L'Actualité économique

Échantillonnage de Gibbs et autres applications économétriques des chaînes markoviennes
Gibbs Sampling and other Applications of Markov Chains in Econometrics

Stephen Gordon and Gilles Bélanger

Volume 72, Number 1, mars 1996

URI: https://id.erudit.org/iderudit/602194ar
DOI: https://doi.org/10.7202/602194ar

See table of contents

Publisher(s)

HEC Montréal

ISSN

0001-771X (print)
1710-3991 (digital)

Explore this journal

Cite this article

Gordon, S. & Bélanger, G. (1996). Échantillonnage de Gibbs et autres applications économétriques des chaînes markoviennes. *L'Actualité économique*, *72*(1), 27–49. https://doi.org/10.7202/602194ar

Article abstract

This survey provides an introduction to Markov Chain Monte Carlo (MCMC) sampling techniques and to their applications to Bayesian econometrics. In describing the Gibbs sampler and the Metropolis-Hastings algorithm, the emphasis is put on how these techniques can be put into practice; the theoretical foundations are outlined using the elementary properties of Markov chains. To illustrate the potential of MCMC techniques, we decribe several examples where their application has produced clear gains over classical methods of inference.

L'Actualité économique, Revue d'analyse économique, vol. 72, nº 1, mars 1996

*Échantillonnage de Gibbs et autres applications économétriques des chaînes markoviennes**

Stephen GORDON
Département d'économique
Université Laval
Gilles BÉLANGER
Département de sciences économiques
Université de Montréal

RÉSUMÉ – Ce survol fournit une introduction aux techniques d'échantillonnage de type *Markov Chain Monte Carlo* (MCMC) et leurs applications à l'économétrie bayesienne. Par ce survol notre but n'est pas d'expliquer les fondements théoriques derrière les méthodes de type MCMC, mais bien de faire un exposé pratique des techniques qui s'y rapportent. Nous chercherons surtout à mettre en valeur la facilité et l'étendue des applications par l'utilisation d'exemples simples.

ABSTRACT – *Gibbs Sampling and other Applications of Markov Chains in Econometrics.* This survey provides an introduction to Markov Chain Monte Carlo (MCMC) sampling techniques and to their applications to Bayesian econometrics. In describing the Gibbs sampler and the Metropolis-Hastings algorithm, the emphasis is put on how these techniques can be put into practice; the theoretical foundations are outlined using the elementary properties of Markov chains. To illustrate the potential of MCMC techniques, we decribe several examples where their application has produced clear gains over classical methods of inference.

INTRODUCTION

Les méthodes d'échantillonnage utilisant des chaînes markoviennes ont connu une forte effervescence au cours des dernières années. Les développements rapides de l'informatique ont favorisé et même rendu possible l'essor des techniques *Markov Chain Monte Carlo* (MCMC). Cet état de fait a permis à l'estimation de type bayesienne de se mettre en valeur par la soudaine facilité de ses applications. Grâce à ces méthodes, l'estimation bayesienne est non seulement devenue plus simple qu'avant, mais dans bien des cas, plus facile que sa contrepartie classique.

* Nous tenons à remercier Guy Lacroix, Denis Bolduc ainsi qu'un arbitre anonyme pour leurs commentaires judicieux.

Du point de vue pratique, les difficultés posées par une application de la méthodologie bayesienne sont liées aux intégrales qu'elles nécessitent. Supposons qu'on s'intéresse à l'espérance *a posteriori* d'un paramètre quelconque. L'espérance étant définie par une intégrale, il se peut fort bien que cette dernière n'ait pas de solution analytique. Mais s'il était possible de simuler une séquence d'aléas tirée de la loi *a posteriori,* un estimateur convergent de l'espérance serait simplement la moyenne de l'échantillon artificiel. En ce sens, les techniques d'échantillonnage sont très utiles ; elles servent principalement à simuler des intégrations impossibles à résoudre analytiquement. Sous certaines conditions de régularité, les propriétés d'une chaîne markovienne assurent que ce résultat tient *même si les réalisations sont corrélées.* Heureusement, il est particulièrement facile de simuler des chaînes markoviennes satisfaisant ces conditions de régularité.

Par ce survol notre but n'est pas d'expliquer les fondements théoriques derrière les méthodes de type MCMC, mais bien de faire un exposé pratique des techniques qui s'y rapportent. Nous chercherons surtout à mettre en valeur la facilité et l'étendue des applications, plutôt que de participer à un débat théorique sur les approches classique et bayesienne. Ces discussions théoriques sont exposées dans certains des articles fournis dans la bibliographie qui devraient être consultés pour les justifications théoriques et les postulats de l'approche bayesienne.

1. SURVOL DES MÉTHODES BAYESIENNES

Il existe une vaste littérature sur les fondements philosophiques de l'approche bayesienne à l'inférence statistique[1]. Notre objectif dans cette section est de résumer la méthodologie afin de permettre au lecteur d'apprécier l'utilité des techniques décrites dans ce survol.

Écrivons le modèle économétrique par la densité conditionnelle $p(x|\theta)$, où x représente les données et où θ est le vecteur des paramètres. Du point de vue de l'analyste, les données x sont observées, mais il ne connaît pas θ; la distribution d'intérêt est donc la densité conditionnelle $p(\theta|x)$. Cette densité est dérivée à l'aide du théorème de Bayes :

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \equiv \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d(\theta)} \quad , \tag{1}$$

où $p(\theta)$ est la distribution marginale de θ. Puisque cette distribution représente l'incertitude sur la valeur de θ avant d'avoir observé les données, la distribution marginale $p(\theta)$ est connue comme la distribution *a priori* de θ. La distribution conditionnelle $p(\theta|x)$ est connue comme la distribution *a posteriori* de θ, parce qu'elle décrit l'incertitude de l'analyste après avoir observé les données.

1. Voir Zellner (1971), Leamer (1978), Poirier (1988) ou Bernardo et Smith (1994) pour une présentation détaillée.

Même si la distribution $p(\theta|x)$ incorpore toute l'information disponible sur θ, presque toutes les études empiriques se limitent à quelques caractéristiques clés de cette distribution. Dans ce qui suit, nous supposons que l'analyste s'intéresse à $E[g(\theta)|x]$; cette notation est assez générale pour inclure la moyenne *a posteriori* $E[\theta|x]$ (où $g(\theta)=\theta$), la matrice de covariances $V[\theta|x]$ (où $g(\theta)=(\theta-E[\theta|x])^2$) ou la probabilité $p[\theta\in A|x]$ (où $g(\theta)=\mathbb{I}_{[\theta\in A]}$, la fonction indicatrice de $\theta\in A$).

L'espérance $E[g(\theta)|x]$ est donnée par :

$$E[g(\theta)|x]=\int g(\theta)p(\theta|x)d\theta\equiv\frac{\int g(\theta)p(x|\theta)p(\theta)d\theta}{\int p(x|\theta)p(\theta)d\theta} \tag{2}$$

Il est aussi possible d'appliquer cette approche au problème de sélection de modèle. Posons M^m l'indicatrice associée au modèle m= 1, 2, ..., $K(\equiv\mathbb{I}_{[\text{modèle }\equiv m]})$, où K est le nombre de modèles considérés. À chaque modèle m est associé un vecteur de paramètres θ^m. On peut représenter un modèle économétrique donné par sa densité conditionnelle $p(x|\theta^m,M^m)$. Après avoir observé les données x, l'analyste s'intéresse à $p(M^m|x)$, la probabilité *a posteriori* que le modèle soit valide. Elle est également dérivée en utilisant la règle de Bayes :

$$p(M^m|x)=\frac{p(x|M^m)p(M^m)}{\sum_{j=1}^{K}p(x|M^j)p(M^j)}\equiv\frac{\int p(x|\theta^m,M^m)p(\theta^m|M^m)d\theta^m p(M^m)}{\sum_{k=1}^{K}\int p(x|\theta^k,M^k)p(\theta^k|M^k)d\theta^k p(M^k)} \tag{3}$$

où $p(M^m)$ est la probabilité *a priori* que le modèle m est correct, et $p(\theta^m|M^m)$ est la densité *a priori* des paramètres du modèle m. Notons que l'expression $p(x|M^m)$ est simplement l'intégrale dans le dénominateur de (1). Puisque la densité $p(x|\theta^m,M^m)$ peut être interprétée comme la vraisemblance $L(\theta^m|x,M^m)$, et puisque l'on intègre cette vraisemblance par rapport à la densité marginale $p(\theta^m|M^m)$, on appelle $p(x|M^m)$ la *vraisemblance marginale* du modèle m.

Dans une application de la méthodologie bayesienne, l'analyste doit spécifier explicitement ses croyances *a priori* et il lui faut trouver une façon pour calculer - ou éviter de calculer - les intégrales dans (2) et (3). Comme on le voit dans (1) et (3), les croyances *a priori* jouent un rôle clé dans une analyse bayesienne. L'information *a priori* peut provenir d'études précédentes utilisant des données semblables, elle peut refléter des conditions de régularité imposées par la théorie économique ou tout simplement refléter les croyances subjectives de l'analyste. Si l'analyste veut utiliser une loi *a priori* « non informative » pour refléter l'ignorance totale, il peut adopter une loi diffuse avec une variance très élevée ou même une loi « impropre » (c'est-à-dire, une densité qui ne s'intègre pas à 1) comme la loi $p(\theta)=\kappa$, une constante. Alors la loi *a posteriori* aura la même forme que la fonction de vraisemblance. Il y a d'ailleurs une importante littérature sur la meilleure façon de caractériser l'ignorance, et il peut arriver que le choix d'une loi *a priori* « non informative » ait des conséquences importantes (Phillips, 1991). Pour un traitement plus détaillé, voir Bernardo et Smith (1994).

Du point de vue pratique, les difficultés auxquelles l'analyste doit faire face sont posées par les intégrales dans (1), (2) et (3); les expressions analytiques pour ces intégrales existent seulement pour quelques cas spéciaux, comme le modèle linéaire normal. Pour les modèles plus complexes, ces intégrales doivent être calculées en utilisant des techniques numériques; si la dimension de θ est trop élevée, cette tâche devient plutôt laborieuse. La popularité de la méthodologie classique dans les dernières décennies pourrait être attribuée en partie au fait qu'elle était plus simple à utiliser dans le cadre de modèles complexes.

2. LES APPLICATIONS DES MÉTHODES D'ÉCHANTILLONNAGE

La principale contribution des techniques d'échantillonnage basées sur des chaînes markoviennes est de faciliter les applications empiriques de la méthodologie bayesienne, à un point tel qu'elles deviennent souvent plus simples que leur équivalent classique.

L'idée de base des techniques d'échantillonnage est bien connue. Supposons qu'une variable aléatoire z peut être décrite par la densité $p(z)$, et supposons que nous voulons calculer l'espérance $E[g(z)]=\int g(z)p(z)dz$. Si une expression analytique pour cette intégrale n'existe pas, et si z est un vecteur de dimension assez élevée, calculer cette intégrale devient une tâche ardue. Supposons maintenant qu'il soit possible de simuler une séquence de N réalisations tirée de la loi $p(z)$. Étant donné l'échantillon $\{z^i\}_{i=1}^N$, on note que la statistique $N^{-1}\sum_{i=1}^N g(z^i)$ est un estimateur convergent de l'espérance $E[g(z)]$ si certaines conditions de régularité sont satisfaites. Il est important de noter que la valeur de N est choisie par l'analyste pour s'assurer du niveau de précision désiré[2]. Notons aussi que la faisabilité de cette approche n'est pas affectée par la dimension de z s'il est possible de tirer des aléas de la loi $p(z)$.

2.1.1 *Techniques basées sur des tirages indépendants*

Il est utile de commencer avec le cas le plus simple, où il est possible de produire des tirages aléatoires de θ de la loi *a posteriori* $p(\theta|x)$ en utilisant les algorithmes standard. Supposons que l'analyste veut calculer $E[g(\theta)|x]$, et que l'intégrale (2) n'a pas de solution analytique. Dans ce cas, l'analyste peut adopter la procédure mentionnée ci-haut. Après avoir simulé une séquence de N observations indépendantes $\{\theta^i\}$ de la loi *a posteriori* $p(\theta|x)$, la statistique $N^{-1}\sum_{i=1}^N g(\theta^i)$ converge presque sûrement vers $E[g(\theta)|x]$.

La technique d'intégration Monte Carlo est une méthode d'échantillonnage plus générale. Supposons qu'il soit plus facile de générer des variables aléatoires à partir d'une loi de densité $I(\theta)$ qu'une de densité $p(\theta|x)$. Nous définissons $w(\theta)\equiv p(\theta|x)/I(\theta)$. Étant donné un échantillon $\{\theta^i\}_{i=1}^N$ tiré de $I(\theta)$, on peut montrer

2. Ce point est discuté en détail plus tard.

facilement[3] que la statistique $\bar{g} \equiv N^{-1}\sum_{i=1}^{N} g(\theta^i)w(\theta^i)$ converge presque sûrement vers $E[g(\theta)|x]$. Geweke (1989) montre comment l'analyste peut évaluer la précision de cet estimateur afin de choisir N.

En principe cette méthode est très générale, mais en pratique la vitesse de convergence de l'estimateur $\bar{g}$ est très sensible au choix de $I(\theta)$, la densité d'importance (*importance function*). Geweke (1989) note que la convergence est plus rapide si $I(\theta)$ est de forme similaire à $p(\theta|x)$ et si $I(\theta)>p(\theta|x)$ pour les valeurs extrêmes de θ. Pour les modèles complexes, il est parfois très difficile d'identifier une fonction d'importance efficace.

2.1.2 *Techniques basées sur les chaînes markoviennes*

Il est possible de générer un échantillon artificiel d'observations indépendantes de $p(\theta|x)$ seulement pour quelques cas spéciaux, comme la loi normale. Néanmoins, il est souvent très facile de construire un algorithme pour simuler des observations corrélées. Si certaines conditions de régularité sont satisfaites, la moyenne d'un échantillon des observations corrélées est également un estimateur convergent de la moyenne de la population.

Les algorithmes de ce type se basent sur la théorie des *chaînes markoviennes*. Une séquence de variables aléatoires $\{z^i\}$ est une chaîne markovienne de premier ordre si la distribution conditionnelle $p(z^i|z^{i-1},z^{i-2}, ...)$ est une fonction de z^{i-1} seulement ($=p(z^i|z^{i-1})$); elle est une chaîne markovienne d'ordre p si cette distribution est une fonction des p observations précédentes.

Remarquons qu'un processus autorégressif d'ordre p est un exemple d'une chaîne markovienne d'ordre p. Il est bien connu que la moyenne d'un échantillon de N observations d'un processus AR(p) est un estimateur convergent de l'espérance inconditionnelle pour le processus seulement si le processus est stationnaire. Dans une analyse de la convergence des chaînes markoviennes, on se sert d'un concept semblable, c'est-à-dire, *l'ergodicité*. Une séquence est dite ergodique s'il est possible de passer de n'importe quelle réalisation a dans la chaîne à n'importe quelle réalisation b, même si la transition de $z=a$ à $z=b$ nécessite de passer par quelques réalisations intermédiaires ($a \rightarrow c \rightarrow b$).

Considérons une chaîne markovienne de premier ordre ayant une densité $p(z^i|z^{i-1})$, et supposons que la valeur de z^0 est connue. On note que la séquence des distributions conditionnelles $p(z^i|z^0)$ peut être décrite par

$$
\begin{aligned}
&p(z^1|z^0) \\
&p(z^2|z^0) = \int p(z^2,z^1|z^0)dz^1 && = \int p(z^2|z^1)p(z^1|z^0)dz^1 \\
&p(z^3|z^0) = \int\int p(z^3,z^2,z^1|z^0)dz^2dz^1 && = \int\int p(z^3|z^2)p(z^2|z^1)p(z^1|z^0)dz^2dz^1 \\
&\quad\vdots && \quad\vdots \\
&p(z^i|z^0) = \int ...\int p(z^i,z^{i-1},...,z^1|z^0)dz^{i-1}...dz^1 && = \int ...\int \prod_{j=1}^{t} p(z^j|z^{j-1})dz^{i-1}...dz^1 \quad .
\end{aligned}
$$

3. Voir Kloek et van Dijk (1978) et Geweke (1989).

Si la chaîne markovienne est ergodique on peut montrer[4] que la séquence des lois conditionnelles $\{p(z^i|z^0)\}_{i=1}^{\infty}$ converge vers la loi marginale $p(z)$ pour n'importe quelle valeur z^0. Autrement dit, la dépendance de la chaîne face à z^0 s'estompe progressivement. Supposons que cette convergence se produise avant la réalisation n. La séquence des N réalisations suivantes $\{z^i\}_{i=n+1}^{n+N}$ constitue donc un échantillon d'observations corrélées ayant la même densité inconditionnelle; la moyenne de la statistique $N^{-1}\sum_{i=n+1}^{n+N} g(z^i)$ converge presque sûrement vers $E[g(z)]$.

Exemple 1 : La convergence des chaînes markoviennes

Soit un processus stochastique $z^i \in \{0,1\}$ avec les probabilités conditionnelles $p(z^i = 1 | z^{i-1}=1)=0,5$ et $p(z^i=1|z^{i-1}=0)=0,25$.

Supposons que $z^0=1$. Les probabilités $p(z^i=1)$ sont calculées à partir de la formule récursive $p(z^i=1|z^{i-1}=1)p(z^{i-1}=1|z^0=1) + p(z^i=1|z^{i-1}=0)p(z^{i-1}=0|z^0=1)$:

$$p(z^1=1|z^0=1) = (0,5)(1) + (0,25)(0) = 0,5$$

$$p(z^2=1|z^0=1) = (0,5)(0,5) + (0,25)(0,5) = 0,375$$

$$p(z^3=1|z^0=1) = (0,5)(0,375) + (0,25)(0,625) = 0,3438$$

$$p(z^4=1|z^0=1) = (0,5)(0,3438) + (0,25)(0,6562) = 0,3359$$

$$p(z^5=1|z^0=1) = (0,5)(0,3359) + (0,25)(0,6641) = 0,3340$$

$$\vdots \qquad\qquad \vdots$$

On note que la probabilité $p(z^i=1|z^0=1)$ tend vers 1/3. Supposons maintenant que $z^0=0$. En appliquant
$p(z^i=1|z^{i-1}=1)p(z^{i-1}=1|z^0=0) + p(z^i=1|z^{i-1}=0)p(z^{i-1}=0|z^0=0)$ on obtient :

$$p(z^1=1|z^0=0) = (0,5)(0) + (0,25)(1) = 0,25$$

$$p(z^2=1|z^0=0) = (0,5)(0,25) + (0,25)(0,75) = 0,3125$$

$$p(z^3=1|z^0=0) = (0,5)(0,3125) + (0,25)(0,6875) = 0,3281$$

$$p(z^4=1|z^0=0) = (0,5)(0,3281) + (0,25)(0,6719) = 0,3320$$

$$p(z^5=1|z^0=0) = (0,5)(0,3320) + (0,25)(0,6680) = 0,3330$$

$$\vdots \qquad\qquad \vdots$$

On note encore que $p(z^{i-1}=1|z^0=0)$ tend également vers 1/3. Lorsque le nombre d'itérations augmente, la sensibilité de la distribution de z^i à la valeur de départ diminue.

---

4. Voir Stokey et Lucas (1989).

Les techniques d'échantillonnage décrites plus loin utilisent cette notion de convergence des chaînes markoviennes. Dans chaque cas, on considère un algorithme pour simuler les tirages artificiels des paramètres à partir d'une chaîne markovienne ergodique dont la distribution marginale est la loi *a posteriori* $p(\theta|x)$.

2.2 *Échantillonnage de Gibbs*

L'échantillonneur de Gibbs est la technique MCMC la plus simple. Sa popularité date de l'application de Geman et Geman (1984). L'expression « échantillonneur de Gibbs » vient de l'utilisation que Geman et Geman ont fait de la distribution Gibbs pour modéliser les images satellites, mais son applicabilité est beaucoup plus générale.

Écrivons le vecteur de paramètres par $\theta=(\theta_1,\theta_2,\ldots,\theta_J)'$, où θ_j peut être un élément ou un sous-ensemble de θ. Si le modèle est assez complexe, la distribution marginale $p(\theta_j|x)$ sera non standard. Par contre, la distribution conditionnelle $p(\theta_j|\theta_{-j},x)$ sera souvent standard, où θ_{-j} représente les paramètres du modèle autres que θ_j. Supposons qu'il est possible de simuler des tirages artificiels $\theta_j \sim p(\theta_j|\theta_{-j},x)$. Si tel est le cas, on peut générer une séquence aléatoire selon l'algorithme suivant :

$$
\left.
\begin{array}{l}
\theta_1^i \sim p(\theta_1|\theta_2^{i-1},\theta_3^{i-1},\ldots,\theta_J^{i-1},x) \text{ } \} \text{ une itération} \\
\theta_2^i \sim p(\theta_2|\theta_1^{i},\theta_3^{i-1},\ldots,\theta_J^{i-1},x) \\
\theta_3^i \sim p(\theta_3|\theta_1^{i},\theta_2^{i},\ldots,\theta_J^{i-1},x) \\
\quad \vdots \\
\theta_J^i \sim p(\theta_J|\theta_1^{i},\theta_2^{i},\ldots,\theta_{J-1}^{i},x)
\end{array}
\right\} \text{ un tour} \quad .
$$

Notons que cet algorithme décrit une chaîne markovienne de premier ordre, puisque la distribution conditionnelle d'un tirage dépend de la réalisation précédente. De plus, si la densité $p(\theta_j|\theta_{-j},x)$ est positive pour toutes les valeurs possibles de θ_{-j}, la chaîne est aussi ergodique[5]. Dans ce cas, la chaîne markovienne converge vers sa distribution stable $p(\theta|x)$. Si la séquence converge avant l'itération n, la statistique $N^{-1}\sum_{i=n+1}^{n+N} g(\theta^i)$ converge presque sûrement vers $E[g(\theta)|x]$.

En principe, les valeurs générées à chaque itération peuvent être utilisées pour estimer $E[g(\theta)|x]$. Mais en pratique, le taux de convergence peut être plutôt lent si la corrélation entre θ^i et θ^{i-1} est trop élevée. Puisque la même valeur de θ_j est retenue pour J-1 itérations, la corrélation entre les valeurs de θ simulées par l'algorithme précédent sera grande. Dans les applications de l'échantillonneur de Gibbs, on utilise des valeurs de θ générées à chaques J itérations - un *tour* - pour que chaque tirage soit différent.

---

5. Cette condition est suffisante, mais non nécessaire. Voir Roberts et Smith (1994) pour une analyse des conditions nécessaires pour la convergence des chaînes markoviennes.

Exemple 2 : L'estimation d'un modèle SUR

Soit un modèle multivarié

$$y_t = X_t\beta + u_t \qquad t=1,\ldots,T$$

où :

y_t est un vecteur de dimension l

X_t est une matrice de dimension lxk

β est un vecteur de paramètres de dimension k

u_t est un vecteur aléatoire de dimension l, iid $N(0,\Sigma)$.

Malgré la structure simple de ce modèle, il est difficile d'en faire l'inférence avec un échantillon fini si on utilise les techniques classiques ; les principaux résultats se fondent sur la théorie asymptotique. Ce modèle pose également des problèmes pour l'analyse bayesienne : les distributions *a posteriori* de β et Σ sont typiquement non standard, même si le modèle est simple.

Mais avec l'application de l'échantillonnage de Gibbs, l'estimation du modèle est énormément simplifiée. Si la valeur de Σ est connue, on peut transformer le modèle selon

$$C'y_t = C'X_t + C'u_t$$
$$\tilde{y}_t = \tilde{X}_t + \varepsilon_t$$

où C est la racine Cholesky de Σ^{-1}. Écrivons le modèle comme $\tilde{y}=\tilde{X}\beta+\varepsilon$, où $\tilde{y} = [\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_T]'$, $\tilde{X} = [\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_T]'$ et où $\varepsilon \sim N(0, I_{lT})$. Si les croyances *a priori* de β sont une loi normale multivariée $p(\beta)=N(\hat{\beta},\hat{A}^{-1})$ il a été démontré[6] que la loi *a posteriori* $p(\beta|\Sigma,X,y)$ est une loi normale :

$$p(\beta|\Sigma,X,y)=N(\overline{\beta}, \overline{A}^{-1})$$

où : $\overline{A} = \tilde{X}'\tilde{X} + \hat{A}$

$$\overline{\beta} = \overline{A}^{-1}(\tilde{X}'\tilde{y} + \hat{A}\hat{\beta}) \quad .$$

De la même façon, si β est connu, on peut calculer $u_t=y_t - X_t\beta$. Si on utilise la loi *a priori* $p(\Sigma^{-1})=\text{Wishart}(\hat{v}, \hat{V})$ [7], la loi *a posteriori* $p(\Sigma|\beta,X,y)$ est également une loi Wishart inverse :

$$p(\Sigma^{-1}|\beta,X,y)=\text{Wishart}(\overline{v}, \overline{V})$$

où $\overline{v} = \hat{v} + T$

$$\overline{V} = \hat{V} + \sum_{t=1}^{T} u_t u_t' \quad .$$

6. Voir Zellner (1971).

7. La loi Wishart est l'extension matricielle d'une loi chi-carré.

Même si la loi conjointe $p(\beta,\Sigma|X,y)$ est non standard, il est facile de produire des tirages artificiels des lois conditionnelles $p(\beta|\Sigma,X,y)$ et $p(\Sigma|\beta,X,y)$. La séquence $\{(\beta^i,\Sigma^i)\}$ est générée selon :

$$\begin{aligned}
&\Sigma^1 \sim p(\Sigma|\beta^0,X,y)\\
&\beta^1 \sim p(\beta|\Sigma^1,X,y)\\
&\Sigma^2 \sim p(\Sigma|\beta^1,X,y)\\
&\beta^2 \sim p(\beta|\Sigma^2,X,y)\\
&\quad\vdots\\
&\Sigma^i \sim p(\Sigma|\beta^{i-1},X,y)\\
&\beta^i \sim p(\beta|\Sigma^i,X,y)\\
&\quad\vdots
\end{aligned}$$

Pour calculer $E[g(\beta,\Sigma)|x]$, on se sert de l'échantillon artificiel, $\{(\beta^i,\Sigma^i)\}_{i=1}^N$ tel que généré ci-dessus.

Comme on le voit dans cet exemple et dans les exemples décrits dans la section 4, l'échantillonnage de Gibbs nous permet d'exploiter la structure récursive d'un modèle pour l'estimer, plutôt que d'être forcé d'estimer tous les paramètres simultanément.

Cet algorithme peut être appliqué même si la distribution $p(\theta_j|\theta_{-j},x)$ n'est pas standard. Ritter et Tanner (1992) démontrent comment utiliser une technique d'approximation par des grilles (*Griddy-Gibbs Sampler*). Les étapes de cette technique étant :

1. Évaluer $p(\theta_j|\theta_{-j},x)$ aux valeurs $\theta_j=a_1, a_2, \ldots, a_D$ pour obtenir $w_1, w_2, \ldots, w_D$, où D représente le nombre d'éléments formant la grille (la précision augmente avec D).

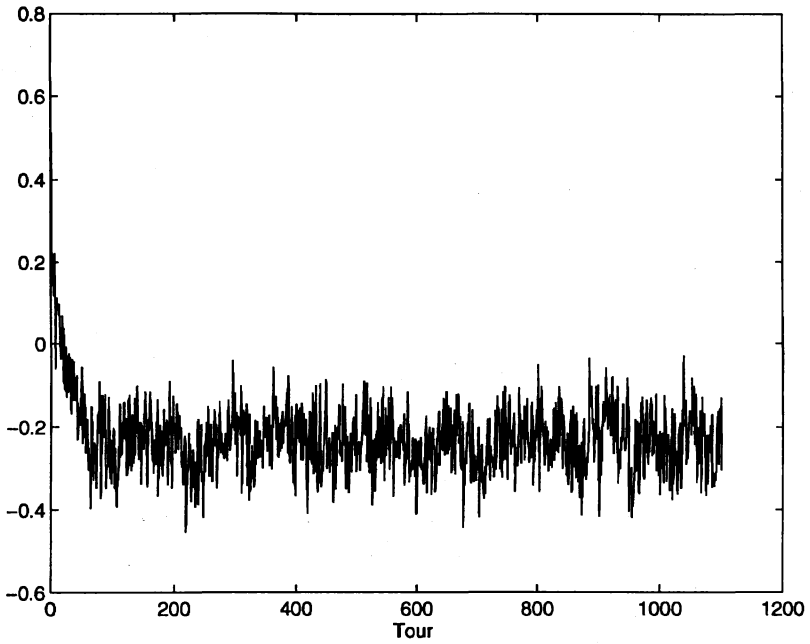
2. Utiliser $w_1, w_2, \ldots, w_D$ pour générer une approximation linéaire par morceaux de la fonction de répartition de $p(\theta_j|\theta_{-j},x)$.

3. Simuler un aléa d'une loi uniforme(0,1) et utiliser cette valeur pour générer une valeur de θ_j en utilisant l'inverse de la fonction de répartition calculée.

On note cependant que cette forme d'approximation est plutôt lente en pratique. On lui préférera généralement les méthodes issues de l'algorithme Metropolis-Hastings.

FIGURE 1

OUTPUT D'UNE APPLICATION DE L'ÉCHANTILLONAGE DE GIBBS



2.3 *L'algorithme Metropolis-Hastings*

Un autre algorithme pouvant être utilisé pour simuler des tirages d'une loi non standard est celui développé par Metropolis *et al* (1956), et Hastings (1970). Un traitement intuitif des fondements théoriques de l'algorithme Metropolis-Hastings est résumé par Chib et Greenberg (1995).

2.3.1 *L'algorithme général*

La base de cet algorithme est une procédure d'acceptation-rejet. Une loi connue - la *distribution génératrice de candidats* (DGC) - est utilisée pour simuler un nouveau candidat pour la chaîne. Ensuite, une règle probabilistique est utilisée pour décider si le candidat devrait être accepté ou non. S'il est accepté, la prochaine valeur de la chaîne prendra la valeur du candidat ; sinon, la chaîne reste à sa valeur actuelle.

Écrivons la DGC par $q(\theta,\theta')$. Si la dernière valeur de la chaîne est égale à θ, le candidat θ' est simulé à partir de la loi $q(\theta,\theta')$. Ce candidat est ensuite accepté avec la probabilité (θ,θ'), où :

$$\alpha(\theta,\theta') = \min\left\{\frac{p(\theta'|x)q(\theta',\theta)}{p(\theta|x)q(\theta,\theta')},1\right\} \quad . \tag{4}$$

Après avoir simulé la valeur du candidat θ', (4) est utilisée pour calculer la probabilité qu'il sera accepté. Une façon simple d'appliquer cette règle probabilistique est simplement de simuler une réalisation U d'une loi uniforme (0,1); le candidat est accepté si $U<\alpha$. Bref, on peut résumer l'algorithme pour simuler la réalisation θ^i de la chaîne markovienne par les étapes suivantes :

1. Étant donné la valeur précédente θ^{i-1}, simuler le candidat θ' à partir de la loi $q(\theta^{i-1},\theta')$.
2. Calculer $\alpha(\theta^{i-1},\theta')$.
3. Simuler U à partir d'une loi uniforme(0,1). Si $U<\alpha$, fixer θ^i à θ'; sinon, fixer θ^i à θ^{i-1}.

On peut montrer[8] que la chaîne markovienne décrite par cet algorithme est ergodique et que sa loi stable est $p(\theta|x)$. Remarquons qu'il n'est pas nécessaire de connaître la valeur de la constante normalisante de la distribution $p(\theta|x)$ dans (1); les termes n'étant pas fonction de θ s'annulent dans (4).

2.3.2 *Le choix d'une DGC*

Il reste à déterminer la forme pour la DGC $q(\theta,\theta')$. Un cas spécial de l'algorithme suggéré par Metropolis *et al.* (1953) est celui où $q(\theta,\theta')=f(\theta'-\theta)$, $f(\bullet)$ étant une densité connue. Dans ce cas, le candidat est simulé à partir d'une marche aléatoire : $\theta'=\theta+z$, où $z\sim f(z)$. Si les innovations de la marche aléatoire viennent d'une distribution symétrique, on note que $f(z)= f(-z)$, alors $q(\theta,\theta')= f(\theta'-\theta)= f(\theta-\theta')=q(\theta',\theta)$. Si c'est le cas, on note que l'expression $q(\theta',\theta)/q(\theta,\theta')$ dans (4) s'annule et que la probabilité d'accepter le candidat devient simplement

$$\alpha(\theta,\theta') = \min\left\{\frac{p(\theta'|x)}{p(\theta|x)},1\right\} \quad . \tag{5}$$

En d'autres termes, si la valeur de la densité « cible » $p(\theta|x)$ est plus élevée avec θ' qu'avec θ, la valeur de l'itération précédente, alors on accepte le candidat. Par contre, si la valeur de la densité diminue de δ%, on ne l'accepte qu'à une probabilité de (1-δ)%.

Le cas suggéré par Hastings (1970) prend la forme $q(\theta,\theta')=f(\theta')$, où f est une densité connue. Ici, la DGC du candidat θ' est indépendante de la valeur de θ. La probabilité d'accepter un candidat est donc

$$\alpha(\theta,\theta') = \min\left\{\frac{p(\theta'|x)f(\theta)}{p(\theta|x)f(\theta')},1\right\} \quad . \tag{6}$$

Comme c'était le cas pour l'intégration de Monte-Carlo, on se sert d'une distribution fixe pour simuler les aléas; ces tirages sont ensuite transformés pour « mimiquer » la distribution d'intérêt. Il est alors raisonnable de penser que les

8. Voir Chib et Greenberg (1995).

critères pour un « bon » choix de $f(\theta)$ sont semblables à ceux mentionnés à la section 2.1.1 : une moyenne semblable et une variance un peu plus élevée que celle de $p(\theta|x)$. Ces caractéristiques sont toujours souhaitables, mais Chib et Greenberg (1995) notent qu'elles sont moins importantes.

Ces deux formulations sont les plus populaires, mais il existe plusieurs DGC utiles, comme la suggestion de Tierney (1994) d'utiliser un processus autorégressif tel $\theta'=A+B\theta+z$, où l'innovation z est tirée d'une loi connue ; la technique de la marche aléatoire étant un cas spécial de cet algorithme. Tierney (1994) ainsi que Chib et Greenberg (1995) suggèrent quelques autres choix pour $q(\theta,\theta')$.

Exemple 3 : Utiliser une pièce de monnaie pour approximer un processus binaire

Illustrons Metropolis-Hastings par une application très simple de l'algorithme : un processus binaire. Posons le paramètre θ pouvant prendre les valeurs 0 ou 1 et x, les données. On cherche $p(\theta=1|x)$. Supposons que la valeur du rapport $p(\theta=1|x)/p(\theta=0|x)$ est de 1/2. Dans ce cas simple, on peut montrer facilement que $p(\theta=1|x)=1/3$.

Supposons que les valeurs du candidat, θ', sont choisies à partir d'une DGC qui génère des valeurs de θ tel que $p(\theta=0)=p(\theta=1)=0{,}5$. Il y a donc quatre combinaisons possibles pour la paire (θ,θ'). La probabilité que le candidat soit accepté peut être dérivée à l'aide de (5) :

si $\theta=1$ et $\theta'=1$, alors $\alpha(\theta,\theta')=p(\theta=1|x)/p(\theta=1|x)=1 \Rightarrow \theta'$ accepté avec probabilité 1,

si $\theta=1$ et $\theta'=0$, alors $p(x|\theta=0|x)/p(\theta=1|x)=2$; $\alpha(\theta,\theta')=1 \Rightarrow \theta'$ accepté avec probabilité 1,

si $\theta=0$ et $\theta'=1$, alors $\alpha(\theta,\theta')=p(\theta=1|x)/p(\theta=0|x)=0{,}5 \Rightarrow \theta'$ accepté avec probabilité 0,5,

si $\theta=0$ et $\theta'=0$, alors $\alpha(\theta,\theta')=p(\theta=0|x)/p(\theta=0|x)=1 \Rightarrow \theta'$ accepté avec probabilité 1.

On peut montrer facilement que cette séquence peut s'exprimer comme une chaîne markovienne avec les probabilités de transition $p(\theta^i=0|\theta^{i-1}=0)=0{,}5$ et $p(\theta^i=0|\theta^{i-1}=1)=0{,}25$. La solution à ce processus se trouve dans l'exemple 1. Il en ressort que

$$\lim_{i\to\infty} p(\theta^i=1|\theta^0=0,x)=\lim_{i\to\infty} p(\theta^i=1|\theta^0=1,x)=1/3 \quad .$$

On voit qu'avec un nombre élevé d'itérations, l'algorithme Metropolis-Hastings nous permet de reproduire la bonne probabilité $p(\theta=1|x)$.

En principe, l'algorithme Metropolis-Hastings converge pour n'importe quel choix de DGC. En pratique, il existe quelques règles informelles permettant d'accélérer la convergence[9]. Par exemple, si la DGC génère trop de valeurs improbables, les candidats seront typiquement rejetés. Dans le cas de la marche aléatoire, si la DGC génère des candidats qui sont presqu'invariablement trop près de la dernière valeur de la chaîne, le passage entre deux valeurs moindrement distantes peut nécessiter plusieurs itérations. Dans les deux cas, les réalisations voisines seront fortement corrélées, ce qui réduit la vitesse de convergence de la statistique $\overline{g}$. Il est plus efficace de faire quelques tours d'estimation préliminaires pour identifier une DGC ayant la bonne forme, moyenne et écart-type. Idéalement, la séquence de tirages de l'argorithme devrait ressembler à celle tracée à la figure 1.

L'algorithme Metropolis-Hastings peut être utilisé pour estimer tous les paramètres du modèle simultanément ou conjointement avec l'échantillonneur de Gibbs. Par exemple, on peut estimer un modèle avec l'échantillonneur de Gibbs même si chaque (ou certaines) lois conditionnelles $p(\theta_j|\theta_{-j},x)$ est (sont) non standard. Dans ce cas, les tirages des lois $p(\theta_j|\theta_{-j},x)$ non standard peuvent être simulés à l'aide de l'algorithme Metropolis-Hastings.

2.4 *L'évaluation de la convergence et la précision numérique*

Dans une application des techniques MCMC - soit l'échantillonnage de Gibbs ou l'algorithme Metropolis-Hastings - la procédure est toujours :

1. Choisir une valeur de départ θ^0.

2. Simuler n réalisations de la chaîne markovienne avec n assez grand pour atténuer l'effet du choix de θ^0.

3. Conserver les N valeurs suivantes. Cet échantillon est utilisé pour calculer la statistique $\overline{g} \equiv N^{-1}\sum_{i=n+1}^{n+N} g(\theta^i)$ pour estimer $E[g(\theta)|x]$.

Puisque la chaîne markovienne est ergodique, la valeur de départ n'a aucune importance si n est suffisamment grand. De la même façon, si N est suffisamment grand, la précision de l'estimateur $\overline{g}$ peut atteindre n'importe quel niveau désiré. Le problème est maintenant de déterminer les valeurs appropriées de n et N.

Il n'existe aucun critère théorique définitif[10] pour décider si une chaîne markovienne a convergé à sa loi stable. Néanmoins, l'analyste pourrait choisir n à partir d'un graphique de la séquence simulée. Considérons l'exemple d'une application de l'échantillonneur de Gibbs dont la séquence simulée apparaît à la figure 1. Étant donné le comportement de cette chaîne, il est raisonnable de fixer n=100 ; après le 100e tour, les valeurs simulées sont généralement dans la même région.

---

9. Voir annexe.

10. Voir la discussion dans le *Journal of the Royal Statistical Society (B), 55, No. 1* (1993).

Une analyse plus rigoureuse devrait être fondée sur la précision avec laquelle la statistique $\bar{g}$ est estimée. Dans le cas où la séquence $\{\theta^i\}$ est simulée à partir de tirages indépendants d'une loi de variance σ^2, il est bien connu que la variance de la moyenne $N^{-1}\sum_{i=n+1}^{n+N}\theta^i$ est simplement $N^{-1}\sigma^2$. Dans le cas d'une séquence où les observations ne sont pas indépendantes, Geweke (1992) note que la variance de la moyenne d'un échantillon dont les observations sont corrélées $\{g(\theta^i)\}$ est $N^{-1}S_g(0)$, où $S_g(0)$ est la densité spectrale de $\{g(\theta^i)\}$ évaluée à 0. À partir de ce résultat, Geweke (1992) suggère la statistique suivante pour tester si les tirages de deux sous-ensembles de la chaîne simulée proviennent de la même distribution. Soient $\bar{g}_A$ et $\bar{g}_B$ les estimateurs calculées selon

$$\bar{g}_A \equiv N_A^{-1}\sum_{i=n+1}^{n+N_A} g(\theta^i)$$

$$\bar{g}_B \equiv N_B^{-1}\sum_{i=N-N_B+1}^{N} g(\theta^i)$$

Si la chaîne a convergé avant l'itération n, on s'attendrait à ce que $\bar{g}_A \simeq \bar{g}_B$. Définissons la statistique[11]

$$CD \equiv \frac{\bar{g}_A - \bar{g}_B}{\sqrt{N_A^{-1}S_A(0) + N_B^{-1}S_B(0)}}$$

Si N_A/N et N_B/N sont fixes, et si $N_A+N_B<N$, alors lorsque $N\to\infty$, $CD \xrightarrow{d} N(0,1)$ si la séquence $\{g(\theta^i)\}_{i=n+1}^{n+N}$ est stationnaire. Les grandes valeurs absolues de CD suggèrent que la séquence n'a pas convergé avant l'itération n; il faut alors augmenter n.

Si on conclut que la séquence a convergé à partir de la n^e itération, la variance de l'estimateur $\bar{g}$ est simplement $N^{-1}S_g(0)$. Notons que la précision à laquelle $E[g(\theta)|x]$ est estimée augmente avec N; l'analyste peut donc atteindre n'importe quel niveau de précision. Par exemple, si on veut un écart-type à 1% de l'estimateur, on fera quelques tours d'estimation préliminaires et il s'agira de retenir les statistiques $\hat{g}$ et $\hat{S}_g(0)$. On peut alors fixer N tel que $[N^{-1}\hat{S}_g(0)]^{1/2}=0{,}01\hat{g}$.

3. EXEMPLES ET APPLICATIONS

L'utilisation des postulats bayesiens aux modèles complexes a longtemps rencontré des difficultés ; calculer un estimateur classique de $g(\theta)$ était typiquement plus facile[12] que de calculer l'espérance *a posteriori* $E[g(\theta)|x]$. On a remarqué que le développement récent des techniques MCMC a simplifié l'analyse bayesienne. De plus, il existe plusieurs cas où l'application des méthodes bayesiennes à travers les techniques MCMC donne des gains considérables par rapport à l'analyse classique. Dans cette section, nous en décrivons quelques exemples. Les gains obtenus par l'application des techniques MCMC sont plus évidents dans

11. *CD* pour *Convergence Diagnostic*, Geweke (1992).
12. Voir le commentaire de Rust dans Poirier (1988).

les modèles avec variables latentes. Supposons que le modèle statistique est décrit par une densité conjointe $p(x,y|\theta)$, où x est observée et où y est une variable latente. Étant donné un échantillon $x_1, ..., x_T$, la fonction de vraisemblance $L(\theta|x)$ est calculée à partir de l'intégrale

$$L(\theta|x)=\int...\int L(\theta|x,y)p(y|\theta)dy_1...dy_T \quad . \tag{7}$$

Évidemment, si l'intégrale multiple (7) n'a pas de solution analytique, il est difficile d'estimer θ à partir de sa fonction de vraisemblance $L(\theta|x)$.

Néanmoins, l'estimation de θ est souvent directe si y est observée avec x, c'est-à-dire, même si la loi *a posteriori* $p(\theta|x)$ est difficile à traiter, la distribution conditionnelle $p(\theta|y,x)$ reste souvent standard. On peut alors se servir de la technique *d'augmentation des données* suggérée par Tanner et Wong (1987). Au lieu d'intégrer la fonction de vraisemblance par rapport à y, on traite les données manquantes comme des paramètres à estimer. L'estimation peut donc utiliser des techniques MCMC selon l'algorithme

$$\begin{aligned} &\theta \sim p(\theta|y,x) \\ &y \sim p(y|\theta,x) \quad . \end{aligned} \tag{8}$$

La séquence générée par (8) converge en distribution à la loi conjointe $p(\theta,y|x)$. Si on ne considère que les valeurs de θ, celles-ci correspondent aux valeurs tirées de la loi marginale $p(\theta|x)$. Par contre, si on s'intéresse aux valeurs réalisées de la variable latente y, on peut estimer $E[y|x]$ de la même façon qu'on estime $E[\theta|x]$.

Dans ce qui suit, nous décrivons brièvement comment cette idée de base a été appliquée à quelques modèles bien connus.

3.1 *Les modèles probit*

Rappelons le modèle linéaire multivarié de l'Exemple 2 :

$$y_t=X_t\beta+u_t \qquad\qquad u_t \sim iid\ N(0,\Sigma) \quad .$$

Dans un modèle probit, on interprète $y_{i,t}$ comme le rendement ou l'utilité de l'option i perçue par l'individu ayant les caractéristiques X_t. Si y_t est observée, le système devient un modèle SUR. Le problème posé par les modèles probit est que y_t n'est pas observée ; on observe seulement quelle option a été choisie. Si l'option j est choisie par l'individu t, la contribution de l'observation t à la fonction de vraisemblance devient simplement la probabilité que $y_{i,t}<y_{j,t}$, $\forall\ i\neq j$. Puisque cette probabilité exige l'intégration d'une densité normale multivariée *pour chaque observation*, il est très difficile d'estimer un modèle probit si la dimension de y_t est importante.

Albert et Chib (1993a) font remarquer que l'augmentation des données facilite l'estimation des modèles de type qualitatif ; McCulloch et Rossi (1994) ont élaboré cette approche dans le contexte des modèles probit. Si y_t est connue,

on peut appliquer l'algorithme décrit dans l'Exemple 2 pour simuler des valeurs de β et Σ de leurs lois conditionnelles. Si on observe seulement le choix j, on note que la distribution conditionnelle de y_t est la loi normale multivariée tronquée $\mathbb{I}_{[y_{i,t}<y_{j,t}]} N(X_t\beta,\Sigma)$, où $\mathbb{I}_{[y_{i,t}<y_{j,t}]}$ est une fonction indicatrice qui assure que le rendement de l'option j est supérieur à celui des autres options. Pour y appliquer l'algorithme de l'échantillonnage de Gibbs, on ajoute l'étape d'augmentation des données à l'algorithme décrit dans l'Exemple 2 :

$$\begin{aligned} \beta &\sim N(\overline{\beta},\overline{\Sigma}) \\ \Sigma^{-1} &\sim \text{Wishart}(\overline{\nu},\overline{V}) \\ y_t &\sim \mathbb{I}_{[y_{i,t}<y_{j,t}]} N(X_t\beta,\Sigma) \quad t=1,\ldots,T \end{aligned}$$

où $\overline{\beta}$, $\overline{\Sigma}$, $\overline{\nu}$ et $\overline{V}$ sont les mêmes expressions que celles données dans l'Exemple 2. McCulloch et Rossi (1994) décrivent comment tirer les y_t d'une loi normale tronquée pour assurer que $y_{i,t}<y_{j,t}$.

On sait que le modèle probit n'est pas identifié[13], cette chaîne ne sera donc pas convergente. On contourne ce problème en appliquant la transformation standard $\tilde{\beta}^i = (1/\Sigma^i_{1,1})\beta^i$, $\tilde{\Sigma}^i = (1/\Sigma^i_{1,1})\Sigma^i$. Les moments *a posteriori* des paramètres du modèle identifié sont calculés à partir de la séquence $\{(\tilde{\beta}^i,\tilde{\Sigma}^i)\}$.

3.2 *Les modèles de volatilité stochastique*

Jacquier, Polson et Rossi (1994) ainsi que Geweke (1994) ont appliqué des techniques MCMC avec augmentation de données à l'estimation d'un modèle de volatilité stochastique. Soit un modèle :

$$x_t = \sqrt{y_t}\,u_t \tag{9a}$$

$$\ln y_t = \beta_1 + \beta_2 \ln y_{t-1} + \sigma v_t \tag{9b}$$

où u_t et v_t sont des aléas indépendants $N(0,1)$; définissons $\beta=(\beta_1,\beta_2)'$. Dans ce modèle simple, le logarithme de la variance de x_t suit un processus stochastique autorégressif; rappelons qu'un modèle GARCH suppose que y_t est une fonction déterministe des observations précédentes. Le modèle (9) peut incorporer plusieurs modifications sans affecter l'intuition de ce qui suit.

On note que la fonction de vraisemblance des paramètres, étant donné $x\equiv\{x_1,\ldots,x_T\}$, est l'intégrale multiple (7); Jacquier, Polson et Rossi (1994) décrivent quelques approches classiques pour estimer le modèle. Ils remarquent aussi que l'estimation du modèle (9) est typiquement fondée sur une approximation de la fonction de vraisemblance. De plus, toutes les inférences doivent se servir d'approximations asymptotiques.

---

13. Ce que l'on observe - le choix de l'individu - reste inchangé si tous les paramètres du modèle sont multipliés par une constante positive.

Il est évident que l'estimation de β et de σ ne pose aucun problème si y est connue ; dans ce cas, on peut se limiter au modèle linéaire courant, $y=Z\beta+u$, où Z est la matrice des variables du côté droit de (9b) et où $u\sim N(0,\sigma^2 I_T)$. Si on suppose que les croyances *a priori* de β prennent la forme $p(\beta)=N(\hat{\beta},\hat{A}^{-1})$ et que σ^2 est connue, on peut diviser toutes les variables par σ pour obtenir $\tilde{y}=\tilde{Z}\beta+\varepsilon$, où $\varepsilon\sim N(0,I_T)$. On obtient donc la distribution conditionnelle $p(\beta|\sigma^2,x,y)=N\left(\overline{\beta},\overline{A}^{-1}\right)$, où $\overline{A}=\hat{A}+\tilde{Z}'\tilde{Z}$ et où $\overline{\beta}=\overline{A}^{-1}(\hat{A}\hat{\beta}+\tilde{Z}'\tilde{y})$. De la même façon, si la loi *a priori* de σ^2 est une loi inverse-gamma, $p(\sigma^{-2})=G(\hat{\nu}/2,\ 2/\hat{\nu}\hat{s}^2)$ on note que la distribution conditionnelle $p(\sigma^{-2}|\beta,x,y)=G(\overline{\nu}/2,\ 2/\overline{\nu}\overline{s}^2)$, où $\overline{\nu}=\hat{\nu}+T$ et où $\overline{s}^2=\overline{\nu}^{-1}\left[\hat{\nu}\hat{s}^2+(y-Z\beta)'(y-Z\beta)\right]$.

Comme dans le modèle probit, l'étape clé de l'algorithme est la simulation des données manquantes. La distribution conditionnelle $p(y|\beta,\sigma^2,x)$ n'a pas une forme standard. Il faut alors simuler chaque y_t de la loi conditionnelle $p(y_t|\beta,\sigma^2,x,y_{-t})$, où y_{-t} représente l'ensemble de toutes les valeurs de y à l'exception de y_t. Jacquier, Polson et Rossi (1994) notent que cette distribution conditionnelle univariée est :

$$p(y_t|\beta,\sigma^2,x,y_{-t}) \propto p(x_t|y_t)p(y_t|y_{t-1})p(y_{t+1}|y_t)$$

$$\propto y_t^{-1/2}\exp\left\{\frac{-x_t^2}{2y_t}\right\}y_t^{-1}\exp\left\{\frac{-(\ln y_t-\mu_t)^2}{2\sigma^2}\right\} \qquad (10)$$

où

$$\mu_t\equiv\frac{\beta_1(1-\beta_2)+\beta_2(\ln y_{t+1}+\ln y_{t-1})}{1+\beta_2^2}\ ,\ \tilde{\sigma}^2\equiv\frac{\sigma^2}{1+\beta_2^2}\ .$$

Puisque la forme de (10) est non standard, Jacquier, Polson et Rossi (1994) se servent de l'algorithme Metropolis-Hastings pour simuler la séquence des y_t. La chaîne markovienne est donc décrite par :

$$\begin{aligned}\beta &\sim p(\beta|\sigma^2,x,y)\\ \sigma^2 &\sim p(\sigma^2|\beta,x,y)\\ y_t &\sim p(y_t|\beta,\sigma^2,x,y_{-t}) \qquad t=1,\ldots,T\end{aligned} \qquad (11)$$

La structure du modèle est semblable à celle des modèles de régimes markoviens introduits par Hamilton (1989). Dans un modèle de régimes markoviens, les cycles économiques sont incorporés dans un modèle de séries chronologiques à l'aide d'une variable latente discrète suivant un processus markovien. Comme dans le cas du modèle de volatilité stochastique, si les variables latentes sont observées, l'estimation du modèle est simple. Albert et Chib (1993b) de même que McCulloch et Tsay (1992) utilisent une étape d'augmentation de données en utilisant une distribution conditionnelle semblable à celle décrite par (10).

3.3 *La sélection de modèle*

On a remarqué dans l'application du théorème de Bayes (3) que la probabilité *a posteriori* d'un modèle m=1,2, ... ,K est calculée à partir des vraisemblances marginales $p(x|M^m)$. Carlin et Polson (1991) ont remarqué que l'intuition de l'augmentation des données peut être adaptée au choix de modèle ; on peut interpréter l'indicatrice M^m comme un paramètre à estimer, et on peut le traiter de la même façon que tous les autres paramètres.

S'il advenait que pour chaque modèle m, le paramètre θ^m était connu, la probabilité $p(x|M^m,\theta^1, \ldots ,\theta^K)$ serait simplement

$$p(x|M^m,\theta^1,\ldots,\theta^K) = \frac{p(x|\theta^m,M^m)p(\theta^m|M^m)p(M^m)}{\sum_{k=1}^{K} p(x|\theta^k,M^k)p(\theta^k|M^k)p(M^k)} \quad . \qquad (12)$$

Étant donné les K probabilités générées par (12), on peut simuler une valeur de M^m de la distribution multinomiale appropriée. Dès que le modèle m est choisi, on peut simuler $\theta^m \sim p(\theta^m|x,M^m)$ à l'aide des techniques décrites ci-haut[14]. On crée donc une chaîne markovienne définie par la séquence :

$$M^m \sim p(M^m|x,\theta^1,\theta^2,\ldots,\theta^K)$$

$$\theta^m \sim p(\theta^m|x,M^m) \quad .$$

L'estimateur pour $p(M^m|x)$ devient simplement le nombre de fois que le modèle m est choisi, divisé par le nombre d'itérations effectuées ; Carlin et Polson (1991) notent que cet estimateur a un écart-type maximal de $(1/4N)^{1/2}$.

Cette approche face au problème de sélection de modèle a été appliquée au problème de choix de variables exogènes d'un modèle[15].

CONCLUSION

En plus des applications illustrées précédemment, les techniques MCMC ont connu beaucoup de succès avec plusieurs autres modèles divers. Une liste partielle de ces applications inclut les modèles ARMA(p,q)[16], les modèles EDTAR[17], les modèles avec erreurs composées[18], ainsi qu'une solution au problème d'identification d'une brisure structurelle[19]. Puisque la popularité des techniques MCMC est assez récente et que les économètres bayesiens sont toujours peu nombreux, il est peu probable qu'on épuisera rapidement la quantité d'applications économétriques possibles de ces techniques.

---

14. Carlin et Chib (1995) décrivent comment simuler des nouvelles valeurs pour les paramètres des modèles qui n'ont pas été choisis.
15. Voir George et McCulloch (1993,1994) ainsi que Geweke (1994b).
16. Chib et Greenberg (1994).
17. Koop et Potter (1994).
18. Koop, Osiewalski et Steel, (1994), (1995).
19. Carlin, Gelfand et Smith (1992).

Les méthodes d'échantillonnage pourraient devenir, d'ici quelques années des instruments privilégiés de l'inférence statistique. L'étendue des applications et la facilité d'utilisation, en font des concurrents sérieux aux méthodes standard actuelles. Bien qu'elles demandent un investissement initial pour la compréhension et la mise en pratique, elles deviennent extrêmement faciles d'application par la suite. L'économètre appréciera sa flexibilité face aux modèles variés auxquels il est confronté.

ANNEXE

LE CALIBRAGE DE L'ALGORITHME METROPOLIS-HASTINGS

Cette annexe démontre comment utiliser des graphiques pour calibrer les DGC et accélérer la vitesse de convergence. Dans les graphiques qui suivent, on a tenté de simuler des aléas d'une $N(0,1)$ à partir de l'algorithme Metropolis-Hastings. Les figures 2 et 3 se basent sur une estimation utilisant la technique des chaînes d'indépendance (DGC fixe), tandis que celle de la marche aléatoire est utilisée à la figure 4.

La figure 2 est un exemple typique d'une estimation faite avec une DGC ayant une variance trop forte. Les candidats proposés sont trop souvent improbables et il en résulte un taux d'acceptation très faible. On remarque qu'il y a beaucoup de plateaux dans la distribution. Notons qu'une évaluation par la marche aléatoire avec une trop grande variance de la DGC générerait un graphique similaire.

La figure 3 rend compte d'un problème différent. Ici, on voit clairement que la moyenne de la DGC est trop élevée. En effet, le bas du graphique est principalement composé de plateaux, tandis que le haut est composé de pointes. Ceci démontre que les candidats sont rarement générés vers le bas, mais y restent plus longtemps ; l'inverse étant vrai pour le haut du graphique.

La figure 4 ne s'applique qu'à la technique de la marche aléatoire. Dans cet exemple, la DGC a une variance trop faible. Les changements de valeurs sont tellement faibles à chaque tour d'estimation qu'il en faut plusieurs pour couvrir tout le spectre de la distribution. Graphiquement, on le voit par les longues traînées que semblent faire les valeurs retenues à chaque tour.

FIGURE 2

METROPOLIS-HASTINGS : DGC FIXE AVEC UNE VARIANCE TROP ÉLEVÉE

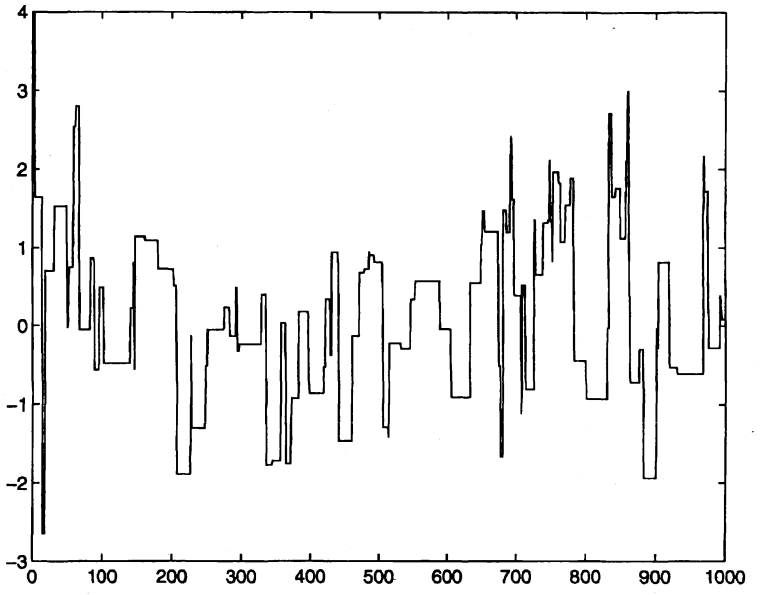


FIGURE 3

METROPOLIS-HASTINGS : DGC FIXE AVEC UNE MOYENNE TROP ÉLEVÉE

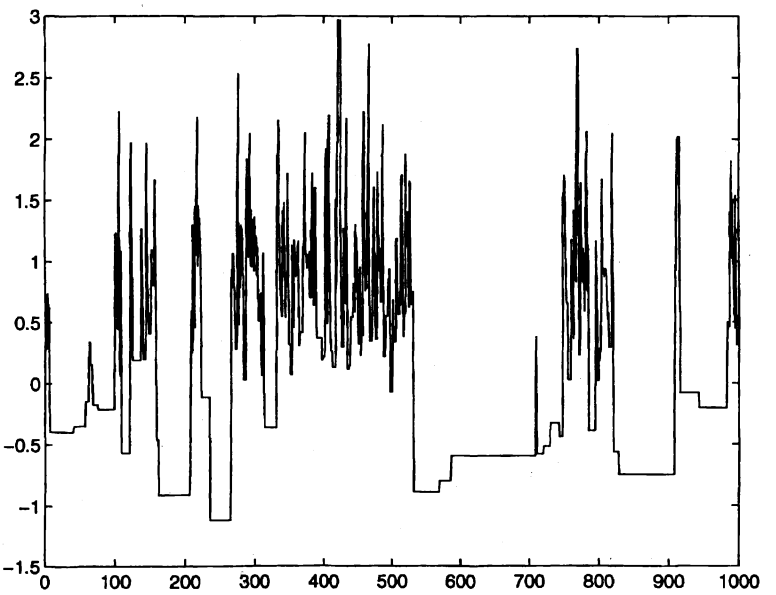
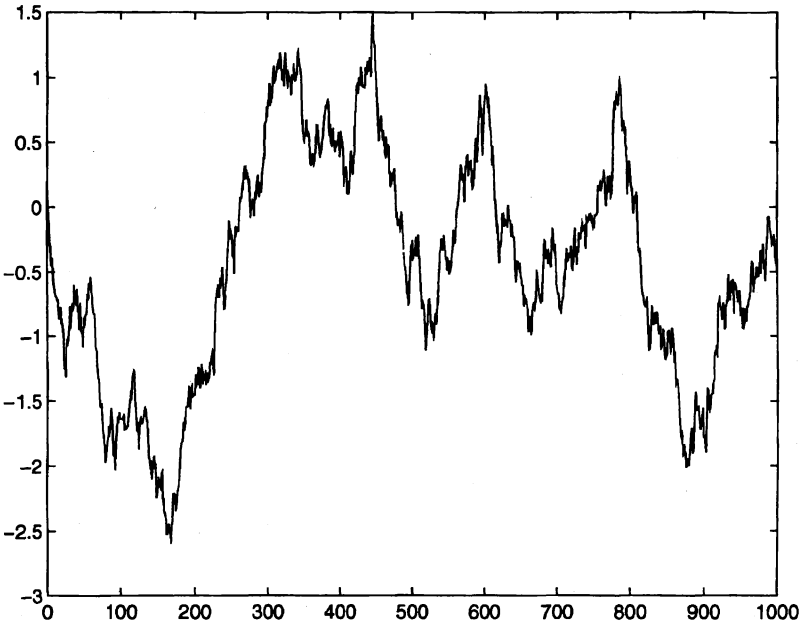


FIGURE 4

METROPOLIS-HASTINGS : MARCHE ALÉATOIRE AVEC UNE VARIANCE TROP FAIBLE



BIBLIOGRAPHIE

ALBERT, J., et S. CHIB (1993a), « Bayesian Analysis of Binary and Polychotomous Response Data », *Journal of the American Statistical Association* 88 : 669-679.

ALBERT, J., et S. CHIB (1993b), « Bayes Inference via Gibbs Sampling of Autoregressive Time Series subject to Markov Mean and Variance Shifts », *Journal of Business and Economic Statistics,* 11 : 1-15.

BERNARDO, J.M., et A.F.M. SMITH (1994), *Bayesian Theory*, Chichester, UK : John Wiley and Sons.

CARLIN, B.P., et S. CHIB (1995), « Bayesian Model Choice via Markov Chain Monte Carlo », *Journal of the Royal Statistical Society-B* 57 : 473-484.

CARLIN, B.P., A.E. GELFAND, et A.F.M. SMITH (1992), « Hierarchical Bayesian Analysis of Changepoint Problems », *Journal of the Royal Statistical Society-A (Applied Statistics)* 41 : 389-405.

CARLIN, B.P., et N.G. POLSON (1991), « Inference for Nonconjugate Bayesian Models Using the Gibbs Sampler », *Canadian Journal of Statistics* 19 : 399-405.

CASELLA, G., et E. GEORGE (1992), « Explaining the Gibbs Sampler », *The American Statistician* 46 : 167-174.

CHIB, S., et E. GREENBERG (1994), « Bayes Inference in Regression Models with ARMA(p,q) Errors », *Journal of Econometrics* 64 : 183-206.

CHIB, S., et E. GREENBERG (1995), « Understanding the Metropolis-Hastings Algorithm », *The American Statistician* 49 : 327-335.

GELFAND, A., et A. SMITH (1990), « Sampling-Based Approaches to Calculating Marginal Densities », *Journal of the American Statistical Association* 85 : 398-409.

GEMAN, D., et S. GEMAN (1984), « Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images », *IEEE Transaction on Pattern Analysis and Machine Intelligence* 6 : 721-741.

GEORGE, E.I., et R.E. MCCULLOCH (1993), « Variable Selection via Gibbs Sampling », *Journal of the American Statistical Association* 88 : 881-889.

GEORGE, E.I., et R.E. MCCULLOCH (1994), « Fast Bayes Variable Selection », miméo.

GEWEKE, J. (1989), « Bayesian Inference in Econometric Models using Monte Carlo Integration », *Econometrica* 57 : 1317-1339.

GEWEKE, J. (1992), « Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments » dans *Bayesian Statistics, 4*, J.M. BERNARDO, J.O. BERGER, A.P. DAWID and A.F.M. SMITH (ed.) Oxford : Oxford University Press : 169-193.

GEWEKE, J. (1994a), « Bayesian Comparison of Econometric Models », Federal Reserve Bank of Minneapolis Reseach Department Working Paper 532.

GEWEKE, J. (1994b), « Variable Selection and Model Comparison in Regression », Federal Reserve Bank of Minneapolis Reseach Department Working Paper 539.

HAMILTON, J. (1989), « A New Approach to the Economic Analysis of Non-Stationary Time Series and the Business Cycle », *Econometrica* 57 : 357-84.

HASTINGS, W.K. (1970), « Monte Carlo Sampling Methods Using Markov Chains and their Applications », *Biometrika* 57 : 97-109.

JACQUIER, E., N.G. POLSON, et P.E. ROSSI (1994), « Bayesian Analysis of Stochastic Volatility Models » (avec discussion), *Journal of Business and Economic Statistics* 12 : 371-417.

KLOEK, T., et H.K. VAN DIJK (1978), « Bayesian Estimates of Equation System Parameters : an Application of Integration by Monte Carlo », *Econometrica* 46 : 1-20.

KOOP, G., J. OSIEWALSKI, et M.F.J. STEEL (1994), « Bayesian Efficiency Analysis with a Flexible Functional Form : the AIM Cost Function », *Journal of Business and Economic Statistics* 12 : 339-346.

KOOP, G., J. OSIEWALSKI, et M.F.J. STEEL (1995), « The Components of Output Growth : a Cross-Country Analysis », miméo.

KOOP, G., et S. POTTER (1994), « Bayesian Analysis of Endogenous Delay Threshold Models Using the Gibbs Sampler », miméo.

LEAMER, E.E. (1978), *Specification Searches*. John Wiley & Sons.

METROPOLIS, N., A.W. ROSENBLUTH, M.N ROSENBLUTH, A.H. TELLER, et E. TELLER (1953), « Equations of State Calculations by Fast Computing Machines », *Journal of Chemical Physics* 21 : 1087-1092.

MCCULLOCH, R., et P.E. ROSSI (1994), « An Exact Likelihood Analysis of the Multinomial Probit Model », *Journal of Econometrics* 64 : 207-240.

MCCULLOCH, R., et R. TSAY (1992), « Statistical Inference of Markov Switching Models with Application to U.S. GNP », Graduate School of Business Working paper, University of Chicago.

PHILLIPS, P.C.B. (1991), « To Criticize the Critics : an Objective Bayesian Analysis of Stochastic Trends » (avec discussion), *Journal of Applied Econometrics* 6 : 333-364.

POIRIER, D.J. (1988), « Frequentist and Subjectivist Perspectives on the Problems of Model Building in Economics », *Journal of Economic Perspectives* 2 : 121-170.

RITTER, Christian, et Martin A. TANNER (1992), « Facilitating the Gibbs Sampler : the Gibbs Stopper and the Griddy-Gibbs Sampler », *Journal of the American Statistical Association* 87 : 861-868.

ROBERTS, G.O., et A.F.M. SMITH (1994), « Simple Conditions for the Convergence of the Gibbs Sampler and Metropolis-Hastings Algorithms », *Stochastic Processes and their Applications* 49 : 207-216.

STOKEY, N.L., et R.E. LUCAS (1989), *Recursive Methods in Economic Dynamics*, Harvard University Press.

TANNER, M.A., et W.H. WONG (1987), « The Calculation of Posterior Distributions by Data Augmentation », *Journal of the American Statistical Association* 82 : 528-549.

TIERNEY, L. (1994), « Markov Chains for Exploring Posterior Distributions » (avec discussion), *Annals of Statistics* 22 : 1701-1762.

ZELLNER, Arnold (1971), *An Introduction to Bayesian Inference in Econometrics*, John Wiley & Sons.