression, c'est-à-dire la dépendance du coût d'investissement par rapport à la sixième variable explicative qui dans ce cas correspond au nombre de réacteurs construits par un ingénieur donné. L'objectif est d'obtenir une estimation de la valeur-p pour l'hypothèse nulle $\beta_6 = -0,01$. De manière semblable à l'exemple précédent, un échantillon sera généré de la distribution cible et une valeur-p empirique sera utilisée comme estimateur.

Pour cet exemple, il a été démontré dans Bédard et Fraser (2008) que la performance de l'algorithme avec ajustement directionnel est supérieure à celle des algorithmes IS et RWMH particuliers considérés dans cet article. En fait, l'approche RWMH peut présenter de graves problèmes de convergence pour cet exemple et les détails en seront abordés au chapitre 3.

Afin de déterminer l'effet du paramètre λ sur la performance de l'algorithme DA, les mêmes diagnostics de convergence empiriques qu'à la section 2.3.1 ont été utilisés. Les figures 2.9 et 2.10 démontrent que la distance de saut carrée moyenne et le taux d'acceptation sont maximisés lorsque λ est petit, autour de 2 ou 3. La figure 2.11 montre que l'écart-type de la valeur-p calculée dépend de façon moins marquée de λ, mais atteint un minimum pour $\lambda = 2$ ou 3. Les relations de l'écart-type de la valeur-p en fonction de λ pour d'autres valeurs de β_6 ont été omises puisque similaires.

De manière semblable à l'exemple 2.3.1, la figure 2.12 de la moyenne des degrés de liberté des densités instrumentales donne un aperçu du comportement des queues de la densité cible. Dans ce cas, les queues de la densité cible sont nettement moins épaisses qu'à l'exemple 2.3.1. En plus, il semble que les queues ont un taux de décroissance rapide jusqu'à $\lambda = 13$ et ce taux ralentit à partir de $\lambda > 13$.

Dans cette situation, il existe un choix optimal pour le paramètre λ, tout comme pour l'exemple 2.3.1, et il se situe autour des mêmes valeurs $\lambda \in [2;4]$.

2.4 Convergence de l'algorithme DA

Dans la section 2.3, trois diagnostics de convergence ont été utilisés pour juger de l'importance du paramètre λ dans deux exemples réalistes. À ce point, il semble que le choix de ce paramètre peut légèrement influencer la performance de l'algorithme et qu'un choix sécuritaire semble être situé entre 2 et 4. Dans cette section, un exemple sera construit pour démontrer que le choix de ce paramètre peut parfois être crucial et se trouver très loin de ces valeurs « sécuritaires ».

Cependant, un résultat théorique est d'abord nécessaire. Au chapitre 1, des notions d'uniformité ergodique et géométrique ont été présentées de façon générale. Bien que leur application soit souvent difficile en pratique, dans le cas d'un algorithme IS et par conséquent dans le cas de la méthode DA, il existe certaines conditions simples qui garantissent ces propriétés.

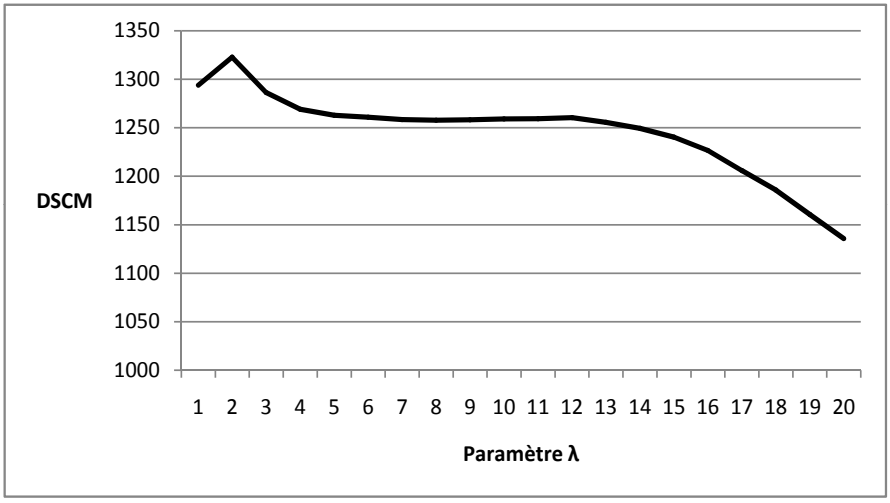

Figure 2.9 – Graphique de la DSCM en fonction de λ (algorithme DA, exemple 2).

Théorème 8. *L'échantillonneur indépendant est uniformément ergodique s'il existe une constante* $\beta > 0$ *telle que*

$$\frac{q(x)}{\pi(x)} \geq \beta, \quad \forall x \in \mathsf{X}, \tag{2.11}$$

et alors $||P^N(x,\cdot) - \pi(\cdot)|| \leq (1-\beta)^N$*, où N est le nombre d'itérations.*

D'un autre côté, si $\operatorname{ess\,inf}\{q(x)/\pi(x)\} = 0$ *en* π*-mesure, l'algorithme n'est même pas géométriquement ergodique. La borne inférieure essentielle est définie comme le supremum des presque minorants d'une fonction, et donc, dans ce cas,* $\operatorname{ess\,inf}_\pi\{q(x)/\pi(x)\} = sup\{a \in \mathbb{R} | \pi(\{x | \frac{q(x)}{\pi(x)} < a\}) = 0\}$

Démonstration. Voir Mengersen et Tweedie (1996). □

Donc, afin de garantir une convergence uniformément ergodique, il suffit de trouver une constante $0 < \beta \leq 1$ (la dernière borne résulte du fait que les deux fonctions sont des densités normalisées et intègrent à 1) telle que le ratio des densités instrumentale

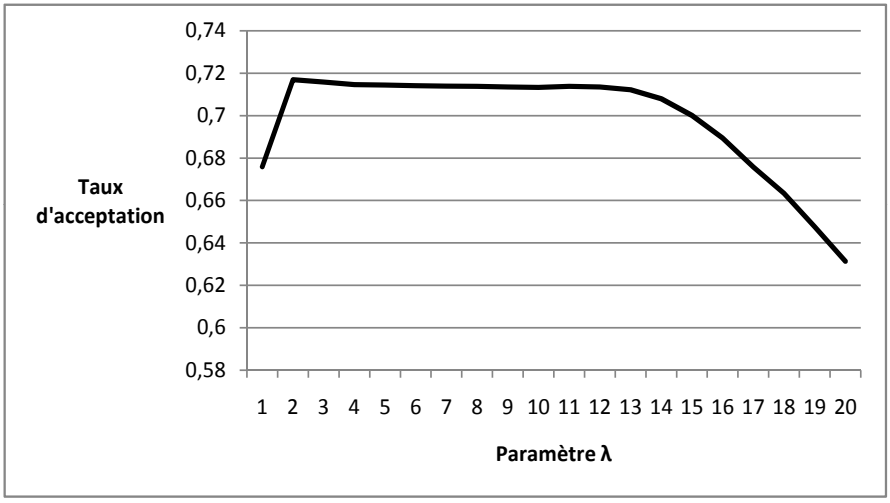

Figure 2.10 – Graphique du taux d'acceptation en fonction de λ (algorithme DA, exemple 2).

et cible en soit supérieur pour tous les éléments de l'espace. En général, $\mathsf{X} = \mathbb{R}^n$ avec mesure sous-jacente de Lebesgue et il n'existe pas d'ensemble borné A tel que $q(A) = 0$ et $\pi(A) > 0$, puisque cela violerait la condition de ϕ-irréductibilté. En plus, la démonstration du théorème 8 dans Mengersen et Tweedie (1996) suppose $q(x) > 0, \pi(x) > 0$ pour tout $x \in \mathsf{X}$. Ainsi, afin de trouver la borne inférieure essentielle du ratio, il est souvent nécessaire d'en comprendre le comportement limite. Ce concept sera exploré tout au long de cette section.

En un premier lieu, en revenant à la section 2.3.1, on se rappelle que l'échantillonneur indépendant (IS) avec distribution instrumentale normale semblait inefficace selon les diagnostics de convergence présentés aux figures 2.6 et 2.8. Selon le graphique 2.5 de la moyenne des degrés de liberté proposés par la méthode DA pour cet exemple, il semble que les queues de la densité cible s'apparentent en moyenne à celles d'une densité Student avec degrés de liberté entre 27 et 33. Si on considère momentanément pour

Figure 2.11 – Graphique de l'écart-type σ de la valeur-p pour $H_0 : \beta_6 = -0,01$ (algorithme DA, exemple 2).

l'exemple 1 que $\pi_1(\mathbf{x}) \approx Student_{27}$, alors

$$\frac{q_{50}(\mathbf{x})}{\pi_1(\mathbf{x})} \approx \frac{\frac{\Gamma\left(\frac{50+3}{2}\right)}{\pi^{3/2}\Gamma\left(\frac{50}{2}\right)}\left(1+\frac{\mathbf{x}'\mathbf{x}}{50+3}\right)^{-\frac{50+3}{2}}(50+3)^{-3/2}}{\frac{\Gamma\left(\frac{27+3}{2}\right)}{\pi^{3/2}\Gamma\left(\frac{27}{2}\right)}\left(1+\frac{\mathbf{x}'\mathbf{x}}{27+3}\right)^{-\frac{27+3}{2}}(27+3)^{-3/2}}.$$

Il est facile de voir que la borne inférieure essentielle de ce ratio est 0. Par exemple, sur une ligne $\mathbf{x} = \mathbf{0} + t\mathbf{1}$, la limite lorsque $t \to \infty$ est donnée par

$$\lim_{t\to\infty}\left((50+3+3t^2)^{\left(\frac{-50-3}{2}\right)}(27+3+3t^2)^{\left(\frac{27+3}{2}\right)}\right) = 0.$$

Ainsi, il semble probable que la raison expliquant la piètre convergence de l'algorithme IS avec densité instrumentale normale (degrés de liberté 50) soit la décroissance rapide de ses queues. Donc, des algorithmes présentés, seuls les méthodes DA et IS avec distribution instrumentale Cauchy pourraient potentiellement converger de façon uni-

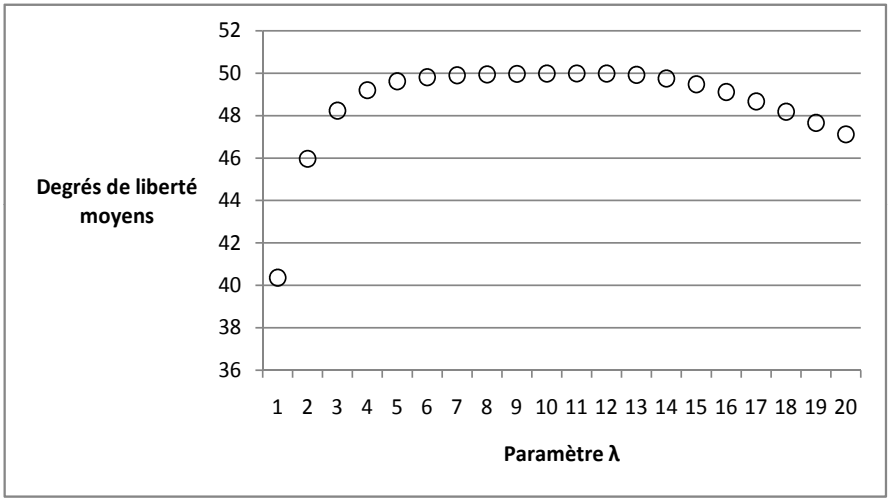

Figure 2.12 – Graphique des degrés de liberté moyens en fonction de λ (algorithme DA, exemple 2). L'écart-type type n'est pas montré puisque plus petit que la largeur des symboles.

formément ergodique et cela semble être corroboré par les diagnostics de convergence présentés à la section 2.3.1.

À ce point, il est temps d'examiner un inconvénient potentiel de la méthode DA. Tel que mentionné à la section 2.1, l'algorithme approxime la densité cible en se basant sur le ratio (2.6) évalué à une distance $s^* = \lambda\sqrt{n}$:

$$\begin{aligned}\frac{\pi\left(\hat{\mathbf{x}}+\hat{H}^{-1/2}\left(\mathbf{u}_{j+1}^{prop}\cdot s^*\right)\right)}{\pi(\hat{\mathbf{x}})} &= \frac{q_{f_{j+1}^{prop}}\left(\mathbf{u}_{j+1}^{prop}\cdot s^*\right)}{q_{f_{j+1}^{prop}}(\mathbf{0})} \\ &= \left(1+\frac{\left(\mathbf{u}_{j+1}^{prop}\cdot s^*\right)'\left(\mathbf{u}_{j+1}^{prop}\cdot s^*\right)}{f_{j+1}^{prop}+n}\right)^{-\frac{f_{j+1}^{prop}+n}{2}}.\end{aligned}$$

Dans certains cas, il peut arriver que la densité cible décroisse rapidement sur une région, s'apparentant à une densité normale, mais que les queues restent épaisses au-delà de cette région. Dans ce contexte, la méthode DA proposera une densité sous-optimale

(quasi-normale) si s^* se trouve dans la région de forte décroissance tandis qu'elle proposera une densité plus optimale (plus épaisse) si s^* se trouve dans la deuxième région. Cette idée est illustrée à l'aide de l'exemple unidimensionnel suivant.

Soit la fonction de densité suivante

$$f(x) = \begin{cases} \frac{3}{16}(2-x^2), & |x| < 1 \\ \frac{3}{16}\frac{1}{x^2}, & |x| \geq 1 \end{cases} . \tag{2.12}$$

Pour cet exemple, la densité est représentée à la figure 2.13, illustrant la symétrie et un mode unique à 0. De plus, cette densité est lisse dans le sens qu'aux points $\{-1,1\}$, la fonction est continue et la valeur des dérivées premières coïncident. La matrice hessienne du négatif du log de la fonction au mode est donnée par

$$\begin{aligned} \frac{d^2}{dx^2}(-\log(f(x)))\Big|_{x=0} &= \frac{d^2}{dx^2}\left(-\log\frac{3}{16} - \log(2-x^2)\right)\Big|_{x=0} \\ &= \frac{d}{dx}\left(\frac{2x}{2-x^2}\right)\Big|_{x=0} \\ &= \frac{4+2x^2}{(2-x^2)^2}\Big|_{x=0} = 1. \end{aligned}$$

Il n'est donc pas nécessaire d'ajuster la densité cible dans cet exemple puisque le mode et la matrice hessienne correspondent déjà à ceux de la densité instrumentale, qui dans ce cas sera de la forme suivante

$$q_f(x) = \frac{\Gamma\left(\frac{f+1}{2}\right)}{\pi^{1/2}\Gamma\left(\frac{f}{2}\right)}\left(1+\frac{x^2}{f+1}\right)^{-\frac{f+1}{2}}(f+1)^{-1/2} .$$

Dans le cas unidimensionnel, la direction proposée par l'algorithme DA est choisie au hasard entre -1 et 1. Le paramètre $s^* = \lambda\sqrt{1} = \lambda$ peut être choisi près du mode, par exemple $\lambda = 1$. D'autres valeurs de λ petites produiront des résultats semblables, mais $\lambda = 1$ ici simplifie l'illustration analytiquement.

Ainsi, les quantités nécessaires de (2.7) pour l'évaluation des degrés de liberté proposés seront simplement

$$r^2 = 2\log\left\{\pi(0)/\pi\left(u_{j+1}^{prop}\cdot s^*\right)\right\} = 2\,\log(2) \tag{2.13}$$

$$Q^2 = \left(u_{j+1}^{prop}\cdot s^*\right)\left(u_{j+1}^{prop}\cdot s^*\right) = 1. \tag{2.14}$$

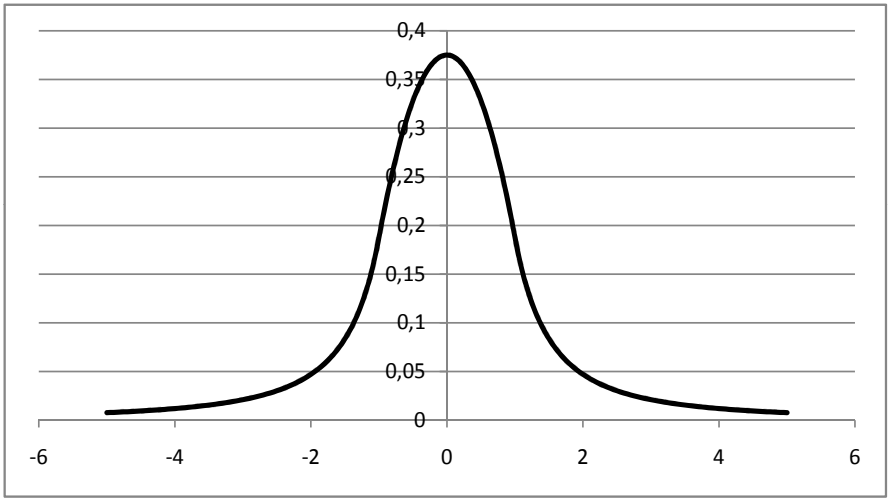

Figure 2.13 – Graphique de la fonction de densité (2.12).

L'équation (2.7) ne sera jamais vraie pour aucune valeur de f puisque la partie gauche est tout au plus 1 (selon une des définitions de la constante $e = \lim_{n\to\infty}(1 + 1/n)^n$). Cependant, la partie gauche croît avec f et par conséquent l'algorithme proposera la plus grande valeur possible $f = 50$.

Toutefois, les queues de la densité cible sont dérivées de densités épaisses de Pareto et donc il semble qu'une densité instrumentale presque normale ne soit pas adéquate. Selon le théorème 8, il serait nécessaire de trouver la borne inférieure essentielle du ratio afin de juger de la convergence de l'algorithme.

Le ratio des deux distributions est donné par :

$$\begin{aligned}\frac{q(x)}{\pi(x)} &= \frac{\frac{\Gamma\left(\frac{f+1}{2}\right)}{\pi^{1/2}\Gamma\left(\frac{f}{2}\right)}\left(1+\frac{x^2}{f+1}\right)^{-\frac{f+1}{2}}(f+1)^{-1/2}}{\frac{3}{16}\frac{1}{x^2}} \\ &= \frac{16\Gamma\left(\frac{f+1}{2}\right)(f+1)^{\frac{f+1}{2}}}{3\pi^{1/2}\Gamma\left(\frac{f}{2}\right)(f+1)^{1/2}}\frac{x^2}{(f+1+x^2)^{\frac{f+1}{2}}},\end{aligned}$$

pour $|x| \geq 1$. En prenant la limite, il en résulte que :

$$\begin{aligned}
\lim_{x\to-\infty}\left(\frac{q(x)}{\pi(x)}\right) = \lim_{x\to\infty}\left(\frac{q(x)}{\pi(x)}\right) &= \lim_{x\to\infty}\left(C\frac{x^2}{(f+1+x^2)^{\frac{f+1}{2}}}\right) \\
&= C\lim_{x\to\infty}\left(\frac{2x}{\left(\frac{f+1}{2}\right)(f+1+x^2)^{\frac{f-1}{2}}2x}\right) \\
&= C\lim_{x\to\infty}\left(\frac{1}{\left(\frac{f+1}{2}\right)(f+1+x^2)^{\frac{f-1}{2}}}\right)
\end{aligned}$$

et finalement

$$\lim_{n\to\infty}\left(\frac{q(x)}{\pi(x)}\right) = \begin{cases} C & \text{si } f=1 \\ 0 & \text{si } f>1 \end{cases} = \begin{cases} \frac{32}{3\pi\sqrt{2}} & \text{si } f=1 \\ 0 & \text{si } f>1 \end{cases}.$$

En examinant le comportement limite, il est clair qu'un algorithme DA proposant une densité instrumentale normale ne converge pas de façon uniformément ni géométriquement ergodique car il n'existe pas de $\beta > 0$ tel que $\frac{q(x)}{\pi(x)} \geq \beta$ pour tout $x \in \mathbb{R}$. La seule densité instrumentale garantissant une ergodicité uniforme est la densité Cauchy. Dans ce cas,

$$\beta = \frac{q(0)}{\pi(0)} = \frac{8}{3\pi\sqrt{2}} \approx 0,6$$

et donc la vitesse de convergence devrait être, selon le théorème 8, de l'ordre de $(1-0,6)^N = 0,4^N$, où N est le nombre d'itérations.

Il peut être intéressant de simuler cet exemple et d'utiliser certaines des mesures de convergence établies précédemment afin d'illustrer l'idée d'un point de vue pratique. La figure 2.14 présente les degrés de liberté moyens proposés, la valeur $s(x=1)$ (dont la valeur exacte est $\frac{3}{16}$) obtenue et son écart-type ainsi que l'autocorrélation en fonction de λ. De ces données, il est clair que la densité instrumentale optimale est utilisée lorsque $\lambda = 42$, ce qui est loin des valeurs $[2;4]$ établies à la section précédente. Donc, un certain degré de prudence est nécessaire lorsque la méthode DA est employée. Il est à mentionner, au cas où l'absence d'espérance mettrait en doute la signification pratique du dernier

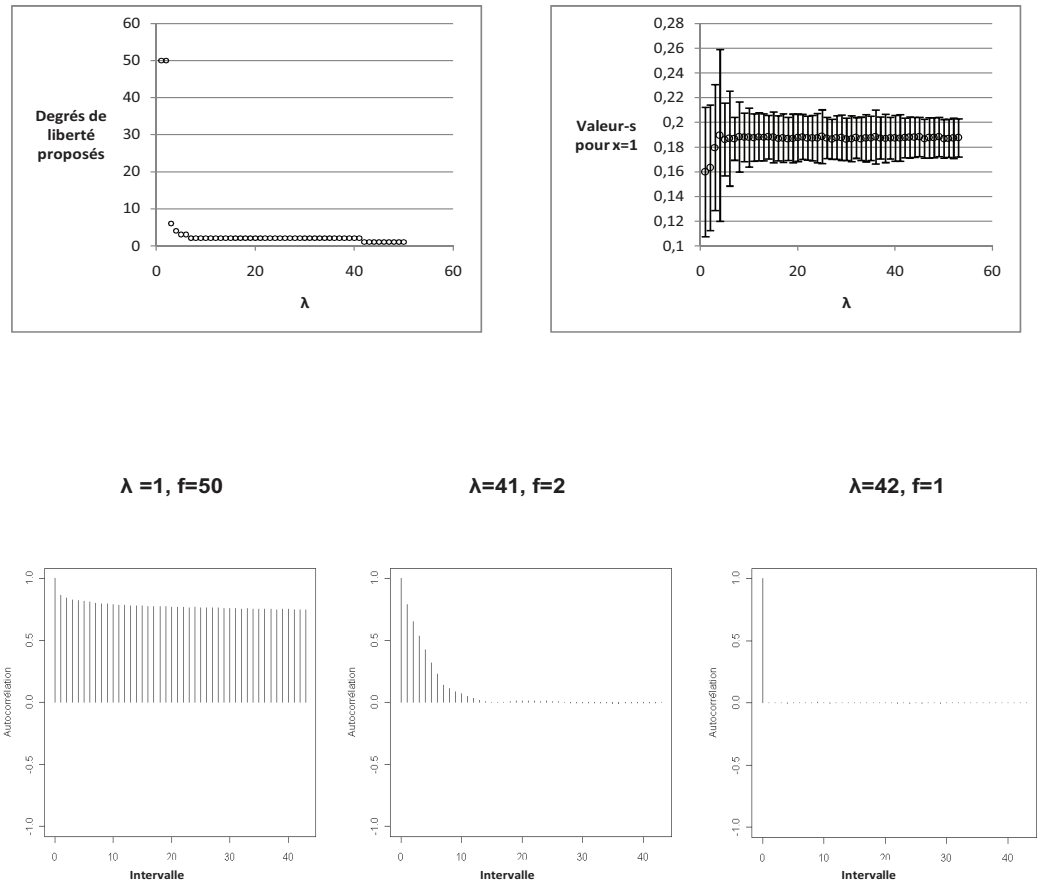

Figure 2.14 – Graphique des degrés de liberté proposés, de la valeur $s(x = 1)$ obtenue (avec un écart-type) et de l'autocorrélation en fonction de λ (algorithme DA, exemple 2.12).

exemple, que les résultats sont sensiblement les mêmes pour une distribution possédant une espérance ainsi qu'une variance et construite de manière identique :

$$f(x) = \begin{cases} \frac{3}{16}(3-2x^2), & |x| < 1 \\ \frac{3}{16}\frac{1}{x^4}, & |x| \geq 1 \end{cases} .$$

À l'inverse, il est anticipé qu'il existe également des exemples où l'algorithme DA serait amené à proposer une densité instrumentale avec des queues épaisses alors que le

choix optimal se trouverait être une densité aux queues minces.

L'utilité de l'exemple précédent était d'illustrer le fait que l'algorithme DA, étant basé seulement sur le ratio des densités par rapport au mode, peut proposer une densité instrumentale sous-optimale. Afin de réduire le risque de telles situations, il est toujours recommandé de simuler l'algorithme avec plusieurs valeurs de λ. Ainsi, il est possible d'étudier le comportement des queues de la densité cible ainsi que l'optimalité de certaines mesures de convergence.

Une autre manière d'éviter une densité instrumentale inefficace est de créer un algorithme avec un mélange entre une densité instrumentale Cauchy et celle proposée par l'algorithme DA. La densité Cauchy possède des queues épaisses, plus épaisses qu'une vaste majorité de densités. En utilisant un tel mélange, il est plus probable d'obtenir une convergence uniformément ergodique de l'algorithme. Par exemple, si on désigne la densité instrumentale globale induite par la méthode DA par $q_f(x)$ et la densité Cauchy par $q_1(x)$ alors un mélange $q(x)$ peut être construit selon

$$q(x) = pq_1(x) + (1-p)q_f(x),$$

où $0 < p < 1$.

Si l'algorithme avec densité instrumentale Cauchy est uniformément ergodique, il en résulte qu'il existe un $\beta > 0$ tel que $q_1(x)/\pi(x) \geq \beta$ pour $x \in \mathbb{R}^n$ et alors

$$\begin{aligned} \frac{q(x)}{\pi(x)} &= \frac{pq_1(x) + (1-p)q_f(x)}{\pi(x)} \\ &\geq p\beta + (1-p)\frac{q_f(x)}{\pi(x)} \\ &\geq p\beta \end{aligned}$$

et la vitesse de convergence de cet algorithme sera $||P^N(x,\cdot) - \pi(\cdot)|| \leq (1-p\beta)^N$.

Il est clair que cette méthode est moins efficace si la méthode DA est déjà uniformément ergodique, mais selon notre expérience l'impact n'est pas énorme pour des valeurs de p petites, typiquement $p \leq 0,05$. En fait, si la méthode DA est déjà uniformément

ergodique, il en resulte qu'il existe $\beta' > 0$ tel que $q_f(x)/\pi(x) \geq \beta'$ et que la vitesse de convergence de cet algorithme est $(1-\beta'+p(\beta'-\beta))^N$.

En contrepartie, si l'approche DA n'est pas uniformément ergodique, l'addition de la densité Cauchy (ou possiblement d'autres densités aux queues épaisses, par exemple les densités de Pareto) pourrait rendre l'algorithme uniformément ergodique. Dans ce cas, la vitesse de convergence sera plus faible que celle d'un algorithme IS avec distribution instrumentale Cauchy, par un facteur de p. Cette dernière méthode est simple, rapidement implémentable et résulte en une approche un peu plus versatile se situant entre celle du DA et du IS. Il est clair qu'elle est utile seulement lorsque l'on est incertain de la convergence de la méthode DA et elle représente en quelque sorte un compromis entre une plus grande probabilité d'uniformité ergodique versus une pire vitesse de convergence.

En conclusion, la méthode DA est une approche intéressante qui vise à combiner la versatilité de l'algorithme RWMH avec la haute performance de l'algorithme IS. Elle se base sur un paramètre λ qui doit être choisi de façon éclairée. En plus, l'approche DA peut dans certains contextes proposer une densité très sous-optimale ce qui peut être évité par une meilleure étude de la densité cible ou l'introduction d'un mélange avec une densité aux queues plus épaisses. En fait, il n'existe pas de méthode infaillible et même un algorithme des plus versatiles comme celui du RWMH peut présenter des problèmes de convergence importants comme dévoilé dans le prochain chapitre.

CHAPITRE 3

COMPARAISON ENTRE L'ALGORITHME DA ET DES MÉTHODES LOCALES

Au dernier chapitre, il a été question de l'algorithme avec ajustement directionnel (DA) qui visait à établir une méthode combinant la versatilité de l'algorithme Metropolis de type marche aléatoire (RWMH) et l'efficacité de l'échantillonneur indépendant (IS). Les méthodes globales, c'est-à-dire les algorithmes où la distribution instrumentale ne dépend pas de l'état présent (comme les méthodes IS et DA), nécessitent une connaissance plus détaillée de la densité cible. Les méthodes locales, c'est-à-dire les algorithmes où la distribution instrumentale est dépendante de l'état présent (comme l'algorithme RWMH), sont plus versatiles et facilement applicables dans plusieurs contextes.

Dans Bédard et Fraser (2008), l'algorithme DA a été comparé et jugé supérieur aux algorithmes IS avec distribution instrumentale $Student_7$ et RWMH avec distribution instrumentale normale aux composantes indépendantes pour les deux exemples spécifiques des sections 2.3.1 et 2.3.2. Cependant, comme mentionné au premier chapitre, il existe plusieurs méthodes MCMC avancées, la plupart dérivées des algorithmes élémentaires présentés à la section 1.2. Par conséquent, il peut être intéressant de comparer l'approche DA avec d'autres algorithmes plus sophistiqués. L'analyse présentée dans ce chapitre s'attardera uniquement sur des méthodes alternatives locales puisque leur versatilité est comparable et/ou supérieure à l'algorithme DA. En effet, il est fort probable qu'il existe des méthodes alternatives globales dont la performance est supériure à celle de la méthode DA, mais elles nécessiteront considérablement plus d'informations au sujet de la densité cible. Donc, l'idée est de déterminer si des méthodes plus facilement applicables que celle du DA peuvent mener à des résultats semblables. Ce chapitre vise à explorer cette comparaison et à cette fin, l'exemple 2 (section 2.3.2) du chapitre dernier sera utilisé comme paradigme.

3.1 Algorithmes RWMH

À titre de rappel, la distribution cible est une distribution de paramètres de régression et est donnée par

$$\pi_2\left(b,a\left|\mathbf{d}^0\right.\right)db\,da = c\prod_{i=1}^{n} h\left(e^a d_i^0 + X_i b\right) e^{a(n-r)} db\,da.$$

La coefficient β_6, correspondant au nombre de réacteurs construits par un ingénieur, est toujours la variable d'intérêt. Pour échantillonner de cette distribution, un choix de méthode locale des plus simples est l'algorithme RWMH avec distribution instrumentale normale aux composantes indépendantes. Cependant, appliqué à cet exemple, cet algorithme présente des problèmes de convergence assez importants. En effet, la figure 3.1 tirée de Bédard et Fraser (2008) illustre ce fait.

Ce graphique présente les valeurs-p obtenues par différentes méthodes en fonction des hypothèses H_0 sur la variable β_6. Les astérisques décrivent la courbe obtenue par une approximation fréquentiste, soit la valeur-p de troisième ordre. Les détails de cette technique ne sont pas abordés dans le présent mémoire, mais le lecteur intéressé est invité à consulter Fraser et Reid (1993). Dans le contexte présent, ces approximations sont de l'ordre de $O(n^{-3/2})$, soit de $O(32^{-3/2})$, et sont incluses à titre de comparaison. La ligne continue correspondant à la méthode DA démontre que cette dernière produit des résultats qui concordent bien avec les valeurs de troisième ordre. D'un autre coté, les courbes pointillées illustrent deux simulations indépendantes de l'approche RWMH qui mènent à des résultats très différents. La différence marquée entre des simulations indépendantes et la discordance entre la méthode RWMH et les deux précédentes indiquent que cet algorithme versatile possède une vitesse de convergence médiocre.

Il est généralement accepté que le choix de la distribution instrumentale pour un algorithme RWMH est crucial. Dans l'article Bédard et Fraser (2008), la matrice de

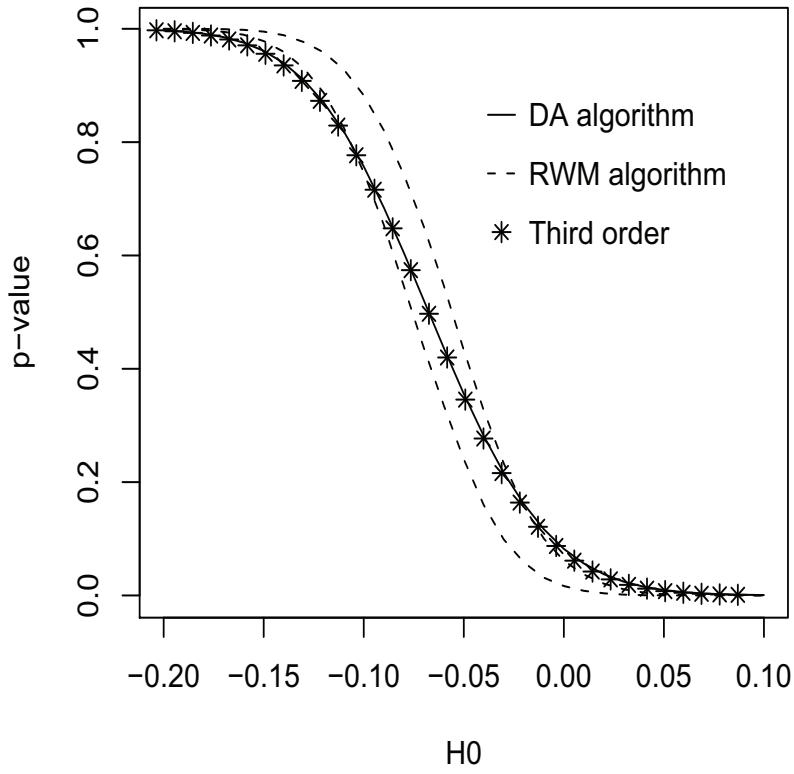

Figure 3.1 – Graphique des valeurs-p obtenues par la méthode DA (ligne pleine), la méthode RWMH (lignes pointillées) ainsi que les approximations de troisième ordre (astérisques) en fonction des hypothèses H_0 sur β_6.

covariance de la distribution instrumentale prenait la forme suivante

$$0,0001 \begin{bmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{bmatrix}.$$

La matrice identité multipliée par un facteur est souvent utilisée comme première approche dans un tel contexte dû à la simplicité de générer des valeurs aléatoires de la distribution instrumentale qui s'en suit. Cependant, cette méthode ne converge pas de façon satisfaisante comme démontré par le graphique précédent. À titre illustratif, la figure 3.2 démontre l'autocorrélation des composantes, obtenue grâce à la fonction CODA du progiciel R. Une période de *burn-in* de 10 000 a été utilisée et 40 000 itérations ont ensuite été analysées.

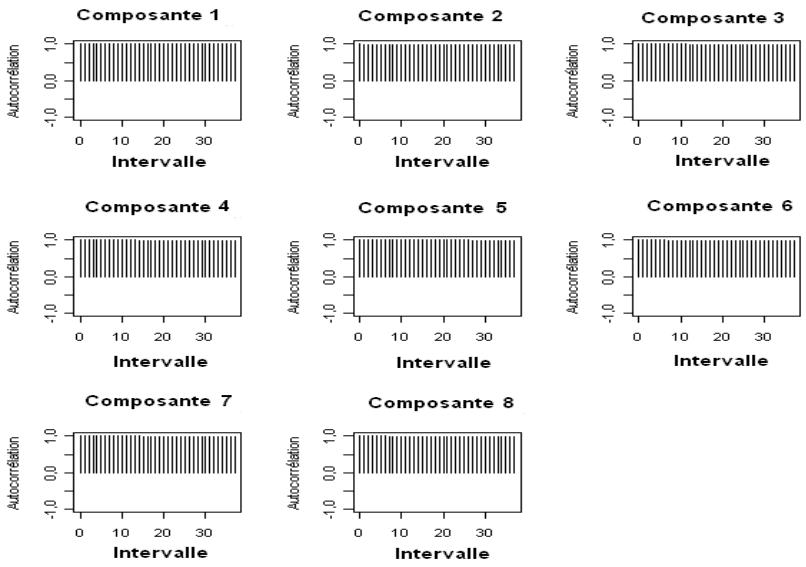

Figure 3.2 – Graphique de l'autocorrélation des valeurs générées par la méthode RWMH.

Donc, l'autocorrélation de chaque composante ne s'estompe pas au fur et à mesure que l'intervalle augmente, ce qui corrobore les conclusions précédentes au sujet de la piètre convergence de cette méthode.

Une manière couramment employée pour tenter d'améliorer la vitesse de convergence est d'optimiser le facteur multipliant la matrice de covariance de la densité instrumentale. Une notion déjà utilisée auparavant, la distance de sauts carrée moyenne (DSCM) peut servir de critère d'optimisation. L'idée est de déterminer la valeur de la variance instrumentale maximisant la DSCM et ensuite d'utiliser cette variance optimale dans les simulations subséquentes. La relation entre la DSCM et la variance instrumentale est illustrée dans la figure 3.3 et la relation similaire du taux d'acceptation est également présentée dans la figure 3.4.

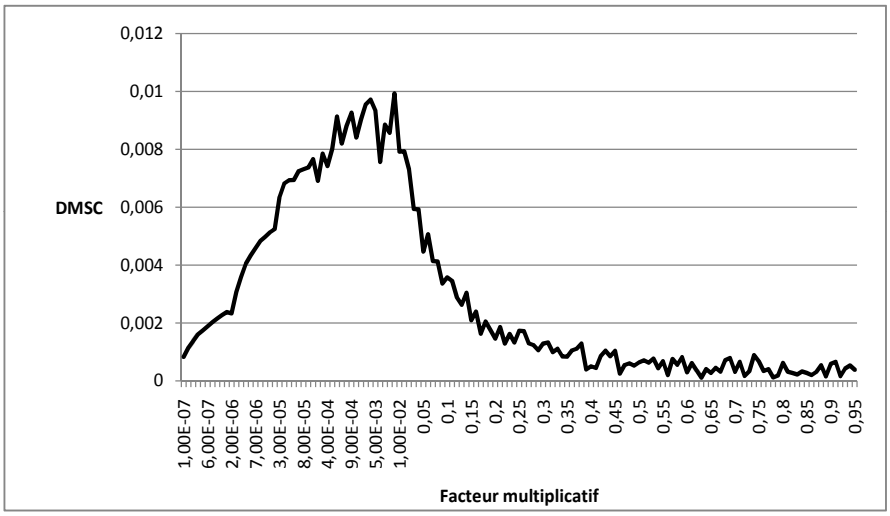

Figure 3.3 – Graphique de la DSCM en fonction du facteur multiplicatif de la matrice de covariance (algorithme RWMH, exemple 2).

La distance de saut carrée moyenne semble être maximale sur l'intervalle $[10^{-4}; 10^{-3}]$ et la valeur initialement choisie dans l'article Bédard et Fraser (2008) appartient à cet intervalle. Le taux d'acceptation correspondant semble raisonnable, soit entre $[0,15; 0,35]$. Il est à noter que le taux d'acceptation tend vers 1 lorsque le facteur multiplicatif tend vers 0. Tel que mentionné au chapitre 1, la méthode RWMH avec

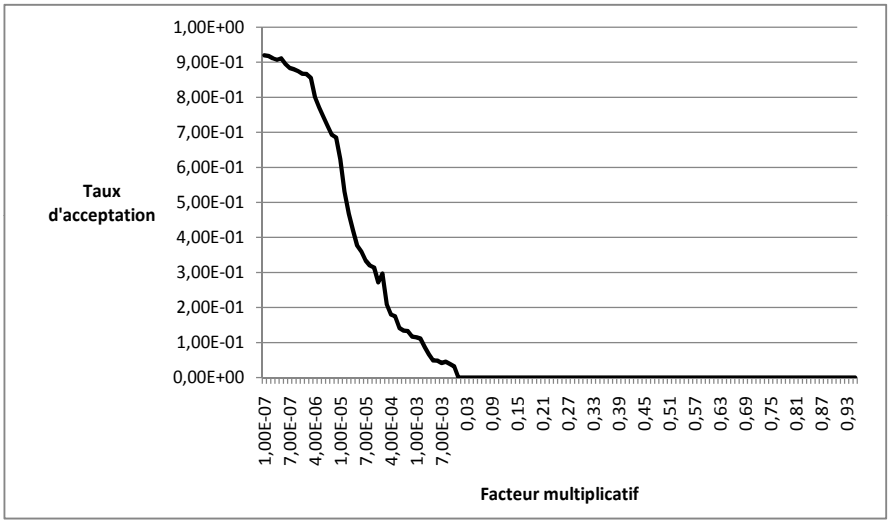

Figure 3.4 – Graphique du taux d'acceptation en fonction du facteur multiplicatif de la matrice de covariance (algorithme RWMH, exemple 2).

distribution instrumentale normale est un cas d'algorithme symétrique et la probabilité d'acceptation se simplifie à

$$\min\left[1, \frac{\pi(y)}{\pi(x)}\right] .$$

Par conséquent, plus la variance de la distribution instrumentale est petite, plus un état proposé y a tendance à être proche d'un état actuel x et donc $\frac{\pi(y)}{\pi(x)} \approx 1$. En d'autres mots, une faible variance de la distribution instrumentale mène à un fort taux d'acceptation de sauts qui sont en moyenne très petits. Par conséquent, il est attendu que le graphique du taux d'acceptation ne possède pas de maximum, puisque cette quantité peut être arbitrairement proche de 1. D'un autre côté, une variance trop forte mène au rejet de la plupart des sauts proposés puisqu'ils se trouveront généralement dans une région de faible densité de la distribution cible. Donc, pour de grandes variances, le taux d'acceptation tend vers 0. Dans le cas d'un algorithme RWMH avec densité instrumentale normale et densité cible n-dimensionnelle avec composantes indépendantes et identiquement dis-

tribuées (i.i.d.), il a été montré que le taux d'acceptation optimal asymptotique (AOAR) est de 0,234 lorsque $n \to \infty$ (Roberts *et al.* (1997)). Bien que prouvé en présumant la condition i.i.d., il apparaît que ce taux optimal est assez robuste et peut s'appliquer dans des contextes déviant de cette supposition (Roberts et Rosenthal (2001)). En général, il est donc courant d'ajuster la variance instrumentale afin d'obtenir un taux d'acceptation d'environ 0,25. Pour l'exemple en cours, l'utilisation d'une variance maximisant la DSCM mène à un taux d'acceptation raisonnablement près de 0,25, mais l'algorithme reste sous-optimal.

Une autre manière d'améliorer la convergence d'un algorithme RWMH est de modifier la forme de la matrice de covariance instrumentale afin de la faire correspondre à la forme de la matrice de covariance cible. La motivation derrière cette technique est à la fois empirique (voir Tierney (1994)) et théorique puisqu'il a été montré que la matrice de covariance instrumentale asymptotiquement optimale est proportionnelle à la matrice de covariance cible au moins dans le cas où les densités cible et instrumentale sont normales (voir Roberts *et al.* (1997)). La seule difficulté réside dans le fait qu'aucune information au sujet de la forme de cette matrice n'est connue à priori.

Une méthode facilement implémentable afin de remédier à ce problème est d'utiliser l'estimateur empirique de la variance pour un échantillon généré par l'algortihme précédent. Bien que l'approche RWMH avec densité instrumentale normale aux composantes indépendentes ne donne pas de résultats convaincants, il est possible que la variance empirique des valeurs générées soit tout de même suffisamment près de la variance cible pour améliorer la convergence d'un algorithme subséquent. Par conséquent, dix essais indépendants ont été simulés avec une période de chauffe de 10 000 et avec un nombre de valeurs générées de 40 000 par essai. Les matrices de variance/covariance obtenues pour chaque simulation ont été comparées entre elles et une forme approximative moyenne

de la variance, qu'on peut appeler A, a été construite :

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0,01 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0,5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0,1 \end{bmatrix}.$$

Cette matrice a ensuite été employée comme matrice de covariance pour la distribution instrumentale normale d'un nouvel algorithme RWMH. Le facteur multiplicatif a été optimisé par un exercice identique à celui appliqué à la première méthode de ce chapitre. Cette nouvelle approche a été simulée avec les mêmes conditions décrites au chapitre 2, c'est-à-dire avec une période de chauffe de 10 000 et 4 000 000 valeurs générées divisées en séries de 950. Le tableau 3.I rapporte aussi les mêmes mesures de convergence qui ont été présentées à la section 2.3.

Le nouvel algorithme augmente la distance de saut carrée moyenne de plus de 50 fois et semble diminuer l'écart-type de la valeur-p pour l'hypothèse $\beta_6 = -0.1$. Pour des raisons de cohérence, cette hypothèse sera maintenue à travers ce chapitre pour toutes les méthodes qui seront présentées.

En somme, cet ajustement de la matrice de covariance a permis d'améliorer quelque peu la convergence de l'algorithme, mesurée selon les critères cités. Cependant, la valeur-p demeure éloignée de l'approximation de troisième ordre, soit 0,75283. À présent, il reste à savoir s'il est possible de faire encore mieux et surtout à quel coût.

3.2 Algorithme Metropolis adaptatif

Une approche qui repose sur les mêmes fondements que l'idée précédente est l'algorithme Metropolis adaptatif (AM) développé par Haario *et al.* (2001). La particularité de cette méthode est le fait que la matrice de covariance est mise à jour à chaque temps (ou saut) de la chaîne de Markov. Pour l'exemple en cours, à chaque temps j, une nouvelle valeur $\mathbf{y}_{j+1}$ est proposée d'une distribution instrumentale $N(\mathbf{x}_j, C_{j+1})$, où C_{j+1} est calculée à partir des valeurs antécédentes $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_j$. Cette nouvelle valeur est acceptée avec probabilité :

$$\alpha(\mathbf{x}_j, \mathbf{y}_{j+1}) = \min\left[1, \frac{\pi(\mathbf{y}_{j+1})}{\pi(\mathbf{x}_j)}\right] .$$

Il est à noter que cet algorithme, même si sa probabilité d'acceptation est évocatrice de certains algorithmes RWMH, n'est pas symétrique ni reversible. Toutefois, il existe des résultats théoriques qui garantissent une convergence vers la distribution cible sous certaines conditions qui seront mentionnées un peu plus loin. En général, il est préférable de générer les premières j_0 valeurs à l'aide d'une distribution instrumentale fixe et les valeurs subséquentes à l'aide d'une distribution mise à jour selon la méthode décrite. Donc, la forme de la matrice de covariance est donnée par

$$C_j \;=\; \begin{cases} C_0, & j \leq j_0 \\ s_n Cov(\mathbf{x}_0, \ldots, \mathbf{x}_{j-1}) + s_n \varepsilon I_n, & j > j_0 \end{cases} .$$

Le terme C_0 représente la connaissance à priori de la matrice de covariance et j_0 indique le temps à partir duquel l'algorithme adaptatif est appliqué. Le paramètre s_n est un facteur multiplicatif qui dépend de la dimension et qui peut être optimisé. Un élément $\varepsilon > 0$ est introduit afin d'assurer la non singularité de la matrice C_j et afin d'assurer la convergence théorique de l'algorithme. En effet, $\varepsilon > 0$ ainsi qu'un support mesurable borné $S \subset \mathbb{R}^n$ pour π sont nécessaires afin de démontrer, pour une fonction f mesurable et bornée, que

$$\lim_{N\to\infty} \frac{1}{N+1}\left(f(\mathbf{x}_0) + f(\mathbf{x}_1) + \ldots + f(\mathbf{x}_N)\right) = \int_S f(\mathbf{x})\pi(d\mathbf{x}) ,$$

presque toujours (voir Haario *et al.* (2001) pour de plus amples détails).

D'un côté pratique, ces deux conditions ne semblent pas être toujours cruciales puisque l'algorithme converge de façon similaire sans ces restrictions dans tous les exemples testés par les auteurs. Toutefois, afin de concilier théorie et pratique, il est toujours possible de choisir S arbitrairement grand et ε arbitrairement petit par rapport à S.

Finalement, il vaut la peine de mentionner que le calcul de la covariance ne requiert pas autant de ressources que l'on pourrait l'imaginer à priori. La matrice de covariance et la moyenne satisfont les formules récursives suivantes

$$C_{j+1} = \frac{j-1}{j} C_j + \frac{s_n}{j} \left(j\overline{\mathbf{x}}_{j-1}\overline{\mathbf{x}}'_{j-1} - (j+1)\overline{\mathbf{x}}_j\overline{\mathbf{x}}'_j + \mathbf{x}_j\mathbf{x}'_j + \varepsilon I_n \right)$$

et

$$\overline{\mathbf{x}}_{j+1} = \frac{1}{j+1} \left(j\overline{\mathbf{x}}_j + \mathbf{x}_{j+1} \right) .$$

L'emploi de ces équations élimine la nécessité de stocker les valeurs générées et permet de diminuer considérablement le coût de simulation. Les résultats de cette méthode sont rapportés au tableau 3.I ; ceux-ci sont beaucoup plus convaincants que ceux obtenus à la section 3.1, la valeur-p concordant avec l'approximation de troisième ordre. De plus, la DSCM est augmentée de près de 100 000 fois et l'écart-type de la valeur-p est réduit de moitié. Le paramètre s_n a été choisi tel qu'il serait asymptotiquement optimal dans un contexte où la densité cible est normale, c'est-à-dire égal à $2,38^2/8$, tel que justifié dans Tierney (1994) et Roberts *et al.* (1997).

Aussi, on remarque que la matrice de covariance (3.1) calculée par l'algorithme est

très différente de la matrice A utilisée précédemment :

$$\begin{bmatrix} 675,85 & -8,74 & -12,26 & -5,16 & -2,21 & 5,75 & -15,11 & -0,53 \\ -8,74 & 0,11 & 0,10 & 0,07 & 0,03 & -0,07 & 0,20 & 0,01 \\ -12,26 & 0,10 & 0,74 & 0,05 & -0,01 & -0,11 & 0,17 & 0,02 \\ -5,16 & 0,07 & 0,05 & 0,30 & -0,02 & -0,08 & 0,20 & 0,00 \\ -2,21 & 0,03 & -0,01 & -0,02 & 0,17 & -0,01 & 0,08 & 0,00 \\ 5,75 & -0,07 & -0,11 & -0,08 & -0,01 & 0,10 & -0,16 & -0,01 \\ -15,11 & 0,20 & 0,17 & 0,20 & 0,08 & -0,16 & 0,56 & 0,01 \\ -0,53 & 0,01 & 0,02 & 0,00 & 0,00 & -0,01 & 0,01 & 0,02 \end{bmatrix} . \quad (3.1)$$

On constate, entre autres, la variance très grande de la première composante ainsi qu'une covariance considérable entre la première et la plupart des autres composantes. En rétrospective, il n'est pas surprenant que les deux premiers algorithmes démontrent une convergence lente puisque la variance de la distribution instrumentale est vastement différente de celle de la distribution cible.

3.3 Algorithmes Metropolis avec essais multiples

Bien que la méthode AM produise des résultats à priori satisfaisants, il peut être intéressant de poursuivre l'analyse de cet exemple et d'examiner la performance de versions spécifiques d'autres algorithmes locaux. Les méthodes suivantes seront computationnellement plus intenses et demanderont davantage de ressources que celle du RWMH avec densité instrumentale aux composantes indépendantes. Cependant, il serait intéressant de voir si un meilleur compromis entre efficacité et effort computationnel peut être atteint. En plus, il serait intéressant de déterminer si d'autres algorithmes peuvent agir à titre de remplacement aux méthodes adaptatives et si ces dernières sont véritablement nécessaires dans ce contexte.

3.3.1 Algorithme MTM

Un algorithme de type marche aléatoire à essais multiples (*Multiple-Try Metropolis Algorithm (MTM)*) sera la première approche utilisée. Cet algorithme a été développé par Liu *et al.* (2000) et continue d'être un sujet d'intérêt en recherche MCMC. L'idée fondamentale de cette méthode est d'élargir l'ensemble des valeurs proposées, c'est-à-dire proposer plus d'un candidat par itération, afin d'améliorer l'exploration de l'espace.

Étant donné la valeur présente $\mathbf{x}$, l'approche MTM génère k valeurs, $\{\mathbf{y}_1,\ldots,\mathbf{y}_k\}$, de la distribution instrumentale $T(\mathbf{x},\cdot)$. Ensuite, un candidat $\mathbf{Y}=\mathbf{y}$ est choisi parmi les k valeurs proposées avec probabilité proportionnelle à un poids $w(\mathbf{y}_l,\mathbf{x})$. Ce poids est défini en général par

$$w(\mathbf{x},\mathbf{y}) = \pi(\mathbf{x})T(\mathbf{x},\mathbf{y})\lambda(\mathbf{x},\mathbf{y}) \ .$$

La densité instrumentale $T(\mathbf{x},\mathbf{y})$ ne doit pas nécessairement être symétrique, mais la condition $T(\mathbf{x},\mathbf{y})>0 \Leftrightarrow T(\mathbf{y},\mathbf{x})>0$ doit être remplie. La fonction $\lambda(\mathbf{x},\mathbf{y})$ est une fonction symétrique et nonnégative choisie par l'utilisateur avec condition $T(\mathbf{x},\mathbf{y})>0 \Rightarrow \lambda(\mathbf{x},\mathbf{y})>0$.

Subséquemment, $k-1$ valeurs $\{\mathbf{x}_1^*,\ldots,\mathbf{x}_{k-1}^*\}$ sont générées de la distribution $T(\mathbf{y},\cdot)$ et $\mathbf{x}_k^*=\mathbf{x}$. Finalement, $\mathbf{y}$ est accepté avec probabilité

$$\alpha(\mathbf{x}_1^*,\ldots,\mathbf{x}_k^*,\mathbf{y}_1,\ldots,\mathbf{y}_k) = \min\left\{1, \frac{w(\mathbf{y}_1,\mathbf{x})+\ldots+w(\mathbf{y}_k,\mathbf{x})}{w(\mathbf{x}_1^*,\mathbf{y})+\ldots w(\mathbf{x}_k^*,\mathbf{y})}\right\} \ .$$

Il est démontré dans Liu *et al.* (2000) que cette méthode avec les conditions citées plus haut produit une chaîne de Markov réversible par rapport à la distribution cible. Il existe plusieurs options pour le choix de $\lambda(\mathbf{x},\mathbf{y})$ et cela donne naissance à plusieurs types d'algorithmes MTM.

Une méthode MTM populaire se base sur le fait que si $T(\mathbf{x},\mathbf{y})$ est symétrique, alors $\lambda(\mathbf{x},\mathbf{y})=1/T(\mathbf{x},\mathbf{y})$ l'est aussi et donc que $w(\mathbf{y}_l,\mathbf{x})=\pi(\mathbf{y}_l)$ est un choix valide. Cela donne lieu à la probabilité d'acceptation suivante

$$\alpha(\mathbf{x}_1^*,\ldots,\mathbf{x}_k^*,\mathbf{y}_1,\ldots,\mathbf{y}_k) = \min\left\{1, \frac{\pi(\mathbf{y}_1)+\ldots+\pi(\mathbf{y}_k)}{\pi(\mathbf{x}_1^*)+\ldots+\pi(\mathbf{x}_k^*)}\right\}$$

Cette dernière approche a été appliquée à l'exemple en cours avec $k = 2$ valeurs proposées. Le tableau 3.I résume les résultats obtenus. Il apparaît que cet algorithme n'apporte qu'une très légère amélioration par rapport à l'algorithme initial RWMH $(\mathbf{x}_j,\ 0,0001I_8)$. La DSCM est augmentée de trois fois et l'écart-type de la valeur-p est diminué de près de 15%. Cependant, la valeur-p globale reste éloignée de l'approximation de troisième ordre.

Il a été montré que le taux d'acceptation asymptotiquement optimal pour cet algorithme se situe à 0,32 sous des conditions similaires à celles du taux optimal asymptotique de la méthode RWMH (voir Bédard *et al.* (2010b)). Dans le contexte présent, la variance instrumentale a donc été ajustée afin de produire un taux d'acceptation global se situant près de cette valeur.

3.3.2 Algorithme MTM hit-and-run

Il existe plusieurs autres types d'algorithmes MTM, dont une variante qui permet de proposer plusieurs valeurs dépendantes à l'intérieur d'une même itération. En effet, il peut arriver que la distribution instrumentale soit dépendante d'une variable auxiliaire, disons $\mathbf{e}$, telle que $T(\mathbf{x},\mathbf{y}) = \int T_{\mathbf{e}}(\mathbf{x},\mathbf{y}) f_{\mathbf{x}}(\mathbf{e}) d\mathbf{e}$ où $f_{\mathbf{x}}(\mathbf{e})$ est la densité de cette variable. Dans un tel cas, il est possible de générer un ensemble de k valeurs $\{\mathbf{y}_1,\ldots,\mathbf{y}_k\}$ de la distribution instrumentale $T_{\mathbf{e}}(\mathbf{x},\cdot)$. Le candidat $\mathbf{Y} = \mathbf{y}$ est choisi avec probabilité proportionnelle à un poids $w_{\mathbf{e}}(\mathbf{y_l},\mathbf{x})$. Ce poids est défini en général tel que

$$w_{\mathbf{e}}(\mathbf{x},\mathbf{y}) = \pi(\mathbf{x}) f_{\mathbf{x}}(\mathbf{e}) T_{\mathbf{e}}(\mathbf{x},\mathbf{y}) \lambda_{\mathbf{e}}(\mathbf{x},\mathbf{y}) \ ,$$

où $\lambda_{\mathbf{e}}(\mathbf{x},\mathbf{y})$ est une fonction positive et symétrique.

Ensuite, $k-1$ valeurs $\{\mathbf{x}_1^*,\ldots,\mathbf{x}_{k-1}^*\}$ sont générées à partir de la distribution $T_{\mathbf{e}}(\mathbf{y},\cdot)$. Le candidat $\mathbf{y}$ est accepté avec probabilité

$$\alpha_e(\mathbf{x},\mathbf{x}_1^*,\ldots,\mathbf{x}_{k-1}^*,\mathbf{y}_1,\ldots,\mathbf{y}_k) = \min\left\{1, \frac{w_{\mathbf{e}}(\mathbf{y},\mathbf{x}) + \sum_{l=1}^{k-1} w_{\mathbf{e}}(\mathbf{y}_l,\mathbf{x})}{w_{\mathbf{e}}(\mathbf{x},\mathbf{y}) + \sum_{l=1}^{k-1} w_{\mathbf{e}}(\mathbf{x}_l^*,\mathbf{y})}\right\} \ .$$

Cette forme du poids donné aux valeurs générées permet d'assurer la réversibilité de la méthode (voir Liu *et al.* (2000) pour de plus amples détails).

Un algorithme appelé MTM hit-and-run (MTMHR) repose sur l'idée précédente et tente d'améliorer l'exploration de l'espace de manière directionnelle. La première étape est la génération d'une direction, c'est-à-dire un vecteur unité $\mathbf{e}$ à partir d'une distribution $f(\mathbf{e})$. Ensuite, un choix populaire est de générer $(r_1,\ldots,r_k) \sim N(0,\sigma^2)$ et de poser $\mathbf{y}_l = \mathbf{x} + r_l\mathbf{e}$. Un candidat $\mathbf{Y} = \mathbf{y}$ est choisi et enfin, $k-1$ valeurs $\{\mathbf{x}_1^*,\ldots,\mathbf{x}_{k-1}^*\}$ sont générées à partir de $T_{\mathbf{e}}(\mathbf{y},\cdot)$ et l'algorithme est exécuté selon les étapes précédentes.

Il est à noter que si $f_{\mathbf{x}}(\mathbf{e}) = f_{\mathbf{y}}(\mathbf{e}) = f(\mathbf{e})$, c'est-à-dire que la distribution du vecteur est indépendante de $\mathbf{x}$ ou $\mathbf{y}$, alors $f(\mathbf{e})$ étant constante peut être exclue de la fonction de poids. Aussi, si $T_{\mathbf{e}}(\mathbf{x},\cdot)$ est symétrique alors $\lambda_{\mathbf{e}}(\mathbf{x},\mathbf{y})$ peut être choisie $1/T_{\mathbf{e}}(\mathbf{x},\cdot)$ et la forme de la probabilité d'acceptation devient alors :

$$\alpha(\mathbf{x},\mathbf{x}_1^*,\ldots,\mathbf{x}_{k-1}^*,\mathbf{y}_1,\ldots,\mathbf{y}_k) = \min\left\{1, \frac{\pi(\mathbf{y}) + \sum_{l=1}^{k-1}\pi(\mathbf{y}_l)}{\pi(\mathbf{x}) + \sum_{l=1}^{k-1}\pi(\mathbf{x}_l^*)}\right\}.$$

Cette derniere version de l'algorithme MTMHR avec $k = 2$ a été appliquée à l'exemple en cours et la variance a été ajustée afin de produire un taux d'acceptation près de la valeur asymptotiquement optimale de 0,46 présentée dans Bédard *et al.* (2010b). Les résultats sont indiqués dans le tableau 3.I. Il apparaît que cet algorithme est presqu'équivalent à l'algorithme initial RWMH$(\mathbf{x}_j,\ 0,0001I_8)$. La DSCM est augmentée de deux fois, mais l'écart-type de la valeur-p obtenue est identique sinon légèrement supérieur. La valeur-p globale reste encore éloignée de l'approximation de troisième ordre.

3.3.3 Algorithme Metropolis avec rejet différé

Enfin, une méthode alternative aux approches de type MTM est l'algorithme avec rejet différé (*Delayed Rejection Metropolis-Hastings* (DR)). Ce type d'algorithme a été développé par Mira (2001) et repose sur un concept similaire à la méthode MTM. En effet, l'objectif est toujours une meilleure exploration de l'espace par une génération de valeurs multiples à l'intérieur d'une même itération. Cependant, les candidats sont maintenant proposés de manière consécutive plutôt que simultanément. L'idée se base sur le fait qu'un rejet suggère que le candidat proposé se trouve dans une région où $\pi(\mathbf{x})$

est faible. En utilisant cette information, une deuxième valeur est proposée, en général d'une distribution moins étendue, avant d'incrémenter le temps. Dans ce type d'algorithme (tout comme les méthodes MTM), il est clair que le taux d'acceptation doit toujours être analysé avec prudence puisqu'il sera artificiellement élevé dû aux nombreuses valeurs proposées avant un rejet.

Si l'on suppose qu'au temps j, $\mathbf{X}_j = \mathbf{x}$, la première étape de la méthode DR est de proposer une nouvelle valeur $\mathbf{y}_1$ à partir d'une distribution instrumentale $q_1(\mathbf{x}, d\mathbf{y}_1)$ et de l'accepter avec la probabilité habituelle

$$\alpha_1(\mathbf{x},\mathbf{y}_1) = \min\left(1, \frac{\pi(\mathbf{y}_1)q_1(\mathbf{y}_1,\mathbf{x})}{\pi(\mathbf{x})q_1(\mathbf{x},\mathbf{y}_1)}\right) = \min\left(1, \frac{N_1}{D_1}\right) .$$

Si la valeur $\mathbf{y}_1$ est acceptée, le temps est incrémenté et $\mathbf{X}_{j+1} = \mathbf{y}_1$. Si elle est rejetée, une nouvelle valeur $\mathbf{y}_2$ est proposée d'une distribution $q_2(\mathbf{x}, \mathbf{y}_1, d\mathbf{y}_2)$ et elle est acceptée avec probabilité

$$\begin{aligned}\alpha_2(\mathbf{x},\mathbf{y}_1,\mathbf{y}_2) &= \min\left(1, \frac{\pi(\mathbf{y}_2)q_1(\mathbf{y}_2,\mathbf{y}_1)q_2(\mathbf{y}_2,\mathbf{y}_1,\mathbf{x})[1-\alpha_1(\mathbf{y}_2,\mathbf{y}_1)]}{\pi(\mathbf{x})q_1(\mathbf{x},\mathbf{y}_1)q_2(\mathbf{x},\mathbf{y}_1,\mathbf{y}_2)[1-\alpha_1(\mathbf{x},\mathbf{y}_1)]}\right) \\ &= \min\left(1, \frac{N_2}{D_2}\right) .\end{aligned}$$

Cette forme de la probabilité d'acceptation est une manière possible de préserver la réversibilité de la chaîne (voir Tierney et Mira (1999) pour plus de détails). On remarque que si $\mathbf{y}_1$ est rejeté cela implique que $N_1 < D_1$ et donc $\alpha_1(\mathbf{x},\mathbf{y}_1) = N_1/D_1$; la deuxième probabilité d'acceptation se simplifie donc à

$$\begin{aligned}\alpha_2(\mathbf{x},\mathbf{y}_1,\mathbf{y}_2) &= \min\left(1, \frac{N_2}{q_2(\mathbf{x},\mathbf{y}_1,\mathbf{y}_2)[\pi(\mathbf{x})q_1(\mathbf{x},\mathbf{y}_1) - \pi(\mathbf{y}_1)q_1(\mathbf{y}_1,\mathbf{x})]}\right) \\ &= \min\left(1, \frac{N_2}{q_2(\mathbf{x},\mathbf{y}_1,\mathbf{y}_2)[D_1 - N_1]}\right) .\end{aligned}$$

De façon générale, la i-ème valeur $\mathbf{y}_i$ est proposée de la distribution $q_i(\mathbf{x}, \mathbf{y}_1, \ldots, d\mathbf{y}_i)$ si la valeur précédente $\mathbf{y}_{i-1}$ est rejetée. La probabilité d'acceptation pour la nouvelle

valeur devient

$$\begin{aligned}\alpha_i(\mathbf{x},\mathbf{y}_1,\ldots,\mathbf{y}_i) &= \min\left(1,\left\{\frac{\pi(\mathbf{y}_i)q_1(\mathbf{y}_i,\mathbf{y}_{i-1})q_2(\mathbf{y}_i,\mathbf{y}_{i-1},\mathbf{y}_{i-2})\cdots q_i(\mathbf{y}_i,\mathbf{y}_{i-1},\ldots,\mathbf{x})}{\pi(\mathbf{x})q_1(\mathbf{x},\mathbf{y}_1)q_2(\mathbf{x},\mathbf{y}_1,\mathbf{y}_2)\cdots q_i(\mathbf{x},\mathbf{y}_1,\ldots,\mathbf{y}_i)}\right.\right.\\ &\qquad\left.\left.\frac{[1-\alpha_1(\mathbf{y}_i,\mathbf{y}_{i-1})][1-\alpha_2(\mathbf{y}_i,\mathbf{y}_{i-1},\mathbf{y}_{i-2})]\cdots[1-\alpha_{i-1}(\mathbf{y}_i,\ldots,\mathbf{y}_1)]}{[1-\alpha_1(\mathbf{x},\mathbf{y}_1)][1-\alpha_2(\mathbf{x},\mathbf{y}_1,\mathbf{y}_2)]\cdots[1-\alpha_{i-1}(\mathbf{x},\mathbf{y}_1,\cdots,\mathbf{y}_{i-1})]}\right\}\right)\\ &= \min\left(1,\frac{N_i}{D_i}\right)\end{aligned}$$

et

$$D_i = q_i(\mathbf{x},\mathbf{y}_1,\ldots,\mathbf{y}_i)(D_{i-1}-N_{i-1})\ .$$

La grande utilité de ces formules récursives est naturellement de réduire le coût computationnel de la méthode.

Une manière simple d'appliquer l'algorithme DR est de proposer des valeurs indépendantes des précédentes à chaque étape, c'est-à-dire $q_i(\mathbf{x},\mathbf{y}_1,\ldots,d\mathbf{y}_i) = q_i(\mathbf{x},d\mathbf{y}_i)$. Dans un tel cas, la variance de chaque distribution instrumentale pourrait être sélectionnée telle que $\sigma_1 > \sigma_2 > \sigma_3 > \ldots > \sigma_k$, où k est le nombre d'essais permis. Bien que relativement facile à implémenter, cette méthode a été montrée équivalente à un algorithme RWMH lorsque la dimension de la distribution cible est grande (voir Bédard *et al.* (2010a)). De façon intuitive, il apparaît qu'un tel algorithme n'utilise pas toute l'information contenue dans le rejet des variables passées. Une variante de cet algorithme permet justement de corriger cet aspect par une approche directionnelle. La méthode DR avec distribution instrumentale antithétique (delayed rejection with antithetic proposal (DR anti)) propose un maximum de deux candidats dans des directions opposées. En d'autres mots, le rejet du premier candidat est considéré comme une indication qu'il se situait dans une région de faible densité de $\pi(\mathbf{x})$ et qu'une région de forte densité se trouve fort probablement dans la direction opposée. Bien que plusieurs choix de distributions instrumentales sont possibles, la présente analyse considérera la distribution instrumentale normale seulement.

Comme d'habitude, on suppose que $\mathbf{X} = \mathbf{x}$ est la valeur présente. La première étape propose une direction $\mathbf{Z}_{j+1} = \mathbf{z}$ provenant d'une distribution $N(\mathbf{0}, I_n)$. Par la suite, un pre-

mier candidat $\mathbf{Y}_{j+1}^{(1)} = \mathbf{y}^{(1)}$ est obtenu en posant $\mathbf{y}^{(1)} = \mathbf{x}_j + \sigma_1 \mathbf{z}$, où $\sigma_1 > 0$. La probabilité d'acceptation associée est identique à celle d'un algorithme symétrique, c'est-à-dire

$$\alpha_1\left(\mathbf{x}, \mathbf{y}^{(1)}\right) = \min\left(1, \frac{\pi\left(\mathbf{y}^{(1)}\right)}{\pi(\mathbf{x})}\right) .$$

En cas de rejet seulement, une deuxième valeur $\mathbf{Y}_{j+1}^{(2)} = \mathbf{y}^{(2)}$ est proposée, $\mathbf{y}^{(2)} = \mathbf{x}_j - \sigma_2 \mathbf{z}$ où $\sigma_2 > 0$. Donc, le deuxième candidat se trouve effectivement dans la direction opposée au premier. Ensuite, afin de préserver la réversibilité de la chaîne, cette deuxième valeur est acceptée avec probabilité

$$\alpha_2\left(\mathbf{x}, \mathbf{y}^{(2)}\right) = \min\left(1, \frac{\pi(\mathbf{y}^{(2)})\left[1 - \frac{\pi((1+\sigma_1/\sigma_2)\mathbf{y}^{(2)} - \sigma_1\mathbf{x}/\sigma_2)}{\pi(\mathbf{y}^{(2)})}\right]_+}{\pi(\mathbf{x})\left[1 - \frac{\pi((1+\sigma_1/\sigma_2)\mathbf{x} - \sigma_1\mathbf{y}^{(2)}/\sigma_2)}{\pi(\mathbf{x})}\right]_+}\right) ,$$

où $[x]_+$ signifie $\max(0, x)$.

Les résultats de cette méthode appliquée à l'exemple en cours sont illustrés au tableau 3.I. La variance a été ajustée similairement aux méthodes précédentes afin de produire un taux d'acceptation global près de la valeur 0,39, démontrée asymptotiquement optimale dans Bédard *et al.* (2010a). Il apparaît que cet algorithme est aussi presqu'équivalent à l'algorithme initial RWMH($\mathbf{x}_j$, $0,0001 I_8$). La DSCM est augmentée de deux fois, mais l'écart-type de la valeur-p obtenue en est légèrement supérieur. D'un autre côté, la valeur-p globale se rapproche un peu plus de l'approximation de troisième ordre.

À première vue, il semble que les algorithmes MTM, MTMHR et DR anti n'offrent pas une véritable amélioration par rapport à l'algorithme RWMH($\mathbf{x}_j$, $0,0001 I_8$) initial. Il est aussi nécessaire de se rappeler que les méthodes MTM et MTMHR nécessitent à peu près deux fois plus de ressources puisqu'il faut générer deux valeurs par itération et évaluer la densité cible à chacune d'entre elles. Si on suit la convention de diviser les mesures de convergence par le nombre d'essais, elles représentent en fait des méthodes dont la performance est inférieure à celle de la méthode RWMH pour cet exemple. D'un autre côté, la méthode DR anti ne requiert qu'une simulation par itération et un facteur

Tableau 3.I – Résumé des résultats de méthodes locales (exemple 2) (Période de chauffe 10 000, Itérations 4 000 000).

Algorithme	valeur-p pour $\beta_6 = -0,1$
RWMH - Instrumentale $N(\mathbf{x}_j,\ 0,0001I_8)$	0,89338
{écart-type de la valeur-p}	$(0,12942)$
{DSCM}	0,000247
{Taux d'acceptation}	0,345
RWMH - Instrumentale $N(\mathbf{x}_j,\ 0,005A)$	0,84919
{écart-type de la valeur-p}	$(0,09242)$
{DSCM}	0,011210
{Taux d'acceptation}	0,356
AM - Instrumentale $N(\mathbf{x}_j, C_j)(\varepsilon = 0,001, j_0 = 40\ 000, s_n = 0,7)$	0,75339
{écart-type de la valeur-p}	$(0,063330)$
{DSCM}	29,97311
{Taux d'acceptation}	0,061
MTM 2 essais - Instrumentale $N(\mathbf{x}_j,\ 0,0003I_8)$	0,90904
{écart-type de la valeur-p}	$(0,10568)$
{DSCM}	0,000737
{Taux d'acceptation}	0,348
MTMHR 2 essais - $(r_1, r_2 \sim N(0,\ 0,036)$ et $f_e(e)$ uniforme)	0,85552
{écart-type de la valeur-p}	$(0,13330)$
{DSCM}	0,000571
{Taux d'acceptation}	0,478
DR anti - $(\mathbf{Z}_{j+1} \sim N(0, I_8), \sigma_1 = \sigma_2 = 0,013)$	0,79176
{écart-type de la valeur-p}	$(0,16370)$
{DSCM}	0,000488
{Taux d'acceptation}	0,403
DA	0,75683
{écart-type de la valeur-p}	$(0,01081)$
{DSCM}	1322,69
{Taux d'acceptation}	0,716
Approximation de trosième ordre	0,75283

de 2 semble assez exagéré dans cette situation, quoiqu'une probabilité d'acceptation plus complexe doit être évaluée en cas de rejet. De toute manière, ces trois algorithmes locaux semblent être moins performants que la méthode AM et encore moins performants que la méthode DA.

3.4 Conclusion et analyse

Pour des raisons illustratives, il peut être intéressant de revenir à l'algorithme global DA dont la performance semblait très satisfaisante. Il est possible d'obtenir une matrice de covariance empirique à partir de valeurs générées par cette méthode. En effet, en utilisant 40 000 itérations, un calcul empirique mène à une matrice B :

$$B = \begin{bmatrix} 978,32 & -12,66 & -17,65 & -7,33 & -3,23 & 8,24 & -21,86 & -0,69 \\ -12,66 & 0,17 & 0,15 & 0,10 & 0,04 & -0,10 & 0,29 & 0,01 \\ -17,65 & 0,15 & 1,04 & 0,07 & 0 & -0,16 & 0,25 & 0,03 \\ -7,33 & 0,10 & 0,07 & 0,42 & -0,04 & -0,11 & 0,28 & 0 \\ -3,23 & 0,04 & 0 & -0,04 & 0,24 & -0,01 & 0,11 & 0 \\ 8,24 & -0,10 & -0,16 & -0,11 & -0,01 & 0,14 & -0,23 & -0,01 \\ -21,8 & 0,29 & 0,25 & 0,28 & 0,11 & -0,23 & 0,81 & 0,02 \\ -0,69 & 0,01 & 0,03 & 0 & 0 & -0,01 & 0,02 & 0,03 \end{bmatrix} \quad (3.2)$$

On remarque tout d'abord que la forme de la matrice de covariance empirique générée par l'algorithme AM, la seule méthode utilisée qui possédait une convergence raisonnable, est comparable à B. Par conséquent, il est fort probable qu'aucune méthode locale ne converge de manière satisfaisante sans une étape adaptative. En d'autres mots, afin d'utiliser une méthode locale quelconque, il apparaît nécessaire pour cet exemple d'avoir au préalable une estimation fiable de la matrice de covariance. En retrospective, la piètre convergence des algorithmes locaux utilisés à la section 3.3 n'est donc pas surprenante, car pour cet exemple, la matrice de covariance de la distribution cible B est vastement différente de I_8.

Le tableau 3.II présente un résumé de certaines des méthodes locales présentées plus tôt, cette fois avec matrice de covariance instrumentale B et ajustées afin d'approximativement obtenir un taux d'acceptation asymptotiquement optimal. On note, pour toutes les méthodes, une amélioration remarquable des mesures de convergence utilisées. La valeur-p globale est très près de l'approximation de troisième ordre, la DSCM est augmentée de près de 700 000 fois et l'écart-type de la valeur-p est diminué de près de 6 fois. L'algorithme MTM, après l'ajustement de ses mesures d'efficacité par un facteur de 2, semble identique à l'algorithme RWMH. La méthode DR anti apparaît légèrement supérieure quant à elle, puisque dans ce cas, un facteur de 2 semble démesuré. L'efficacité des méthodes locales semble rester inférieure à celle de la méthode DA, même après l'ajustement par la matrice de covariance B. Cependant, l'écart paraît petit à toutes fins pratiques et il est anticipé qu'un nombre légèrement plus élevé d'itérations pour la méthode RWMH ou une méthode avec plus de 2 essais multiples mènent à des mesures de convergence semblables à la méthode DA.

En conclusion, cet exemple démontre bien l'importance du choix de la distribution instrumentale pour une méthode locale. Bien que versatiles, ces méthodes peuvent être très sous-optimales surtout quand la variance de la distribution instrumentale est loin de celle de la distribution cible. Pour cet exemple, l'idée initiale était d'examiner si des méthodes locales pouvaient être aussi performantes que la méthode DA. En réalité, il apparaît que la méthode RWMH$(\mathbf{x}_j,\ 0{,}0001I_8)$, une des méthodes les plus versatiles, ne converge pas de façon satisfaisante. Une légère modification telle que l'algorithme RWMH$(\mathbf{x}_j,\ 0{,}005A)$ ou des méthodes plus avancées telles que le MTM, MTMHR ou DR anti n'améliorent pas grandement la performance. L'algorithme AM est le seul qui, grâce à sa nature adaptative, semble être en voie de convergence. À la lumière de ces résultats, il semble qu'une étape adaptative est cruciale dans cet exemple, un fait qui est démontré par la matrice de covariance (3.2) obtenue à partir de valeurs générées par l'algorithme convergent DA. Dans cette situation, une estimation efficace (contrairement à la matrice A qui a été obtenue par une méthode sous-optimale) de la matrice de cova-

Tableau 3.II – Résumé des résultats de méthodes locales (exemple 2) avec ajustement de la matrice de covariance (Période de chauffe 10 000, Itérations 4 000 000).

Algorithme	valeur-p pour $\beta_6 = -0,1$
RWMH - Instrumentale $N(\mathbf{x}_j,\ 0,7B)$	0,75502
{écart-type de la valeur-p}	$(0,02932)$
{DSCM}	139,17
{Taux d'acceptation}	0,257
MTM 2 essais - Instrumentale $N(\mathbf{x}_j, B)$	0,75675
{écart-type de la valeur-p}	$(0,02328)$
{DSCM}	224,86
{Taux d'acceptation}	0,312
DR anti - $(\mathbf{Z}_{j+1} \sim N(\mathbf{0}, B), \sigma_1 = \sigma_2 = 0,95)$	0,75573
{écart-type de la valeur-p}	$(0,02056)$
{DSCM}	270,73
{Taux d'acceptation}	0,401
DA	0,75683
{écart-type de la valeur-p}	$(0,01081)$
{DSCM}	1322,69
{Taux d'acceptation}	0,716
Approximation de trosième ordre	0,75283

riance semble nécessaire en un premier lieu. Une fois cette estimation obtenue, des méthodes locales RWMH, MTM et DR anti convergent de façon raisonnable. D'entre elles, l'algorithme DR anti semble la méthode la plus performante et mène à des mesures de convergence près de celles de la méthode DA. En plus d'explorer une comparaison entre la méthode DA et certaines méthodes locales, ce chapitre visait aussi à présenter une explication pour la piètre performance de l'algorithme RWMH($\mathbf{x}_j$, $0,0001I_8$) dans cet exemple, un problème qui avait été laissé ouvert dans Bédard et Fraser (2008).

Une question reste toutefois sans réponse, soit comment obtenir une estimation fiable de la matrice de covariance de la distribution cible. Précédemment, B a été calculée à des fins d'illustration et il est clair qu'en pratique on ne dispose pas d'un algorithme déjà convergent pour calculer cette matrice. Une réponse à cette question peut être d'utiliser l'inverse de la matrice hessienne comme approximation de par sa relation à la matrice d'information. Cette méthode est bien sûr vastement dépendante de la densité cible. Une autre réponse peut être l'algorithme AM qui souvent peut donner rapidement une estimation suffisamment fiable de la matrice de covariance. Il est possible qu'ensuite cela soit plus avantageux, en termes de coût computationnel, d'utiliser cette estimation dans un algorithme de type RWMH que de laisser poursuivre l'algorithme AM. Une autre possibilité est de créer un algorithme qui combine l'étape adaptative à une méthode DR ou MTM.

De toute manière, une méthode locale semble ardue dans le contexte de cet exemple. Pour y arriver, un coût computationnel et analytique considérable est nécessaire. Par conséquent, il est plus avantageux d'utiliser une méthode globale telle que le DA.

D'un autre côté, dans un contexte où une distribution instrumentale générique (telle que $N(\mathbf{0}, I_n)$) est déjà suffisamment adéquate et qu'aucune estimation de la matrice de covariance est nécessaire, une méthode locale peut être à meilleur marché.

Par exemple, le tableau 3.III résume les mêmes mesures d'efficacité dans le contexte de l'exemple de la section 2.3.1. Dans ce cas, les méthodes locales se comportent déjà de façon suffisamment satisfaisante en utilisant une matrice de covariance instrumentale

Tableau 3.III – Résumé des résultats de méthodes locales (exemple 1) (Période de chauffe 10 000, Itérations 4 000 000).

Algorithme	valeur-s pour $\beta = 1$
RWMH - Instrumentale $N(\mathbf{x}_j,\ 0,3I_3)$	0,10794
{écart-type de la valeur-s}	$(0,01483)$
{DSCM}	0,123851
{Taux d'acceptation}	0,251
MTM 2 essais - Instrumentale $N(\mathbf{x}_j,\ 0,4I_3)$	0,10837
{écart-type de la valeur-s}	$(0,01192)$
{DSCM}	0,194470
{Taux d'acceptation}	0,334
DR anti - $(\mathbf{Z}_{j+1} \sim N(\mathbf{0},I_3), \sigma_1 = \sigma_2 = 0,6)$	0,10851
{écart-type de la valeur-s}	$(0,01086)$
{DSCM}	0,233924
{Taux d'acceptation}	0,389
DA	0,10764
{écart-type de la valeur-s}	$(0,007945)$
{DSCM}	0,744361
{Taux d'acceptation}	0,665

I_3. Elles sont légèrement moins efficaces que la méthode DA et d'une certaine manière, cette situation est semblable à celle décrite dans les tableau 3.II. Les méthodes globales peuvent généralement posséder une vitesse de convergence supérieure aux méthodes locales (qui en fait peuvent être tout au plus géométriquement ergodiques et ce seulement sous certaines conditions particulières (voir Mengersen et Tweedie (1996))). Toutefois, la facilité d'application de ces méthodes locales et leur versatilité sont souvent leurs principaux attraits.

Dans le cas de l'exemple de la section 2.3.1, il peut être avantageux d'utiliser une méthode locale, puisque le temps d'exécution ainsi que le temps d'implémentation (temps de codage) en seront inférieurs à ceux de la méthode DA et donc, pour une précision donnée, un nombre d'itérations suffisamment élevé sera rapidement atteignable. D'un côté pratique, il peut être préférable d'opter pour un algorithme légèrement moins efficace mais plus facilement implémentable puisque le temps de simulation est généralement à meilleur marché que le temps de l'utilisateur.

De toute manière, la distribution cible dictera dans chaque cas la méthode appropriée, mais il est clair que l'algorithme DA tout comme une méthode locale peuvent constituer un choix optimal selon le contexte.

CONCLUSION

En première partie, ce mémoire a effectué un survol des algorithmes MCMC ainsi que des grands théorèmes et résultats portant sur leur convergence. Il est à espérer que ce chapitre a permis au lecteur d'apprécier l'importance de ces méthodes, leur facilité d'application ainsi que les raisons de leur popularité dans la communauté scientifique et surtout bayésienne.

Au deuxième chapitre, il a été question d'un algorithme Metropolis avec ajustement directionnel (DA) récemment développé dans Bédard et Fraser (2008). L'idée de cette méthode était de combiner la versatilité de l'approche Metropolis-Hastings de type marche aléatoire (RWMH) avec la performance supérieure de l'échantillonneur indépendant (IS). Pour ce faire, la méthode DA employait une densité instrumentale Student centrée au mode de la densité cible avec des degrés de liberté distincts selon la direction. Ces degrés de liberté étaient choisis tels que les rapports à un point $s*$ et au mode des densités instrumentale et cible coïncidaient. En plus, les auteurs de la méthode DA avaient choisi de faire correspondre la courbure des deux densités au mode par l'emploi d'une reparamétrisation menant à des matrices hesiennes identiques.

De par sa nature, l'algorithme DA a un rendement efficace dans le cas de distributions lisses et unimodales. À priori, cette condition est peu restrictive puisqu'en pratique plusieurs problèmes de ce type existent. Toutefois, une utilisation consciencieuse de la méthode DA nécessite une première étape d'optimisation afin d'identifier le mode et de s'assurer qu'il est unique. À cette fin, il existe des algorithmes préétablis dans la plupart des progiciels de statistique, comme par exemple la fonction nlm ou optim du logiciel R. Les deux exemples étudiés plus tôt dans ce mémoire portaient sur des problèmes de régression. Dans ces cas, l'utilisateur dispose souvent de la valeur des estimateurs par moindres carrés. Cela peut servir de point de départ dans un algorithme d'optimisation et, cette valeur étant généralement près du maximum, la convergence est la plupart du temps rapide. Cependant, lorsque ce type d'estimateur n'est pas disponible ou si la fonc-

tion de densité est complexe ou possède des extremums locaux, le problème devient beaucoup plus difficile à cerner. Dans ces cas, il est possiblement plus avantageux de considérer un algorithme local comme le RWMH.

La deuxième étape de l'algorithme DA, c'est-à-dire l'évaluation de la matrice hessienne au mode, représente une difficulté tout aussi importante. Dans le logiciel R, il est possible d'évaluer numériquement cette matrice à l'aide de la fonction fdHess. Selon le contexte, cette technique peut être très instable. Par exemple, l'évaluation de la matrice hessienne pour l'exemple de la section 2.3.2 a nécessité plusieurs ajustements avant d'obtenir des résultats fiables. D'un autre côté, il est possible d'utiliser un logiciel analytique comme Maple afin d'éviter les difficultés de la différenciation numérique. Enfin, même si l'ajustement des courbures semble bénéfique à priori, son impact véritable n'est pas tout à fait clair. De futures études sont nécessaires afin de déterminer l'importance de l'utilisation de matrices hessiennes identiques et surtout si les avantages conférés en termes de convergence l'emportent sur les difficultés numériques potentielles.

Quoiqu'il en soit, un des objectifs du chapitre 2 était d'étudier l'évolution de certaines mesures de convergence empiriques par rapport au point $s* = \lambda\sqrt{n}$. En utilisant deux exemples tirés de Bédard et Fraser (2008), l'efficacité de l'algorithme semblait être optimisée pour des valeurs de λ entre 2 et 4. Sur cet intervalle, les mesures empiriques indiquaient une amélioration modeste mais soutenue de la performance. Ces deux exemples semblaient indiquer que le choix de λ était important mais non pas capital. Cependant, le dernier exemple unidimensionnel a démontré que l'algorithme DA pouvait dans certaines situations mener à une convergence fortement dépendante de λ. Dans ce cas, un choix de $\lambda = 1$ résultait en un algorithme très sous-optimal tandis qu'un choix de $\lambda \geq 42$ représentait la situation idéale. En pratique, la meilleure approche est donc de simuler l'algorithme DA en employant différentes valeurs de λ tout en surveillant le comportement des critères de convergence.

Finalement, le troisième chapitre visait à comparer l'approche DA avec des algorithmes locaux beaucoup plus versatiles et facilement implémentables. L'exemple de la

section 2.3.2 a été utilisé comme paradigme pour la majeure partie de la discussion. Dans ce contexte, plusieurs méthodes locales ont été trouvées inefficaces, un fait qui avait été démontré dans Bédard et Fraser (2008) pour l'algorithme RWMH. Pour cet exemple, une étape adaptative semblait nécessaire avant toute méthode locale et donc l'approche DA était préférable. D'un autre côté, dans le contexte de l'exemple de la section 2.3.1, les méthodes locales semblaient répondre au problème de façon raisonnable sans ajustement de la matrice de covariance. Dans ce cas, une méthode locale pouvait constituer un choix moins dispendieux en termes de temps d'exécution et d'implémentation.

En conclusion, l'algorithme DA est une alternative attrayante aux algorithmes MCMC traditionnels. Bien qu'une attention particulière doit être portée quant au choix du paramètre λ, cette méthode semble efficace, surtout en grande dimension. Elle comporte certains désavantages comme un temps d'exécution considérable et des difficultés numériques potentielles, mais elle peut constituer un choix optimal dans certains contextes.

BIBLIOGRAPHIE

Bédard, M., Douc, R. et Moulines, E. (2010a). Scaling analysis of delayed rejection MCMC methods. Article soumis.

Bédard, M., Douc, R. et Moulines, E. (2010b). Scaling analysis of multiple-try MCMC methods. *Stochastic Processes and their Applications*. À paraître.

Bédard, M. et Fraser, D. A. S. (2008). On a directionally adjusted Metropolis-Hastings algorithm. *International Journal Of Statistical Sciences*, **9 (Special Issue)**, 33–57.

Bédard, M., Fraser, D. A. S. et Wong, A. (2007). Higher accuracy for bayesian and frequentist inference : Large sample theory for small sample likelihood. *Statistical Science*, **22**, 301–321.

Chan, K. et Geyer, C. (1994). Discussion de l'article Tierney (1994). *Annals of Statistics*, **22**, 1747–1758.

Cogburn, R. (1972). The central limit theorem for Markov processes. Dans *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, 485–512. University of California Press.

Cowles, M. et Carlin, B. (1996). Markov chain Monte Carlo convergence diagnostics :a comparative review. *Journal of the American Statistical Association*, **91**, 883–904.

Cox, D. R. et Snell, E. J. (1981). *Applied Statistics : Principles and Examples*. Chapman and Hall. 192p.

Damerdji, H. (1994). Strong consistency of the variance estimator in steady-state simulation output analysis. *Mathematics of Operations Research*, **19**, 494–512.

Flegal, J. M. et Jones, G. L. (2010). Batch means and spectral variance estimators in Markov chain Monte Carlo. *Annals of Statistics*, **38**, 1034–1070.

Fraser, D. et Reid, N. (1993). Ancillaries and third order significance. Rapport technique, Université de Toronto.

Gelman, A., Carlin, J. B., Stern, H. S. et Rubin, D. B. (1995). *Bayesian Data Analysis*. New York : Chapman.

Gelman, A. et Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 473–483.

Genz, A. (1972). An adaptive multidimensional quadrature procedure. *Computer Physics Communications*, **4**, 11–15.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. Dans *Bayesian Statistics 4*, 169–193. Oxford University Press.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo (with discussion). *Statistical Science*, **7**, 473–511.

Haario, H., Saksman, E. et Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, **7**, 223–242.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their application. *Biometrika*, **57**, 97–109.

Kipnis, C. et Varadhan, S. R. S. (1986). Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Communications in Mathematical Physics*, **104**, 1–19.

Liu, J. S., Liang, F. et Wong, W. H. (2000). The multiple-try method and local optimization in Metropolis sampling. *Journal of the American Statistical Association*, **95**, 121–134.

Liu, J. S., Wong, W. H. et Kong, A. (1995). Covariance structure and convergence rate of the Gibbs sampler with various scans. *Journal of the Royal Statistical Society : Series B*, **57**, 157–169.

Matthews, P. (1993). A slowly mixing Markov chain with implications for Gibbs sampling. *Statistics and Probability Letters*, **17**, 231–236.

Mengersen, K. et Tweedie, R. L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *Annals of Statistics*, **24**, 101–121.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. et Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1092.

Mira, A. (2001). On Metropolis-Hastings algorithms with delayed rejection. *Metron*, **LIX**, 231–241.

Raftery, A. et Lewis, S. (1992). How many iterations in the Gibbs sampler ? Dans *Bayesian Statistics 4*, 763–773. Oxford University Press.

Roberts, G. O. (1999). A note on acceptance rate criteria for CLTs for Metropolis-Hastings algorithms. *Journal of Applied Probability*, **36**, 1210–1217.

Roberts, G. O., Gelman, A. et Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, **7**, 110–120.

Roberts, G. O. et Rosenthal, J. S. (1997). Geometric ergodicity and hybrid Markov chains. *Electronic Communications in Probability*, **2**, 13–25.

Roberts, G. O. et Rosenthal, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, **16**, 351–367.

Roberts, G. O. et Rosenthal, J. S. (2004). General state space Markov chains and MCMC algorithms. *Probability Surveys*, **1**, 20–71.

Tierney, L. (1994). Markov chains for exploring posterior disiributions (with discussion). *Annals of Statistics*, **22**, 1701–1762.

Tierney, L. et Mira, A. (1999). Some adaptive Monte Carlo methods for Bayesian inference. *Statistics in Medicine*, **8**, 2507–2515.

Annexe I

Programmes R

```
#Programme R pour la méthode DA, pour s*=1 ... 20 (exemple 1)



ADR <- function (BurnIN, Itns, s1, s2,d) {

#Période de chauffe =10000, itérations=4M, s* minimum=sqrt(3), s*maximum=20*sqrt(3), d=sqrt(3)


    Sl<-seq(s1,s2,d)

    #Définition de la densité

    Pii<-function(par) {
        x<-c(-3,-2,-1,0,1,2,3)
        y<-c(-2.68, -4.02, -2.91, 0.22, 0.38, -0.28, 0.03)

        exp(-7*par[3])*prod(c( (1+ (  y-par[1]-par[2]* x)^2  /7/exp(2*par[3]))^-4))}

    #Définition du log de la densité

    Piil<-function(par) {
        x<-c(-3,-2,-1,0,1,2,3)
        y<-c(-2.68, -4.02, -2.91, 0.22, 0.38, -0.28, 0.03)

        log(exp(-7*par[3])*prod(c( (1+ (  y-par[1]-par[2]* x)^2  /7/exp(2*par[3]))^-4)))}


    #Détermination du mode

    MM<-optim(c(-1.32,0.67,1), Pii, method="BFGS", control=list(fnscale=-1))
    MMM<-as.function(MM[1])
    Xhat<-MMM()

    #Calcul de la matrice hessienne, de sa racine carrée et sa racine carrée inverse

    G<-fdHess(Xhat, Piil)
    T<-as.function(G[3])
    H<-T()
    pd1 <- pdSymm(diag(3))
    matrix(pd1) <- -H
    Hf<- pdMatrix(pd1,0.5)
    Hfinv<-solve(Hf)

    #Initialisation de variables

    p<-(1)
    AvDOF<-(1)
```

```
Svalue<-(1)
D3d<-(0)
D1d<-(0)
ACCEP<-(1)
StDOF<-(1)
StSval<-(1)
StD3<-(1)
StD1<-(1)


#Boucle pour un s* donné

while (s2+1>=s1) {

    print(s1)

    #Valeur initiale et initialisation de variables

    VI<-c(1,1,1)
    j<-(1)
    SS<-rep(0,Itns)
    F<-(1:Itns)

    #Standardisation de la valeur initiale et calcul du degré de liberté optimal

    Xstar<-Hf %*%(VI-Xhat)
    Ustar<-Xstar/(sqrt(sum(Xstar^2)))
    Rsquare<-rep(1,50)
    Qsquare<-rep(1,50)
    Rsquare<- Rsquare*2*log(   (Pii(Xhat))/(Pii( Xhat + Hfinv%*% (Ustar*s1)))   )
    Qsquare<- Qsquare*(t(Ustar*s1))%*%(Ustar*s1)
    Min<-(1:50)
    Min<-abs(((Min+3)*log(1+Qsquare/(Min+3)))-Rsquare)
    Minpt<-which.min(Min)
    F[j]<-Minpt

    #Initialisation d'autres variables
    Di3d<-(0)
    Di1d<-(0)
    ACPT<-(0)

    #Boucle pour les itérations

    while ( j <Itns+BurnIN+1){
        VIS<-VI

        #Proposition de valeur et calcul du degré de liberté optimal

        R<- c(rnorm(1),rnorm(1), rnorm(1))
        UstarN<- R/ sqrt(sum(R^2))

        RsquareN<-rep(1,50)
        QsquareN<-rep(1,50)


        RsquareN<- RsquareN*2*log(   (Pii(Xhat))/(Pii( Xhat + Hfinv%*% (UstarN*s1)))   )
        QsquareN<- QsquareN*(t(UstarN*s1))%*%(UstarN*s1)
```

```
            MinN<-(1:50)
            MinN<-abs(((MinN+3)*log(1+QsquareN/(MinN+3)))-RsquareN)
            MinptN<-which.min(MinN)
            F[j+1]<-(MinptN)



            #Proposition d'une distance radiale

            S<-rchisq(1, MinptN, ncp=0)
            SN<-(MinptN+3)^0.5* sqrt(sum(R^2))/sqrt(S)
            Ystar<-UstarN*SN


            #Calcul de alpha

            Td<-function(x,y)     ( gamma((y+3)/2)/ pi^1.5 /gamma(y/2) )* (
(1+( sum(x^2)/(y+3)) ) ^-((y+3)/2)) * (y+3)^-1.5
            Num<-  Pii(Xhat+Hfinv%*%Ystar)*Td(Xstar, Minpt)
            Den<-   Pii(Xhat+Hfinv%*%Xstar)*Td(Ystar,MinptN)
            alpha<-min(1,Num/Den)


            #Acceptation ou refus

            T<-runif(1)
            if (T<alpha)
            {VI<-Xhat+(Hfinv%*%Ystar)
            Minpt<-MinptN
            Xstar<-Ystar
            ACPT<-ACPT+1
            }

            #Distance et valeur-s

            Di3d<- (sum((VIS-VI)^2))+Di3d
            if (j>BurnIN) {
            if (VI[2] >=1) SS[j-BurnIN]<-(1)}


            j<-j+1
            }
            #Fin boucle itérations

          FN<-(0)
          SSN<-(0)
          Dis3d<-(0)
          Dis1d<-(0)
          JK<-Itns/1000
          for (g in 1:JK){

          #Séparation en séries de 950
          lo<-50+1000*(g-1)
          hi<-1000*g-1
          FN[g]<-mean(F[lo:hi])
          SSN[g]<-mean(SS[lo:hi])
          }
```

```
        #Enregistrement de sorties
        AvDOF[p]<-mean(FN)
        Svalue[p]<-mean(SSN)
        StDOF[p]<-sd(FN)
        StSval[p]<-sd(SSN)
        D3d[p]<- Di3d
        ACCEP[p]<-ACPT

        p<-p+1

        s1<-s1+d
        }
        #Fin boucle s*


    #Sorties globales

    par(mfrow= c(2,2))

    plot(Sl,AvDOF)

    for (mn in 1:length(Sl)){
    segments(Sl[mn],AvDOF[mn]+StDOF[mn],Sl[mn], AvDOF[mn]-StDOF[mn])}
    plot(Sl,Svalue)
    for (mn in 1:length(Sl)){
    segments(Sl[mn],Svalue[mn]+StSval[mn],Sl[mn], Svalue[mn]-StSval[mn])}
    plot(Sl,D3d/Itns)
    print("Svalue")
    print(Svalue)
    print("Dis3")
    print(D3d/Itns)
    print("AvDOF")
    print(AvDOF)
    print("StDOF")
    print(StDOF)
    print("StSval")
    print(StSval)
    print("ACEPT")
    print(ACCEP)


}
```

```
#Programme R pour la méthode RWMH avec N(xj, 0,000118)


RWMH <- function (BurnIN, Itns, Fac1, AC) {

    #Période de chauffe =10000, Itns=4M, Fac1=0,0001,AC=diag(1,8)
    #Initialisation de variables
    ACPT<-0
    DIS3d<-0
    A<-Itns/1000
    Accep<-0
    Dis<-0
    j<-1
    P<-0
    SN<-0

    #Facteur d'ajustement * Matrice de covariance
    Sigma<-Fac1*AC


    # Définition de la densité cible
    Pii<-function(par) {
        Func<-(0)
        Par2<-par[1:7]

        for (i in 1:32) {
        Func[i]<-      dt(   exp(par[8])*D[i] + XX[i,]%*%Par2 , 4)
        }
        prod(Func) * exp(25*par[8])
        }


    #Valeur initiale
    VI<-c(b0 ,log(s))

    while  (j < Itns + BurnIN+1) {
        VIS<-VI

        # Saut proposé
        Yn<- rmvnorm(1, mean=VI, Sigma)



        Num<-Pii(Yn)          #Numérateur du ratio alpha
        Den<-Pii(VI)            #Dénominateur du ratio alpha


        alpha<- min (1,(Num/Den))


        #Acceptation ou refus
        C<- runif(1);
         if (C < alpha) {
        VI<-Yn
        Accep<-Accep+1 }
```

```
        #Calcul distance et valeur-p si l'itération est au-delà de la période de chauffe

        if (j>BurnIN){
        Tk<-(VI[6])/(sqrt(CK[6,6])*exp(VI[8]))
        TO<-(-0.0878 +0.1)/(sqrt(CK[6,6])*s)
        if (Tk<TO) {P[j-BurnIN]<-1} else {P[j-BurnIN]<-0}
        Dis<-Dis+ (sum((VI-VIS)^2))
        }

        j<-j+1
        }


    #Division des sorties en séries de 950

    for (i in 1:A) {

        SN[i]<-sum ( P[ (1000 *(i-1) +51 ) : (1000*i)]) / 950
        }

    #Calcul valeur-p moyenne et écart-type

    MN<-mean(SN)
    STDEV<-sd(SN)

    #Sorties globales

    print("prob moyene")
    print(MN)
    print("standard deviation ")
    print(STDEV)
    print("Accepations total ")
    print(Accep/(Itns+BurnIN))
    print("distance totale ")
    print(Dis/Itns)

}
```