REVUE DE STATISTIQUE APPLIQUÉE

ANAS ALTALEB
CHRISTIAN P. ROBERT

Analyse bayésienne du modèle Logit : algorithme par tranches ou Metropolis-Hastings ?

Revue de statistique appliquée, tome 49, n° 4 (2001), p. 53-70
<http://www.numdam.org/item?id=RSA_2001__49_4_53_0>

Rev. Statistique Appliquée, 2001, **XLIX** (4), 53-70

ANALYSE BAYÉSIENNE DU MODÈLE LOGIT : ALGORITHME PAR TRANCHES OU METROPOLIS-HASTINGS?

Anas Altaleb[1], Christian P. Robert[2]

[1] *Département de Mathématiques Appliquées Probabilités & Statistiques Université de Damas, Syrie*

[2] *Ceremade Université Paris 9 Dauphine xian@ceremade.dauphine.fr*

RÉSUMÉ

L'étude d'un modèle de régression non linéaire généralisé, le modèle de régression logistique, est abordée suivant deux méthodes : celle de Damien et Walker (1996) et une technique générique de Metropolis-Hastings. L'ensemble des résultats exposés est illustré par différentes simulations, qui montrent les performances supérieures de l'algorithme de Metropolis-Hastings en termes de vitesse de convergence vers la loi stationnaire et de rapidité d'exploration de la surface de la loi a posteriori, par rapport à la méthode de Damien et Walker.

Mots-clés : *méthode d'intégration de Monte-Carlo; méthode de Monte-Carlo par chaînes de Markov; Rao-Blackwellisation; échantillonnage par tranches; statistique bayésienne; stationnarité.*

ABSTRACT

In this article, we examine some Markov chain Monte Carlo methods which often appear in the treatment of complex statistical models. We consider a generalized non-linear regression model -the Logit model- according to two methods : Damien and Walkers's (1996) (slice sampler) and a generic Metropolis-Hastings algorithm. The results presented here are illustrated by different simulations, and show the superior performances of the Hastings-Metropolis algorithm in terms of convergence to the stationary distribution and exploration of the posterior distribution surface, against Damien and Walker's method.

Keywords : *Bayesian statistic; Markov chain Monte Carlo; Monte Carlo integration; Rao-Blackwellisation; slice sampler; stationarity; convergence assessment.*

1. Introduction

Du point de vue bayésien, il n'existe pas de différence fondamentale entre l'observation et le paramètre d'un modèle statistique, tous deux étant considérés comme quantités variables. Donc, si on note D la donnée, de loi d'échantillonnage $f(D|\theta)$, et θ le paramètre du modèle considéré (plus, éventuellement, les variables latentes), de loi a priori π, une inférence formelle requiert la mise à jour de la distribution conditionnelle $f(\theta|D)$ du paramètre. La détermination de $\pi(\theta)$ et de

$f(D|\theta)$ donne $f(D,\theta)$ par

$$f(D,\theta) = f(D|\theta)\pi(\theta).$$

Ayant observé D, on peut utiliser le Théorème de Bayes pour déterminer la distribution de θ conditionnellement aux données (ou loi a posteriori)

$$\pi(\theta|D) = \frac{f(D|\theta)\pi(\theta)}{\int f(D|\theta)\pi(\theta)d\theta}. \tag{1}$$

Pour l'approche bayésienne, toutes les caractéristiques de la loi a posteriori sont importantes pour l'inférence : moment, quantiles, etc. Ces quantités peuvent souvent être exprimées en termes d'espérance conditionnelle d'une fonction de θ par rapport à la loi a posteriori

$$E\left[h(\theta)|D\right] = \frac{\int h(\theta)f(D|\theta)\pi(\theta)d\theta}{\int f(D|\theta)\pi(\theta)d\theta}. \tag{2}$$

Mais, il est rare de disposer d'une loi a posteriori $\pi(\theta|D)$ qui soit explicite et il est alors nécessaire de pouvoir simuler un échantillon $(\theta_1, ..., \theta_n)$ qui soit approximativement iid de loi $\pi(\theta|D)$, afin de déterminer soit des régions de confiance, soit la structure générale de la loi (détection de modes, d'asymétries, etc.). (Voir Robert, 1992, Chap. 9.)

Dans la section suivante nous présentons un modèle de régression non linéaire généralisée : le modèle de régression logistique ou modèle Logit. Ensuite, nous posons les conditions nécessaires pour l'utilisation des algorithmes de Monte-Carlo par chaînes de Markov (MCMC) et nous introduisons quelques algorithmes MCMC, en particulier l'algorithme de Metropolis-Hastings à marche aléatoire (où nous approchons la loi a posteriori du modèle Logit par une loi instrumentale normale bidimensionnelle), et la méthode d'échantillonnage de Gibbs du point de vue de Damien et Walker (1996). La dernière section concerne la comparaison de ces algorithmes et quelques critères pour contrôler la convergence des algorithmes MCMC. L'exemple traité dans cet article illustre les comportements et les performances des algorithmes MCMC pour l'approximation de la distribution a posteriori en question. D'un point de vue pratique, on conclut que l'algorithme de Metropolis-Hastings pour une approximation normale bidimensionnelle est beaucoup plus efficace en termes de vitesse de convergence vers la loi stationnaire et de rapidité d'exploration de la surface de la loi a posteriori, que la méthode de Damien et Walker. L'ensemble des résultats exposés est illustré par différentes simulations en utilisant le logiciel CODA (Best *et al.*, 1995) pour contrôler la convergence.

2. Le modèle de régression logistique

2.1. Introduction

Un modèle standard de régression qualitative est le modèle de régression logistique ou modèle Logit, où la loi de y conditionnelle aux variables explicatives $z \in \mathbb{R}^p$ est

$$P(y = 1) = 1 - P(y = 0) = \frac{\exp(z^t\gamma)}{1 + \exp(z^t\gamma)},$$

fondé sur la dépendance logistique entre les variables explicatives et l'observation.

Considérons le cas particulier où $z = (1, x)$ et $\gamma = (\alpha, \beta)$: les variables binaires y_i à valeurs dans $\{0, 1\}$ sont associées à des variables explicatives x_i et sont modélisées suivant une loi de Bernoulli de probabilité conditionnelle

$$y_i|x_i \sim B\left(\frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}\right). \tag{3}$$

Supposons que nos paramètres suivent a priori une loi impropre $\pi(\alpha, \beta) = 1$. La vraisemblance de notre modèle, pour un échantillon $(y_1, x_1), ..., (y_n, x_n)$, est égale à

$$f(y_1, ..., y_n|x_1, ..., x_n, \alpha, \beta) = \prod_{i=1}^{n} \frac{\exp\{(\alpha + \beta x_i)y_i\}}{1 + \exp(\alpha + \beta x_i)}.$$

La loi a posteriori de (α, β), se déduit alors par application formelle du Théorème de Bayes

$$\begin{aligned}\pi(\alpha, \beta \mid D) &\propto f(y_1, ..., y_n|x_1, ..., x_n, \alpha, \beta)\pi(\alpha, \beta) \\ &\propto \prod_{i=1}^{n} \frac{\exp\{(\alpha + \beta x_i)y_i\}}{1 + \exp(\alpha + \beta x_i)} = \frac{\exp\{\sum_{i=1}^{n}(\alpha + \beta x_i)y_i\}}{\prod_{i=1}^{n}\{1 + \exp(\alpha + \beta x_i)\}}.\end{aligned}$$

2.2. Définition de la loi a posteriori

L'utilisation de lois a priori impropres, c'est-à-dire de mesures σ finie de masse infinie sur l'espace des paramètres, implique que la dérivation des lois a posteriori par la relation de proportionnalité

$$\pi(\theta|x) \propto f(x|\theta)\pi(\theta),$$

n'est pas nécessairement acceptable pour mettre en œuvre un algorithme de Metropolis-Hastings sur $f(x|\theta)\pi(\theta)$, car la « loi » correspondante peut ne pas exister, c'est-à-dire, $f(x|\theta)\pi(\theta)$ n'est pas forcément intégrable. On est confronté à cette difficulté

pour l'échantillonnage de Gibbs, par exemple, qui, contrairement aux algorithmes de Metropolis-Hastings, fonctionne avec des lois conditionnelles extraites de $\pi(\theta_1, .., \theta_q)$, elle-même représentée par la relation de proportionnalité ci-dessus. Il peut arriver que ces lois soient clairement définies et simulables, mais qu'elles ne correspondent pas à une loi jointe f, c'est-à-dire que f n'est pas intégrable (voir Robert, 1996, pour des exemples).

Ce fait, assez fréquent dans une approche bayésienne « généralisée », ne représente pas un défaut de l'échantillonnage de Gibbs, ni même un problème de simulation. Il ne faut cependant pas omettre la vérification de l'existence de f, ce que nous faisons dans le lemme suivant. On introduit l'hypothèse

Hypothèse [H]. — *Soit l'échantillon* $(x_1, y_1), .., (x_n, y_n)$ *avec* $n \geqslant 4$. *On suppose qu'il existe des* x_i *négatifs et des* x_i *positifs associés à des* $y_i = 1$ *et à des* $y_i = 0$.

LEMME 1. — *La loi a posteriori du modèle Logit* $\pi(\alpha, \beta \mid D)$ *où* D *représente les données est une vraie loi sous* **[H]**, *c'est-à-dire*

$$\iint \pi(\alpha, \beta \mid D) d\alpha \, d\beta < +\infty,$$

quelle que soit la taille des données.

Preuve. — Soit

$$I = \iint \prod_{i=1}^{n} \frac{\exp\{(\alpha + \beta x_i) y_i\}}{1 + \exp(\alpha + \beta x_i)} d\alpha d\beta.$$

Soit $p = \{i_1, ..., i_p\}$ l'ensemble des indices i pour lesquels $y_i = 1$. L'intégrale s'écrit

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \frac{\exp(\alpha p)\exp(\beta x_0)}{\prod_{i=1}^{n}(1 + \exp(\alpha + \beta x_i))} d\alpha d\beta$$

$$= I_1 + I_2 + I_3 + I_4 \,,$$

avec $x_0 = \sum_{i=1}^{n} x_i y_i$ et

$$I_1 = \int_{-\infty}^{0} \int_{-\infty}^{0} \frac{\exp(\alpha p)\exp(\beta x_0)}{\prod_{i=1}^{n}(1 + \exp(\alpha + \beta x_i))} d\alpha d\beta \,,$$

I_2, I_3 et I_4 correspondant aux autres quadrants de $\mathbb{R}^2$.

Le dénominateur

$$\prod_{i=1}^{n}(1+\exp(\alpha+\beta x_i))$$

se décompose sous la forme

$$\exp(n\alpha)\exp(\beta\sum_{i=1}^{n}x_i)+\ldots+\exp((p+1)\alpha)\left[\sum_{\sigma\in W_{p+1}}\exp(\beta\sum_{j=1}^{p+1}x_{\sigma(j)})\right]+$$

$$\exp(p\alpha)\left[\sum_{\sigma\in W_p}\exp(\beta\sum_{j=1}^{p}x_{\sigma(j)})\right]+\exp((p-1)\alpha)\left[\sum_{\sigma\in W_{p-1}}\exp(\beta\sum_{j=1}^{p-1}x_{\sigma(j)})\right]+\ldots+1,$$

où W_q est l'ensemble des injections σ de $\{1,\ldots,q\}$ dans $\{1,..,n\}$.

Dans I_1, considérons seulement dans le dénominateur les termes impliquant le facteur commun $\exp(\alpha(p-1))$, d'où

$$I_1 \leqslant \int_{-\infty}^{0}\int_{-\infty}^{0}\frac{\exp(\alpha p)\exp(\beta x_0)}{\exp(\alpha(p-1))\left[\sum_{\sigma\in W_{p-1}}\exp(\beta\sum_{j=1}^{p-1}x_{\sigma(j)})\right]}\,d\alpha d\beta$$

$$=\int_{-\infty}^{0}\frac{\exp(\beta x_0)}{\left[\sum_{\sigma\in W_{p-1}}\exp(\beta\sum_{j=1}^{p-1}x_{\sigma(j)})\right]}\,d\beta.$$

Or

$$\sum_{i=1}^{n}x_i y_i=\sum_{k=1}^{p}x_{i_k},$$

où $(i_1,i_2,..,i_p)$ sont les indices des y_i égaux à 1. Par **[H]**, il existe $\tilde{k}\in(i_1,i_2,..,i_p)$ tel que $y_{\tilde{k}}=1$ et $x_{\tilde{k}}>0$. Soit, sans perte de généralité, i_p tel que $y_{i_p}=1$ et $x_{i_p}>0$, et soit $\tilde{\sigma}$ tel que $\tilde{\sigma}(k)=i_k$ avec $k=1,2,\ldots,p-1$.

On a $\tilde{\sigma}\in W_{p-1}$, donc on peut majorer l'intégrale par

$$I_1\leqslant\int_{-\infty}^{0}\frac{\exp(\beta\sum_{k=1}^{p}x_{i_k})}{\left[\exp(\beta\sum_{k=1}^{p-1}x_{\tilde{\sigma}(k)})\right]}\,d\beta=\int_{-\infty}^{0}\frac{\exp(\beta\sum_{k=1}^{p}x_{i_k})}{\left[\exp(\beta\sum_{k=1}^{p-1}x_{i_k})\right]}\,d\beta$$

$$=\int_{-\infty}^{0}\exp(\beta x_{i_p})\,d\beta<\infty.$$

On procède de même pour les autres intégrales.

3. L'algorithme de Metropolis-Hastings

3.1. Définition

L'algorithme de Metropolis-Hastings repose sur l'utilisation d'une densité conditionnelle $q(y|x)$ par rapport à la mesure dominante pour le modèle. Il ne peut être mis en pratique que si $q(.|x)$ est simulable rapidement et est, soit disponible analytiquement à une constante près indépendante de x, soit symétrique, c'est-à-dire tel que $q(y|x) = q(x|y)$. L'algorithme de Metropolis-Hastings associé à la loi objective π et la loi conditionnelle q produit une chaîne de Markov $(x^{(t)})$ fondée sur la transition suivante :

Etant donné $x^{(t)}$

1. Générer $y_t \sim q(y|x^{(t)})$,

2. Prendre $$x^{(t+1)} = \begin{cases} y_t & \text{avec probabilité } \rho(x^{(t)}, y_t) \\ x^{(t)} & \text{avec probabilité } 1 - \rho(x^{(t)}, y_t) \end{cases} \quad [A1]$$

où

$$\rho(x^{(t)}, y_t) = \min\left\{ \frac{\pi(y_t)}{\pi(x^t)} \frac{q(x^{(t)}|y_t)}{q(y_t|x^{(t)})}, 1 \right\}.$$

La loi q est appelée loi instrumentale ou de proposition. Cet algorithme accepte systématiquement les simulations y_t telles que le rapport $\pi(y_t)/q(y_t|x^{(t)})$ est supérieur à la valeur précédente $\pi(x^{(t)})/q(x^{(t)}|y_t)$. Ce n'est que dans le cas symétrique que l'acceptation est gouvernée par le rapport $\pi(y_t)/\pi(x_t)$.

3.2. L'algorithme de Metropolis-Hastings à marche aléatoire

Un cas particulier de l'algorithme de Metropolis-Hastings est l'algorithme à marche aléatoire, pour lequel $q(y|x) = g(|y - x|)$. Par exemple, quand x est continu, $q(.|x)$ peut être une distribution normale multivariée de moyenne x et à matrice de variance-covariance Σ constante. Un choix prudent de distribution instrumentale conduit à générer un petit pas $y - x_t$ qui donne généralement un taux d'acceptation élevé mais aussi une chaîne à mélangeance lente. Un mauvais choix de distribution instrumentale conduit à un pas excessif et à des mouvements du centre à la queue de la distribution, et produit en général de petites valeurs de $\pi(y)/\pi(x_t)$ et un faible taux d'acceptation. Une telle chaîne conduit aussi à une lente mélangeance.

La solution pratique pour éviter ces deux écueils est d'utiliser un paramètre d'échelle pour la loi instrumentale. L'algorithme $[A1]$ autorisant cette dépendance, $q(y|x)$ peut ainsi être de la forme $g_\tau(|y - x|)$, c'est-à-dire que y_t peut s'écrire sous la forme $x_t + \tau\varepsilon_t$, ε_t étant une perturbation aléatoire de loi g, indépendante de x_t, et

τ est un paramètre d'échelle. La chaîne de Markov associée à q_τ est également une marche aléatoire sur $\mathcal{X}$.

L'algorithme $[A1]$ s'exprime alors sous la forme suivante :

1. Générer $y_t \sim g_\tau(y - x^{(t)})$.

2. Prendre $x^{(t+1)} = \begin{cases} y_t & \text{avec probabilité } \min = \left\{1, \dfrac{\pi(y_t)}{\pi(x^{(t)})}\right\} \\ x^{(t)} & \text{sinon} \end{cases}$ $[A2]$

Par rapport aux autres algorithmes de Metropolis-Hastings, l'algorithme de Metropolis-Hastings à marche aléatoire demande une analyse spécifique des taux d'acceptation, du fait de la dépendance de la loi instrumentale à la valeur précédemment acceptée. Un taux d'acceptation élevé n'indique pas que l'algorithme évolue correctement. A l'inverse, si le taux d'acceptation moyen est faible, les valeurs successives de $\pi(y_t)$ sont fréquemment petites par rapport à $\pi(x_t)$, c'est-à-dire que la marche aléatoire se déplace rapidement sur la surface de π (mais peut aussi trop visiter les queues de π). Une méthode automatique de paramétrisation ne peut pas garantir de performances optimales pour l'algorithme de Metropolis-Hastings à marche aléatoire, et les choix de taux opérés ici ne conduisent pas forcément à l'optimalité. Gelman, Gilks et Roberts (1996), recommandent pour les modèles de dimension 1 ou 2 un taux d'acceptation proche de 50%.

3.3. Approximation normale

Soit $\hat{\theta}$ l'estimateur du maximum de vraisemblance de θ fondé sur les données y. On a la loi asymptotique suivante

$$\sqrt{n}(\theta - \hat{\theta}) \approx \mathcal{N}(0, C), \tag{4}$$

où $C = I^{-1}(\theta)$ est l'information de Fisher pour une observation

$$I(\theta|y) = E\left[-\frac{\partial^2 l(\theta|y)}{\partial\theta\partial\theta^t}\right]$$

(dans le cas iid), $l(\theta|y)$ étant la log-vraisemblance. Pour l'approche bayésienne, $\hat{\theta}$ est fixé conditionnellement aux données y et θ est la variable. Connaissant le modèle et les données, la formule (4) implique que la densité a posteriori de θ est asymptotiquement normale de moyenne $\hat{\theta}$, et de matrice de variance-covariance C.

La normalité de l'estimateur du maximum de vraisemblance n'est bien sûr qu'asymptotique. Elle suggère cependant une distribution instrumentale q naturel dont l'effet d'approximation est corrigée par le rapport de Metropolis-Hastings.

3.4. Application au modèle Logit

Dans le cadre du modèle Logit, on peut donc approcher la loi a posteriori de (α, β) par l'algorithme de Metropolis-Hastings à marche aléatoire pour la loi instrumentale suivante :

$$\begin{pmatrix} \alpha^{(t+1)} \\ \beta^{(t+1)} \end{pmatrix} = \mathcal{N}_2 \left(\begin{pmatrix} \alpha^{(t)} \\ \beta^{(t)} \end{pmatrix}, \Sigma^{(t)} \right).$$

Pour calculer $\Sigma^{(t)}$, on prend le développement de Taylor de l'algorithme de la loi objectif au voisinage de $(\hat{\alpha}, \hat{\beta})$, qui est le maximum de vraisemblance, donné par

$$\log \pi(\alpha, \beta) = \log \pi(\hat{\alpha}, \hat{\beta}) + \frac{1}{2} \left(\alpha - \hat{\alpha}, \beta - \hat{\beta} \right) \nabla \nabla^t \log \pi(\hat{\alpha}, \hat{\beta}) \left(\alpha - \hat{\alpha}, \beta - \hat{\beta} \right),$$

où ∇ (*nabla*) représente l'opérateur gradient (et ∇^t l'opérateur divergence). Ensuite on remplace $E\left[\nabla \nabla^t \log \pi(\alpha, \beta)\right]$ par son observation. Cela implique le calcul de

$$\nabla \log \pi(\alpha, \beta) = \begin{pmatrix} \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{\exp(\alpha + \beta x_i)}{(1 + \exp(\alpha + \beta x_i))} \\ \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \frac{x_i \exp(\alpha + \beta x_i)}{(1 + \exp(\alpha + \beta x_i))} \end{pmatrix},$$

$$\nabla \nabla^t \log \pi(\alpha, \beta) = - \begin{pmatrix} \sum_{i=1}^{n} \frac{\exp(\alpha + \beta x_i)}{(1 + \exp(\alpha + \beta x_i))^2} & \sum_{i=1}^{n} \frac{x_i \exp(\alpha + \beta x_i)}{(1 + \exp(\alpha + \beta x_i))^2} \\ \sum_{i=1}^{n} \frac{x_i \exp(\alpha + \beta x_i)}{(1 + \exp(\alpha + \beta x_i))^2} & \sum_{i=1}^{n} \frac{x_i^2 \exp(\alpha + \beta x_i)}{(1 + \exp(\alpha + \beta x_i))^2} \end{pmatrix}.$$

De plus, on ajuste la matrice $\Sigma^{(t)}$ par un facteur d'échelle τ

$$\Sigma^{(t)} = \tau^2 \begin{pmatrix} \sum_{i=1}^{n} \frac{\exp(\alpha + \beta x_i)}{(1 + \exp(\alpha + \beta x_i))^2} & \sum_{i=1}^{n} \frac{x_i \exp(\alpha + \beta x_i)}{(1 + \exp(\alpha + \beta x_i))^2} \\ \sum_{i=1}^{n} \frac{x_i \exp(\alpha + \beta x_i)}{(1 + \exp(\alpha + \beta x_i))^2} & \sum_{i=1}^{n} \frac{x_i^2 \exp(\alpha + \beta x_i)}{(1 + \exp(\alpha + \beta x_i))^2} \end{pmatrix}^{-1}.$$

L'idée sous-jacente est de calibrer le facteur τ en fonction du taux d'acceptation de l'algorithme. Enfin, en substituant à (α, β) $(\alpha^{(t)}, \beta^{(t)})$, valeur courante du paramètre, on obtient $\Sigma^{(t)}$.

4. Méthode d'échantillonnage par tranche

4.1. Principe

Soit π une densité définie sur $\mathbb{R}^d$, à simuler. L'idée principale de Damien et Walker (1996) (voir aussi Damien, Wakefield et Walker, 1999) est d'introduire une variable auxiliaire Y et une densité jointe $g(x, y)$ de (X, Y) de sorte que la densité marginale de X soit donnée par π,

$$\pi(x) = \int g(x, y)dy\,.$$

Cette loi jointe peut ensuite être, par exemple, générée par un algorithme d'échantillonnage de Gibbs, à partir des densités conditionnelles correspondantes, $f(y|x)$ et $f(x|y)$, si elles sont facilement simulables :

1. Simuler $Y^{t+1} \sim g(y|X = x^t)$,
2. Simuler $X^{t+1} \sim g(x|Y = y^{t+1})$. [A3]

Sous la condition de positivité du support de g, (X^t) converge en loi vers π. La méthode de Damien et Walker, pour être utile, doit conduire à des distributions conditionnelles faciles à simuler (voir aussi Besag et Green, 1993). Par exemple, si

$$\pi(x) \propto f_0(x) \prod_{i=1}^{n} f_i(x),$$

où les $f_i(x)$ sont des fonctions inversibles et positives, c'est-à-dire telles qu'il soit aisé d'exhiber l'ensemble $A_i(u) = \{x : f_i(x) > u\}$, on peut représenter π sous la forme

$$\pi(x) \propto f_0(x) \prod_{i=1}^{n} \int \mathbf{1}_{\{u_i \leqslant f_i(x)\}} du_i$$

soit encore comme la loi marginale de

$$g(x, u_1, \ldots, u_n) \propto f_0(x) \prod_{i=1}^{n} \mathbf{1}_{\{u_i \leqslant f_i(x)\}}\,.$$

Remarquons que la densité conditionnelle pour chaque u_i est la densité uniforme sur l'intervalle $(0, f_i(x))$. De même la densité conditionnelle de x est donnée par la densité $f_0(x)$ restreinte à l'ensemble $A(u) = \{x : u_i < f_i(x); \quad i = 1, ..., n\}$. Par conséquent, l'échantillonnage de Gibbs pourra être appliqué facilement si l'inversion des conditions $u_i < f_i(x)$ est aisée.

4.2. Application au modèle Logit

Pour simuler la loi a posteriori du modèle Logit suivant la méthode de Damien et Walker, on remarque, par une simple application de la décomposition précédente que toutes les lois conditionnelles sont disponibles et nous permettent d'approcher la distribution a posteriori par l'échantillonnage de Gibbs de n variables auxiliaires $w_1, w_2, ..., w_n$ indépendantes. On peut en effet écrire la loi a posteriori comme marginale de

$$\pi(\alpha, \beta \mid D) \propto \prod_{i=1}^{n} \mathbf{1}\left(w_i \leqslant \frac{\exp\{(\alpha + \beta x_i) y_i\}}{1 + \exp(\alpha + \beta x_i)}\right).$$

Une mise en œuvre directe de l'échantillonnage de Gibbs dans ce contexte conduit à simuler $\pi(\alpha, \beta, w \mid D)$. La génération de ce modèle consiste en les étapes suivantes, où

$$\rho_i(\alpha, \beta) = \frac{\exp\{(\alpha + \beta x_i) y_i\}}{1 + \exp(\alpha + \beta x_i)}$$

et où w_i dénote les variables auxiliaires. Et

$$\alpha, \beta \mid w_1, ..., w_n \sim \mathcal{U}_{\bigcap_{i=1}^{n} \{(\alpha,\beta) : w_i \leqslant \rho_i(\alpha,\beta)\}},$$

s'écrit $\alpha, \beta \mid w_1, .., w_n \sim \mathcal{U}_{\bigcap_{i=1}^{n} \{\{(\alpha,\beta):w_i \leqslant \rho_i(\alpha,\beta) \text{ et } y_i=1\} \bigcup \{(\alpha,\beta):w_i \leqslant \rho_i(\alpha,\beta) \text{ et } y_i=0\}\}}$.

L'algorithme de Damien et Walker s'écrit alors

1. Simuler les w_i suivant les lois uniformes

$$w_i \mid \alpha, \beta \sim \mathcal{U}_{\{0, \rho_i(\alpha,\beta)\}} \qquad i = 1, .., n.$$

2. Simuler $\alpha \mid \beta, w_1, ..., w_n \sim$ [A4]

$$\mathcal{U}_{\left\{\left\{\bigcap_{y_i=1} \left\{\alpha : \alpha \geqslant \log\left(\frac{w_i}{1-w_i}\right) - \beta x_i\right\}\right\} \bigcap \left\{\bigcap_{y_i=0} \left\{\alpha : \alpha \leqslant \log\left(\frac{1-w_i}{w_i}\right) - \beta x_i\right\}\right\}\right\}}$$

3. Simuler $\beta \mid \alpha, w_1, ..., w_n \sim$

$$\mathcal{U}_{\left\{\begin{array}{l} \bigcap_{y_i=1, x_i \geqslant 0} \left\{\beta : \beta > \dfrac{\log\left(\frac{w_i}{1-w_i}\right) - \alpha}{x_i}\right\} \bigcap_{y_i=0, x_i<0} \left\{\beta : \beta > \dfrac{\log\left(\frac{1-w_i}{w_i}\right) - \alpha}{x_i}\right\} \\ \bigcap_{y_i=1, x_i<0} \left\{\beta : \beta < \dfrac{\log\left(\frac{w_i}{1-w_i}\right) - \alpha}{x_i}\right\} \bigcap_{y_i=0, x_i \geqslant 0} \left\{\beta : \beta < \dfrac{\log\left(\frac{1-w_i}{w_i}\right) - \alpha}{x_i}\right\} \end{array}\right\}}$$

5. Comparaison des algorithmes

5.1. Critères de contrôle de convergence

Même si les méthodes MCMC sont applicables à une vaste classe de modèles, elles souffrent d'un problème pratique important : il s'agit de déterminer le moment où on peut conclure à leur convergence, autrement dit, stopper la chaîne et utiliser les observations pour l'estimation des caractéristiques de la distribution en considérant que l'échantillon est assez représentatif de la distribution stationnaire. Pour le moment, il n'existe pas véritablement de méthode efficace d'affronter ce problème de contrôle de convergence. De plus, ces diagnostics permettent de vérifier des conditions nécessaires, qui ne sont pas suffisantes pour « assurer» la convergence (voir Robert, 1996, Chap.8).

D'après Cowles et Carlin (1996) et Brooks et Roberts (1998), on peut distinguer trois degrés de convergence pour lesquels un contrôle est nécessaire. Le premier décide si les variables $\theta^{(t)}$ sont distribuées suivant la distribution stationnaire π. Pour le second type de convergence, il faut noter que même si $\theta^{(t)} \sim \pi$, l'exploration de la complexité de π par la chaîne $(\theta^{(t)})$ peut être plus ou moins longue. On doit donc s'assurer que la chaîne a bien mis à jour toutes les spécificités de π, comme l'ensemble des modes. Brooks et Roberts (1998) associent cette convergence à la vitesse de mélangeance de la chaîne, au sens vague d'une exploration plus ou moins rapide du support de π. Il s'agit en particulier de déterminer la valeur minimale de T autorisant l'approximation de $E_\pi[h(\theta)]$ par l'estimateur classique de Monte-Carlo soit

$$\delta_T = \frac{1}{T}\sum_{t=1}^{T} h(\theta^t). \tag{5}$$

Bien que (5) converge presque sûrement vers l'espérance a posteriori $E_\pi[h(\theta)]$ quand T tend vers $+\infty$, il est évidemment préférable de pouvoir contrôler la vitesse de convergence ou de manière équivalente la précision de l'approximation de $E_\pi[h(\theta)]$ par (5) au moyen d'un théorème de la limite centrale

$$\sqrt{T}(\delta_T - E_\pi[h(\theta)]) \approx \mathcal{N}(0, \sigma_h^2).$$

Les méthodes des chaînes en parallèle ou « batch sampling » s'inscrivent dans ce troisième type de convergence pour garantir la quasi-indépendance des variables simulées (voir Mengersen, Robert, et Guihenneuc-Jouyaux, 1999)

5.2. Application au modèle Logit

Dans le cadre du modèle Logit, nous avons généré des variables explicatives x_i suivant la loi normale $\mathcal{N}(0, 1)$ puis des variables y_i avec $\alpha = 3$, $\beta = -2$ et $n = 500$.

Le résultat de cette expérience est que la méthode de Damien et Walker demande beaucoup plus d'itérations que l'algorithme de Metropolis-Hastings pour explorer le

support et arriver à des propriétés de convergence proches de celles de l'algorithme de Metropolis-Hastings. Cette lenteur est, en partie, due au nombre de variables auxiliaires, dans le sens où plus la taille des données est grande, plus il y a de variables auxiliaires à simuler, ce qui diminue la vitesse de convergence et la rapidité d'exploration de la surface de la loi a posteriori. À l'inverse, l'algorithme de Metropolis-Hastings permet l'exploration rapide de l'ensemble des modes de la loi objective et une bonne vitesse de convergence vers la distribution stationnaire, avec, comparativement, moins d'itérations. Bien entendu, les deux algorithmes conduisent à la même loi a posteriori, comme le montre la figure 1.

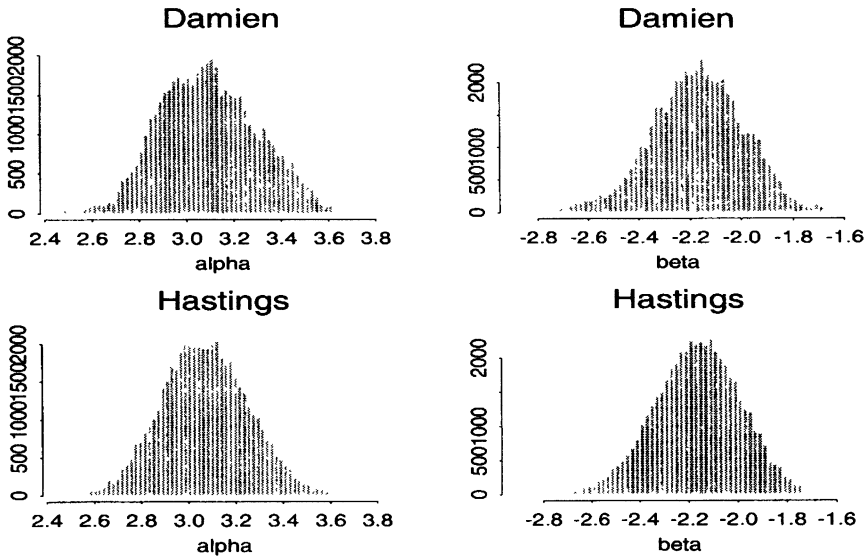

FIGURE 1

Comparaison d'histogrammes des paramètres α et β suivant la méthode de Damien et Walker (en haut) *avec l'algorithme de Metropolis-Hastings* (en bas).

La figure 3 montre on ne peut plus clairement les défauts de la méthode de Damien et Walker par rapport à l'algorithme de Metropolis-Hastings à loi instrumentale normale. Sur 200 itérations, les déplacements de la chaîne $(\alpha^{(t)}, \beta^{(t)})$ sur la surface de π ne recouvrent pas entièrement le support de la distribution. Au contraire, pour l'algorithme de Metropolis-Hastings, la distribution de la chaîne sur les niveaux de la loi a posteriori est satisfaisant et coincide plus nettement avec le support.

La convergence vers cette loi a posteriori, au sens de l'estimation, est également beaucoup plus lente pour la méthode de Damien et Walker, comme illustré par la figure 5 sur la convergence des moyennes empiriques de α et β, et la figure 2 sur la vitesse de variation des chaînes $\alpha^{(t)}$ et $\beta^{(t)}$.

Une comparaison des autocorrélations pour les paramètres α, β renforce ce constat de faible mélangeance et de convergence lente (voir figure 4). Au contraire,

l'algorithme de Metropolis-Hastings associé à la loi normale induit une faible autocorrélation, c'est-à-dire une vitesse de convergence rapide. Il semble évident que la très forte autocorrélation pour la méthode de Damien et Walker limite fortement le déplacement de la chaîne sur le contour de la loi a posteriori, et cela explique la nécessité d'un grand nombre d'itérations pour parvenir à la convergence.

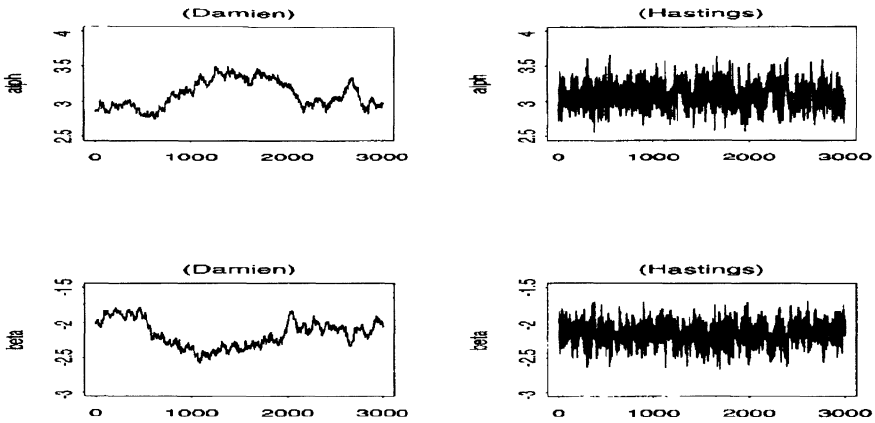

FIGURE 2

Comparaison des traces des paramètres α et β, suivant la méthode e Damien et Walker (à gauche) *et l'algorithme de Metropolis-Hastings* (à droite).

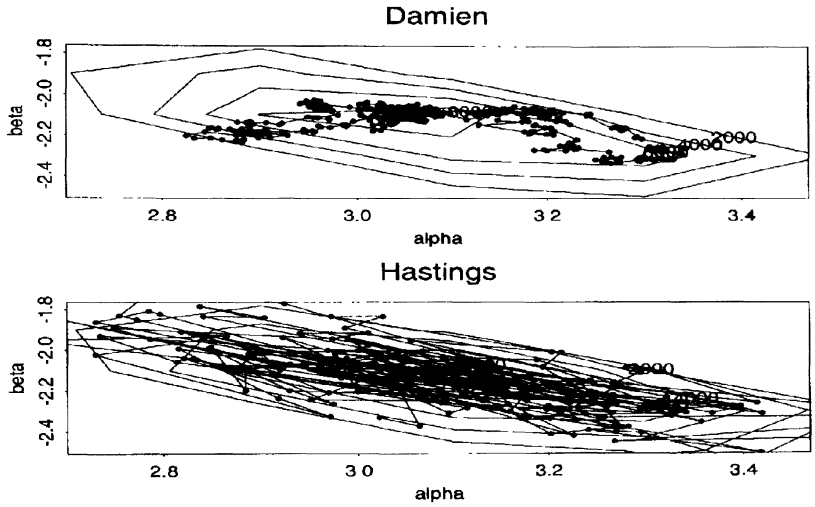

FIGURE 3

Comparaison du déplacement de la chaîne $(\alpha^{(t)}, \beta^{(t)})$ sur le contour de la loi a posteriori suivant la méthode pour 200 itérations de la chaîne.

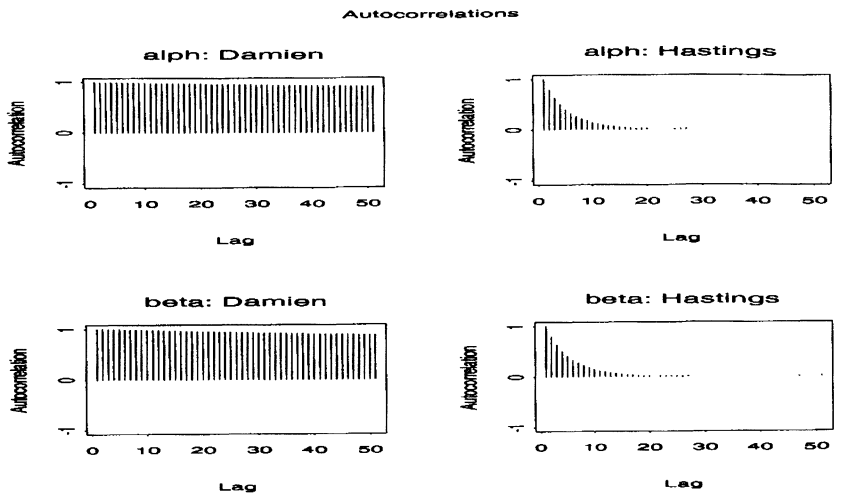

FIGURE 4
Comparaison des autocorrélations pour les paramètres α et β suivant la méthode de Damien et Walker (à gauche) avec l'algorithme de Metropolis-Hastings (à droite).

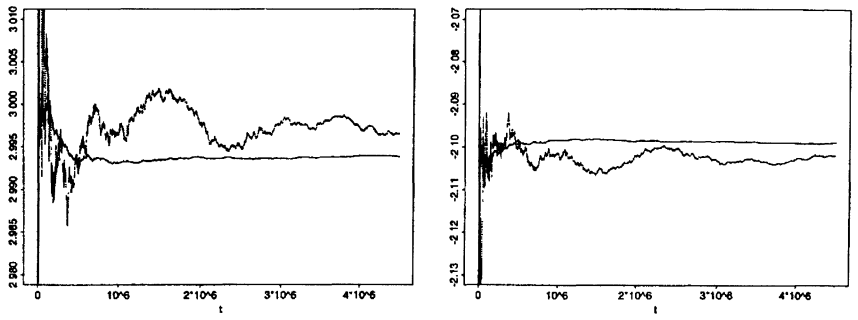

FIGURE 5
(à gauche) *Comparaison des approximations de $E[\alpha]$, suivant les deux méthodes : la méthode de Damien et Walker* (traits pointillés) *et l'algorithme de Metropolis-Hastings* (traits pleins). (à droite) *Comparaison des approximations de $E[\beta]$, pour la méthode de Damien et Walker* (traits pointillés) *et l'algorithme de Metropolis-Hastings* (traits pleins).

On peut également évaluer la convergence par des méthodes plus avancées, comme celle de Geweke (1992). La figure 6 montre que la plupart des valeurs de la statistique Z_n (qui doit être distribuée suivant $\mathcal{N}(0,1)$ en cas de convergence) pour la méthode de Damien et Walker se trouvent en dehors de l'intervalle ± 1.96 de la distribution normale centrée réduite, signifiant l'échec du test de convergence pour l'exemple traité. Par contre, pour l'algorithme de Metropolis-Hastings associé à la loi normale, la majorité des valeurs sont dans l'intervalle ± 1.96, ce qui démontre le fort potentiel de convergence de cette méthode.

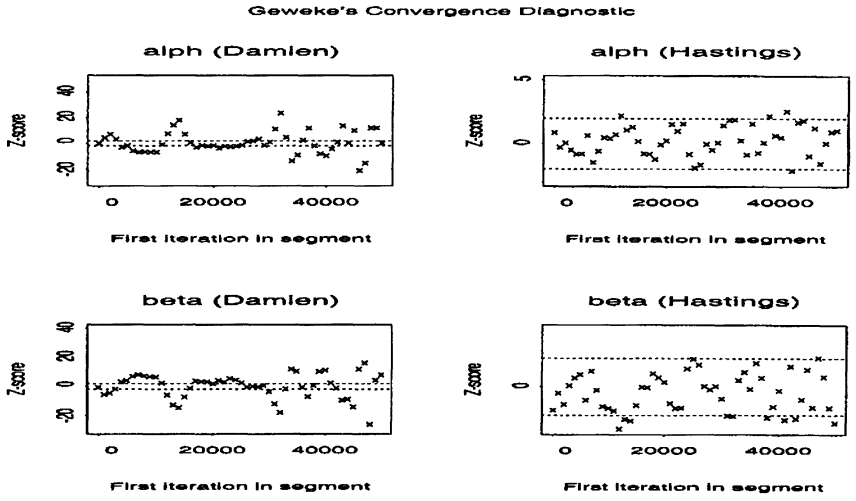

FIGURE 6
Comparaison du diagnostic de convergence de Geweke pour le modèle Logit suivant la méthode de Damien et Walker (à gauche) avec l'algorithme de Metropolis-Hastings (à droite).

De même, l'efficacité de l'algorithme de Metropolis-Hastings, associé à la loi normale, est confirmée par la technique de Raftery et Lewis (1992). Le tableau 1 donne l'évolution de k, pas minimum d'échantillonnage, t_0 nombre minimum d'itérations nécessaires pour atteindre la stationnarité, et T nombre total d'itérations assurant la convergence. En effet, le tableau 1 indique que pour la méthode de Damien et Walker, avec une chaîne test de 50 000 itérations, il faut au total 333 480 (resp. 240 448) itérations, dont 308 (resp. 240) itérations initiales à rejeter et un pas d'échantillonnage de 14 (resp. 16) pour le paramètre α (β, respectivement). Tandis que celle de Metropolis-Hastings avec une loi instrumentale normale nécessite un pas de 2 (resp. 4) avec une proportion de rejet de 17% (resp. 12%) de ce que rejette Damien et Walker, nécessitant un nombre total d'itérations de 17% (resp. 9.5%) de celui-ci et avec un minimum identique pour le paramètre α (β, respectivement). Ces valeurs sont obtenues par le logiciel `CODA` (voir Best *et al.* 1995).

TABLEAU 1
Tableau de diagnostic de Raftery et Lewis correspondant aux données simulées.

Méthode	Variable	Pas minimum	Nombre à rejeter	Nombre total	Nombre minimum	Facteur de dépendance (I)
Damien	α	14	308	333480	3746	89
	β	16	240	240448	3746	64.2
Metropolis-	α	2	18	19302	3746	5.15
Hastings	β	4	20	25292	3746	6.75

6. Conclusion

La comparaison entre la méthode de Damien et Walker et un algorithme arbitraire de Metropolis-Hastings semble a priori donner l'avantage à la méthode de Damien et Walker, puisque cette dernière tire ses distributions conditionnelles de la véritable loi π, tandis que l'algorithme de Metropolis-Hastings est fondé sur une loi instrumentale g qui est une approximation de π. En effet, la méthode de Damien et Walker est, par construction, plus directe que l'algorithme de Metropolis-Hastings, car elle n'est pas soumise à un « mauvais » choix de la loi instrumentale et évite les simulations inutiles (« rejets »). Mais la disponibilité et « l'objectivité » de la méthode de Damien et Walker ne sont pas nécessairement des arguments en sa faveur.

La simulation d'une seule composante à chaque itération pour l'algorithme de Gibbs limite fortement les déplacements possibles de la chaîne (y^t) et fait que les méthodes d'échantillonnage de Gibbs sont lentes à converger, car lentes à explorer la surface de f. Cette lenteur propre à l'échantillonnage de Gibbs conduit à une attraction forte vers le mode local le plus proche, ce qui induit des difficultés énormes à visiter l'ensemble des modes importants de f en cas de multimodalité.

Par contre, les défauts des algorithmes de Metropolis-Hastings sont d'une autre nature, car ils proviennent plus souvent d'un mauvais ajustement entre la loi objective f et la loi instrumentale g que d'une trop forte approximation entre les deux lois. De plus, la liberté donnée par les méthodes de Metropolis-Hastings permet parfois de remédier à ces défauts en augmentant certains paramètres de variation. L'inconvénient essentiel des algorithmes de Metropolis-Hastings par rapport à l'échantillonnage de Gibbs, est plutôt de ne pas toujours saisir les détails de la distribution f du fait d'une échelle de simulation peu exacte.

Dans le contexte de cet article, la méthode de Damien et Walker est employée en introduisant des variables auxiliaires pour obtenir un accès facile aux lois conditionnelles qui sont indispensables à l'application de l'algorithme de Gibbs. La complétion de $\pi(\alpha, \beta)$ en $\pi_1(\alpha, \beta, w)$ passe par la simulation de $\pi_1(\alpha, \beta, w)$ et le rejet des variables auxiliaires qui sont considérées, dans ce cas, comme des paramètres de nuisance. Ceci multiplie le temps de calcul et réduit la vitesse de convergence de l'échantillonnage de Gibbs de façon considérable, car ces variables auxiliaires sont de la même taille que les données.

Bibliographie

BESAG, J. et GREEN, P.J. (1993) Spatial Statistics and Bayesian computation (avec discussion). *Journal of the Royal Statistical Society* (Series B) **55**, 25-38.

BEST, N.G., COWLES, M.K. et VINES, K. (1995) CODA : Convergence diagnosis and output analysis software for Gibbs sampling output, Version 0.30. Tech. Report, MRC Biostatistics Unit, Univ. of Cambridge.

BROOKS, S.P et ROBERTS, G. (1998) Assessing convergence of Markov chain Monte Carlo algorithms. *Statistics and Computing* **8**, 319-335.

CARLIN, B.P. et CHIB, S. (1995) Bayesian model choice through Markov-Chain Monte Carlo. *Journal of the Royal Statistical Society* (Series B), **57**, 473-484.

COWLES, M.K. et CARLIN, B.P. (1996) Markov Chain Monte Carlo convergence diagnostics : a comparative study. *Journal of the American Statistical Association* **91**, 883–904.

DAMIEN, P. et WALKER, S. (1996) Sampling probability densities via uniform random variables and a Gibbs sampler. Tech. Report, Business School, University of Michigan.

DAMIEN, P., WAKEFIELD, J. et WALKER, S. (1999) Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society* (Series B) **61**, 331–344.

GELMAN, A., GILKS, W.R. and ROBERTS, G.O. (1996) Efficient Metropolis jumping rules. In *Bayesian Statistics 5*, J.O. Berger, J.M. Bernardo, A.P. Dawid, D.V. Lindley and A.F.M. Smith (Eds.). 599–608. Oxford University Press, Oxford.

GEWEKE, J. (1992) Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments (with discussion). In *Bayesian Statistics 4*, J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (Eds.). 169–193. Oxford University Press, Oxford.

GILKS, W.R. et ROBERTS, G.O. (1996) Strategies for improving MCMC. In *Markov Chain Monte Carlo in Pratice*, W.R. Gilks, S. Richardson and D.J. Spiegelhalter (Eds.), 89-114. Chapman and Hall, London.

HAMMERSLEY, J.M. et HANCOMB, D.C. (1964) *Monte Carlo Methods*, J. Wiley, New York.

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

MENGERSEN, K.L., ROBERT, C.P. et GUIHENNEUC-JOUYAUX, C. (1999) MCMC convergence diagnostics : a « reviewww» (avec discussion). In *Bayesian Statistics 6*. J.O. Berger, J.M. Bernardo, A.P. Dawid, D.V. Lindley and A.F.M. Smith (Eds.). Oxford University Press, Oxford, 415-441.

RAFTERY, A.E. et LEWIS, S. (1992) How many iterations in the Gibbs sampler? In *Bayesian Statistics 4*, J.O. Berger, J.M. Bernardo, A.P. Dawid and A.F.M. Smith (Eds.), 763-773. Oxford University Press, Oxford.

ROBERT, C.P. (1992) *L'Analyse Statistique Bayésienne.* Economica, Paris.

ROBERT, C.P. (1996) *Méthodes de Monte-Carlo par Chaînes de Markov*. Economica, Paris.

SMITH, A.F.M. et ROBERTS, G.O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (avec discussion). *Journal of the Royal Statistical Society* (Series B) **55**, 3-24.