$N°$**D'ordre :19/2013-M/MT**

République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique Université des Sciences et de la Technologie Houari Boumediene
Faculté de Mathématiques

MÉMOIRE
Présenté pour l'obtention du diplôme de MAGISTER

EN : MATHEMATIQUES

Spécialité : Probabilités et Statistique

par : MESSACI Yasmina

THÈME

L'utilisation de la méthode MCMC et son application actuarielle

Soutenu publiquement, le 07/03/2013 devant le jury composé de :

M.	TATACHAK	Abdelkader	Maître de Conférence	à l'USTHB	Président
M.	BOUKHETALA	Kamal	Professeur	à l'USTHB	Directeur de mémoire
M.	KHALDI	Khaled	Professeur	à l'UMBB	Examinateur

Remerciements

Je tiens à exprimer mes sincères remerciements à Mr K.BOUKHETALA professeur à l'USTHB pour m'avoir proposé ce sujet et accepté de diriger ce travail. Je le remercie aussi pour sa disponibilité et ses nombreux conseils qu'il m'a prodigué durant la réalisation de ce mémoire.

Je tiens aussi à exprimer mes vifs remerciements à Mr A.TATACHAK maitre de conférence à l'USTHB de l'honneur qu'il me fait d'avoir accepté de présider le jury.

J'exprime ma gratitude à Mr K.KHALDI professeur à l'université M'hamed Bougara de Boumerdès pour avoir accepté d'être membre du jury et d'examiner ce travail.

Mes remerciements vont aussi à tous les enseignants du département de Probabilités et Statistique qui ont contribué à ma formation.

A ma famille et mes amies

Table des matières

Table des figures

Liste des tableaux

Résumé

Dans ce mémoire, la méthodologie de simulation de Monte Carlo par chaines de Markov est développée pour l'estimation des paramètres des deux modèles à espace d'état non gaussiens : le modèle à durée conditionnelle stochastique et le modèle à volatilité stochastique dans ses deux versions discrète et continue. La stratégie de simulation utilisée est la stratégie hybride (utilisation simultanée des algorithmes de Metropolis Hastings et Gibbs). Le vecteur d'états est simulé par blocs en appliquant l'algorithme du filtre de Kalman. Plusieurs diagnostics de convergence sont mis en oeuvre pour étudier la convergence de ces méthodes lorsque la taille des données et le nombre d'itérations varient. On montre que lorsque le nombre d'observations disponibles augmente la convergence est plus lente.
Mots clés : Analyse bayésienne, diagnostics de convergence, modèle à durée conditionnelle stochastique, modèle à espace d'état, modèle à volatilité stochastique, Monte Carlo par chaines de Markov, processus de diffusion.

Introduction

Durant les dernières années, les outils mathématiques sont devenus déterminants en finance. Et depuis la théorie financière a été modifiée avec l'apparition de nouveaux marchés et produits. Ce changement (fait suite à une volonté accrue de déréglementation dans les années 1970), rendant volatiles des taux d'intérêt et instables les taux de change. Des marchés organisés ont vu le jour et permis à des intervenants comme les entreprises industrielles et commerciales, les compagnies d'assurance et les banques d'intervenir massivement sur un marché unique et liquide. Le développement important de ces activités a été rendu possible grâce aux progrès technologiques, mais aussi grâce aux outils théoriques qui ont permis de valoriser les nouveaux produits financiers. De nos jours, une large gamme d'outils des mathématiques appliquées est utilisée : en probabilités (mouvement brownien, processus de diffusion, calcul stochastique, méthodes de simulation de type Monte Carlo,...) ou en inférence statistique (estimation de paramètres). En finance et actuariat les données arrivent souvent sequentiellement dans le temps. Les processus stochastiques constituent donc un outil prévilégié pour modéliser ce genre de phénomènes. Parmi les outils théoriques, les modèles à espace d'état qui sont devenus un outil de modélisation de plusieurs modèles financiers et actuariels notamment lorsque les données arrivent séquentiellement dans le temps.

La forme espace d'état fournit un cadre approprié pour la construction de modèles de séries chronologiques, elle offre plus de flexibilité notamment en présence de non stationnarité. Les modèles à espace d'état modélisent une série d'observations par l'intermédiaire d'un processus latent (inobservé). La classe la plus générale de ces modèles est la classe des systèmes non linéaires et non gaussiens. Récemment, les modèles à espace d'état non gaussiens ont joui d'un intérêt croissant dû à l'évolution des techniques de simulation stochastique. En particulier, ils ont été utilisés pour modéliser les différentes caractéristiques de diverses séries financières et actuarielles, telles que : les durées entre les transanctions boursières, la volatilité des rendements d'ac-

tifs. Différentes approches sont utilisées pour l'estimation dans cette classe de modèles, la méthode du maximum de vraisemblance par exemple, mais aussi des méthodes bayésiennes qui présentent plusieurs avantages, notamment depuis le développement rapide des méthodes de Monte Carlo par chaines de Markov. Durbin et Koopman (1998) ont considéré l'analyse des séries temporelles non gaussiennes en utilisant la forme espace d'état avec l'approche classique et bayésienne avec une gamme de méthodes de Monte Carlo par chaines de Markov (MCMC) ayant été employés pour mettre en oeuvre une analyse bayésienne de ces modèles.

Dans ce mémoire, nous étudions l'utilisation de ces méthodes dans deux modèles à espace d'état non gaussiens particuliers. Le premier modèle étant une forme particulière du modèle à durées conditionnelles stochastiques (SCD) qui modélise les durées entre les transactions boursières en fonction d'un processus latent qui est basée sur l'hypothèse de données conditionnellement exponetielles. Une variété de ce modèle a été étudiée par différents auteurs avec des hypothèses de distribution alternatives voir par exemple Strickland, Forbes et Martin (2006) et Bauwens et Veredas (1999). Le second modèle considéré est le modèle à volatilité stochastique (SV) qui modélise la volatilité des rendements d'actifs en tant que processus latent reposant sur une normalité conditionnelle. Plusieurs auteurs se sont intéréssés au modèles SV parmi eux Jacquier, Polson et Rossi (1994) qui ont étudié l'analyse bayésienne des modèles à volatilité stochastique, Kim, Shephard et Chib (1998) qui ont traité les modèles SV en utilisant différentes méthodes de simulation alternatives bayésiennes et classique. Ces deux modèles sont à temps discet, nous étendons aussi notre étude aux modèles à temps continu (processus de diffusion) qui sont aussi très répandus en théorie financière et actuarielle. Dans les modèles en temps continu la caractérisation de la densité a posteriori conjointe est difficile pour plusieurs raisons. Premièrement, les observations sont disponibles en temps discret tandis que le modèle théorique est en temps continu donc l'augmentation des données introduira des variables d'état (latentes). Deuxièment, la densité a posteriori est de dimension très élevée et donc les méthodes de simulation standards ne sont pas fiables. Troisièment, dans plusieurs modèles en temps continu la densité a posteriori est non gaussienne et non linéraire donc les méthodes d'estimation classiques ne sont pas applicables. Eraker (2001) propose les méthodes MCMC pour estimer les paramètres des modèles de diffusion en utilisant des séries chronologiques discrètes et en faisant augmenter les données pour faire disparaître le biais de discrétisation, cette méthode a été appliquée à la classe des modèles à élascticité constante de variance (CEV) et à une classe plus générale de ces modèles à savoir les modèles à volatilité stochastique (SV). Nous prenons

l'exemple du modèle SV en temps continu (écrit sous forme d'une diffusion) et nous étudions l'estimation dans ces modèles en utilisant les méthodes MCMC.

Danc ce travail, nous nous intéressons particulièrement à l'efficience des méthodes bayésiennes en utilisant plusieurs critères d'évaluation ainsi que le comportement de leurs convergence pour les deux modèles à espace d'état non gaussiens SCD et SV cités précédemment. Plusieurs auteurs se sont intéréssés à ce sujet sous plusieurs angles. Kim, Shephard et Chib (1998) ont décrit une variété d'algorithmes de simulation pour l'estimation des modèles à volatilité stochastique. Ils ont d'abord appliqué un algorithme d'acceptation rejet et pour des raisons de lenteur de convergence, ils ont appliqué des algorithmes MCMC, en linéarisant l'équation de mesure et en l'approximant par une mixture de variables aléatoires normales, les résultats obtenus concluent que cette méthode proposée peut réaliser des gains d'efficacité significatifs. Dans le cadre du modèle SCD Strickland et al (2003) ont utilisés les méthodes MCMC pour estimer les paramètres et les variables latentes ces dernières sont d'abord simulées par blocs (multi move), puis élément par élément (single move), ils ont aussi comparé l'approche bayésienne avec l'approche classique (QML). Les critères de comparaison sont le facteur d'inefficacité ainsi que le RMSE. Strickland et al (2008) ont étudié ce problème en définissant plusieurs parametrisations pour les deux modèles SCD et SV et en étudiant l'impact de ces dernières sur la convergence. Le critère de comparaison est le facteur d'inefficacité, les résultats des simulations montrent pour les deux modèles que les gains d'efficacité les plus importants sont obtenus en déplaçant le paramètre d'echelle ou les paramètres d'echelle et de position de l'équation d'état vers l'équation de mesure.
En ce qui nous conserne nous allons étudier le comportement de la convergence en faisant varier T (la taille des données) et M (le nombre d'itérations) simultanément. L'augmentation de T s'interprète comme une augmentation de la quantité d'information connue sur le processus étudié. Donc on peut se poser la question : cette information supplémentaire influe-t-elle sur la vitesse de la convergence ?
Ces conclusions ont été déterminées en comparant les résultats de la convergence en utilisant en plus du facteur d'inefficacité les tests classiques de convergence des méthodes MCMC à savoir : les diagnostics visuels, diagnostics de Gelman, Geweke, Raftery et Heidelberg.

Le présent mémoire est structuré comme suit : le chapitre 1 présente les notions fondamentales de la statistique bayésienne en décrivant les modèles bayésiens et l'inférence bayésienne. Le chapitre 2 décrit les méthodes MCMC

ainsi que les différents algorithmes appartenant à la classe des méthodes MCMC. Ce chapitre décrit également les différents outils utilisés pour tester la convergence de ces méthodes vers la distribution cible. Dans le chapitre 3 sont définis les modèles à espace d'état et les méthodes d'estimation dans ce type de modèles qui consiste à estimer les variables d'état (latentes) ainsi que les paramètres du modèle. Dans le chapitre 4 les résultats précédents sont appliqués aux deux modèles à espace d'état non gaussiens en temps discret particuliers à savoir les modèles SCD et SV ainsi qu'au modèle SV en temps continu, nous rappellons aussi quelques définitions et résultats des processus de diffusion et équations differentielles stochastiques. Le chapitre 5 contient les différents résultats obtenus des simulations de ces deux modèles particuliers ainsi qu'une comparaison des résultats des différents diagnostics de convergence par rapport à la taille des données fournie et le nombre d'itérations utilisé dans les méthodes MCMC.

Chapitre I

Statistique bayésienne

Durant les dernières années il y a eu un intérêt croissant pour les méthodes statistiques bayésiennes et leur champ d'applications ne cesse de s'agrandir. La différence fondamentale entre les approches bayésienne et classique est que dans l'inférence statistique classique les données sont supposées être des réalisations d'une variable aléatoire dont la loi dépend d'un paramètre θ inconnu mais fixé. Dans l'inférence bayésienne θ est considéré lui même comme variable aléatoire et donc ayant une loi de probabilité. La démarche bayésienne consiste à probabiliser le paramètre inconnu en lui associant une loi de probabilité sur l'espace des paramètres dite loi a priori. Cette loi représente l'ensemble des informations a priori disponibles sur le paramètre θ ainsi que les incertitudes qui s'y rattachent.

I.1 Modèles statistiques bayésiens

I.1.1 Définition

Rappelons qu'un modèle statistique classique est défini par la donnée d'une variable aléatoire Y observée un certain nombre de fois $y_1, y_2, ..., y_n$ qui constituent les données et par une famille de distribution de probabilité $\mathcal{P} = \{P_\theta/\theta \in \Theta\}$ censée contenir la vraie loi de probabilité de Y inconnue.
Dans le cas où P_θ est absolument continue la donnée de $\mathcal{P}$ est équivalente à la donnée de la famille des densités $\{f_\theta/\theta \in \Theta\}$.
Un modèle statistique bayésien est défini comme un modèle statistique classique avec la spécification supplémentaire de la distribution conjointe des données et du paramètre, i.e la densité

$$f(y, \theta) = f(y/\theta)\pi(\theta) = f_\theta(y)\pi(\theta) \tag{I.1}$$

de (Y, Θ) (ici Θ est une variable aléatoire et non pas un espace de paramètre)
Dans l'analyse bayésienne $f_\theta(y)$ est interprétée comme la densité conditionnelle de Y sachant $\Theta = \theta$ appelée vraisemblance et notée $L(\theta, y)$ et $\pi(\theta)$ est la distribution a priori du paramètre θ.
Une analyse bayésienne procède suivant les quatre étapes suivantes :

1. Spécification de la vraisemblance $f(y/\theta)$, la distribution de $Y/\Theta = \theta$, donnée conditionnellement au paramètre.
2. Spécification de la distribution a priori $\pi(\theta)$ du paramètre inconnu θ.
3. Observer les données y et calculer la distribution a posteriori $\pi(\theta/y)$, la distribution de $\Theta/Y = y$, du paramètre conditionnellement aux données observées.
4. Faire l'inférence à partir de la distribution a posteriori.

Si dans l'approche classique l'inférence est basée sur la vraisemblance, l'approche bayésienne elle se base sur la distribution a posteriori $\pi(\theta/y)$.
Une autre différence est que l'hypothèse classique d'indépendance et d'équidistribution des observations se transforme en l'hypothèse d'échangeabilité des observations.
Dans un modèle bayésien, la distribution du vecteur des données $(y_1, ..., y_n)$ dépend de la valeur de la variable aléatoire Θ, donc les composantes du vecteur des données ne seront pas en général indépendantes (quoi qu'elles sont généralement supposées indépendantes conditionnellement à θ).

I.1.2 Echangeabilité

L'échangeabilité est la notion de base dans l'analyse bayésienne.

Définition I.1.1. *(Echangeabilité finie)*
Soit $p(y_1, ..., y_n)$ la densité de probabilité jointe (ou fonction de masse pour Y discrète) de $Y_1, ..., Y_n$. Si

$$p(y_1, ..., y_n) = p(y_{z(1)}, ..., y_{z(n)}) \tag{I.2}$$

pour toute permutation z de $\{1, ..., n\}$, alors la séquence $Y_1, ..., Y_n$ est dite échangeable.

L'échangeabilité est donc équivalente à la condition que la densité jointe des données reste la même sous réordonnencement des indices des données.
Remarque Une séquence infinie de variables aléatoires $Y_1, Y_2, ...$ est dite échangeable si n'importe quelle sous-suite finie est échangeable.

Si $Y_1, ..., Y_n$ sachant θ sont iid et $\theta \sim \pi(\theta)$ alors $Y_1, ..., Y_n$ sont échangeables

(voir [10]).

Réciproquement, si les données sont échangeables, elles sont conditionnellement à un certain θ iid. En effet, c'est l'une des conséquences essentielles du théorème de de Finetti.

Théorème I.1.1. *(Représentation de de Finetti) (voir [10])*
Soit $Y_1, Y_2, ...$ une séquence de variables aléatoires échangeables. Alors il existe une mesure de probabilté π et un ensemble Θ tels que

$$p(y_1, ..., y_n) = \int_\Theta \{\prod_{i=1}^{n} p(y_i/\theta)\}\pi(\theta)d\theta \tag{I.3}$$

L'échangeabilité est souvent interprétée comme la version bayésienne de l'hypothèse iid qui est fondamentale dans beaucoup de modèles statistiques. Il s'agit d'une hypothèse raisonnable lorsque les Y_t représentent les résultats des experiences répétées dans des conditions similaires.

I.2 Inférence bayésienne

Le théorème de Bayes est le fondement de l'analyse statistique bayésienne.

Théorème I.2.1. *Soit A et B_0 deux évènements, de probabilités non nulles, alors*

$$P(B_0/A) = \frac{P(A/B_0)P(B_0)}{P(A)} \tag{I.4}$$

Si $(B_n)_{n\in N}$ est un système complet d'évènements alors on a :

$$P(B_0/A) = \frac{P(A/B_0)P(B_0)}{\sum_{n=0}^{+\infty} P(A/B_n)P(B_n)} \tag{I.5}$$

Soit X et Y deux variables aléatoires discrètes à valeurs respectives dans $\mathcal{X} = \{(x_i), i \in N\}$ et $\mathcal{Y} = \{(y_i), i \in N\}$ alors on a comme application directe du théorème précédent

$$\begin{aligned} P(Y = y_j/X = x_i) &= \frac{P(X = x_i/Y = y_j)P(Y = y_j)}{P(X = x_i)} \\ &= \frac{P(X = x_i/Y = y_j)P(Y = y_j)}{\sum_{n=0}^{+\infty} P(X = x_i/Y = y_n)} \end{aligned}$$

On a une version continue de ce théorème dans le cas où les deux variables sont absolument continues de densités respectives f_X et f_Y et de densité conjointe $f_{X,Y}$:

$$f(y/X=x) = \frac{f(x/Y=y)f_Y(y)}{f_X(x)} = \frac{f(x/Y=y)f_Y(y)}{\int f(x/Y=y)f_Y(y)dy} \tag{I.6}$$

Dans un cadre bayésien on a deux variables aléatoires Y et Θ et en utilisant le théorème de Bayes la loi de Θ sachant $Y=y$, dite loi a posteriori de Θ, s'écrit :

$$\pi(\theta/Y=y) = \frac{f_\theta(y)\pi(\theta)}{\int_\Theta f_\theta(y)\pi(\theta)d\theta} \tag{I.7}$$

Le numérateur est la loi du couple (Y,Θ).
Le dénominateur est la loi marginale de Y, dite loi prédictive notée f.

$$f(y) = \int_\Theta f_\theta(y)\pi(\theta)d\theta \tag{I.8}$$

Le dénominateur $\int_\Theta f(y/\theta)\pi(\theta)d\theta$ est alors une constante qui est interprétée comme une constante de normalisation C avec $C = \dfrac{1}{\int_\Theta f_\theta(y)\pi(\theta)d\theta}$ afin que $CL(\theta,y)\pi(\theta)$ soit une densité de probabilité. Elle est dite constante de proportionnalité.
Une difficulté dans les problèmes de l'inférence bayésienne est le calcul de la constante de proportionnalité car elle s'exprime comme une intégrale qui n'est pas toujours facile à calculer, et dans beaucoup de cas on doit faire appel à des méthodes numériques qui nous en donne une approximation. Sa connaissance n'est cependant pas indispensable pour la détermination de la loi a posteriori.
On a

$$\pi(\theta/Y=y) \propto L(\theta,y)\pi(\theta) \tag{I.9}$$

loi a posteriori $\propto$ vraisemblance $\times$ loi a priori.

où le signe $\propto$ représente la proportionnalité.
En conclusion la densité a posteriori dépend essentiellement de la vraisemblance et de la loi a priori.

I.2.1 Distributions a priori

La distribution a priori $\pi(\theta)$ résume la connaissance a priori de θ i.e la connaissance de θ obtenue avant l'observation des données. Le choix d'une

distribution a priori varie d'une personne à l'autre car liée à des considérations subjectives. La principale objection à l'inférence bayésienne est que de tels choix subjectifs de la loi a priori influent nécessairement sur la conclusion générale de l'analyse. Les bayésiens argumentent que l'information a priori est toujours disponible et que l'aspect subjectif est avantageux.
De nos jours, beaucoup de recherches en statistique bayésienne sont axées vers la recherche de règles formelles pour des choix plus objectifs de ces lois. Les distributions a priori peuvent être divisées en deux types : les distributions a priori informatives qui résument l'information existante obtenue par exemple à partir d'expériences précédentes, et distributions a priori non-informatives lorsqu'on a peu ou pas d'information a priori sur le paramètre.

La famille de lois la plus utilisée dans le cas informatif est la famille des lois conjuguées naturelles.

Lois conjuguées naturelles

Supposons que la loi de Y, f_θ appartienne à une famille des lois F.

Définition I.2.1. *La famille de lois F^c sur Θ est dite conjuguée naturelle de la famille F si :*

$$\Pi \in F^c, y \in \mathcal{Y} \Rightarrow \Pi_{/Y=y} \in F^c \tag{I.10}$$

Définition I.2.2. *La famille des lois $F = \{f_\theta(.)/\theta \in \Theta\}$ de R^p est dite appartenir à la famille des lois exponentielles d'ordre k s'il existe des fonctions $h, t_i (1 \leq i \leq k)$ de R^p dans R et c, α_i $(1 \leq i \leq k)$ de Θ dans R telles que :*

$$f_\theta(y) = c(\theta)h(y) \exp\left\{\sum_{i=1}^{k} t_i(y)\alpha_i(\theta)\right\} \forall y \in R^p \tag{I.11}$$

Si $k = 1$, on parle de famille des lois exponentielles d'ordre 1, on a dans ce cas $f_\theta(y) = c(\theta)h(y) \exp\{t(y)\alpha(\theta)\}$. $\alpha_i(\theta) = \psi_i$ sont dits paramètres naturels. Ce changement de variables permet de réécrire (I.11) sous la forme :

$$f_\psi(y) = c'(\psi)h(y) \exp\left\{\sum_{i=1}^{k} \psi_i t_i(y)\right\} \forall y \in R^p \tag{I.12}$$

Théorème I.2.2. *La famille des lois*

$$F^c = \left\{c(\theta)\beta_0 \exp\left\{\sum_{i=1}^{k} \beta_i \alpha_i(\theta)\right\} / \beta_i \in R, 1 \leq i \leq k\right\} \tag{I.13}$$

est la famille des lois conjuguées naturelles de F.

Exemple I.2.1. ***Modèle Binomial***
Soit Y une variable aléatoire de loi $B(n,\theta)$, donc

$$f(y/\theta) = C_n^y \theta^y (1-\theta)^{n-y} \tag{I.14}$$

qui peut être aussi écrite de la façon

$$f(y/\theta) = C_n^y (1-\theta)^n \exp\left\{ y \log(\frac{\theta}{1-\theta}) \right\} \tag{I.15}$$

donc cette loi appartient à la famille des lois exponentielles avec $c(\theta) = (1-\theta)^n$ et $\alpha(\theta) = \log(\frac{\theta}{1-\theta})$, en appliquant le théorème précédent, la loi a priori est donnée par :

$$\pi(\theta) = (1-\theta)^n \exp\{\beta_1 \log \frac{\theta}{1-\theta}\} = \theta^{\beta_1}(1-\theta)^{n-\beta_1} \tag{I.16}$$

qui est proportionnelle à une loi Beta sur $[0,1]$ notée $Be_{[0,1]}(a,b)$ de densité

$$\pi(\theta) = \frac{1}{B(a,b)} \theta^{a-1}(1-\theta)^{b-1} 1_{[a,b]}(\theta) \tag{I.17}$$

La loi a posteriori est donnée par :

$$\begin{aligned} \pi(\theta/Y=y) &= \frac{C_n^y \theta^y (1-\theta)^{n-y} \dfrac{1}{B(a,b)} \theta^{a-1}(1-\theta)^{b-1}}{\int_0^1 C_n^y \theta^y (1-\theta)^{n-y} \dfrac{1}{B(a,b)} \theta^{a-1}(1-\theta)^{b-1} d\theta} \\ &= \frac{1}{B(a+y, b+n-y)} \theta^{a+y-1}(1-\theta)^{n+b-y-1} \end{aligned}$$

donc la densité a posteriori est de loi $B_{[0,1]}(a+y, b+n-y)$

Exemple I.2.2. ***Modèle Gaussien (variance connue)***
On considère le cas où on a un vecteur de n observations $y = (y_1, ..., y_n)'$, tel que $y_i \sim^{iid} N(\mu, \sigma^2)$, avec σ^2 connue, alors la vraisemblance est

$$L(\mu; y, \sigma^2) \propto f(y/\mu) \propto \prod_{i=1}^{n} \exp\left[\frac{-(y_i-\mu)^2}{2\sigma^2} \right] \tag{I.18}$$

Proposition I.2.3. *(voir [11], chapitre2)*
Soit $Y_i \sim^{iid} N(\mu, \sigma^2), i = 1, ..., n$, avec σ^2 connue, et $y = (y_1, ..., y_n)'$.
Si a priori $\mu \sim N(\mu_0, \sigma_0^2)$, alors

$$\mu/y \sim N\left(\frac{\dfrac{\mu_0}{\sigma_0^2} + \bar{y}\dfrac{n}{\sigma^2}}{\dfrac{1}{\sigma_0^2} + \dfrac{n}{\sigma^2}}, \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right)^{-1} \right) \tag{I.19}$$

Exemple I.2.3. ***Modèle Gaussien (moyenne connue)***
On considère un vecteur de n observations $y = (y_1, ..., y_n)'$, *tel que* $y_i \sim^{iid} N(\mu, \sigma^2)$, *avec* μ *connue. On prend comme loi a priori pour* σ^2 *une Inverse-Gamma* $IG(a, b)$ *donc la loi a posteriori de* σ^2 *est donnée par :*

$$\pi(\mu/Y = y) \propto \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{1}{2}\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{\sigma^2}} \frac{e^{-\frac{b}{\sigma^2}}}{(\sigma^2)^{a+1}}$$
$$\propto e^{-\frac{1}{2}(\frac{\sum_{i=1}^{n}(y_i-\mu)^2+2b}{\sigma^2})} \frac{1}{(\sigma^2)^{a+1+\frac{n}{2}}}$$

Donc la loi a posteriori est une $IG(a', b')$ *avec* $a' = a + 1 + \frac{n}{2}$ *et* $b' = \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2 + b$.

Exemple I.2.4. ***Modèle Gaussien (moyenne et variances inconnues)***
Le cas le plus courant en pratique, est lorsque la moyenne et la variance sont inconnues. Donc le vecteur de paramètres dans ce cas est $\theta = (\mu, \sigma^2)'$. *La densité a priori dans ce cas est bivariée, de densité conjointe* $p(\theta) = p(\mu, \sigma^2)$ *et de densité a posteriori* $p(\mu, \sigma^2/y)$.
La vraisemblance est définie par rapport aux deux paramètres inconnus, c'est à dire :

$$L(\mu, \sigma^2; y) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp[\frac{-(y_i - \mu)^2}{2\sigma^2}] \tag{I.20}$$

Pour obtenir une loi a priori conjuguée pour $\theta = (\mu, \sigma^2)'$ *en respectant la vraisemblance normale de l'équation précédente, on factorise la densité a priori conjointe* $p(\mu, \sigma^2)$ *comme le produit de la densité a priori conditionnelle de* μ, $p(\mu/\sigma^2)$ *et la densité a priori de* σ^2, $p(\sigma^2)$.
C'est à dire, $p(\mu, \sigma^2) = p(\mu/\sigma^2)p(\sigma^2)$ *où*

$$\mu/\sigma^2 \sim N(\mu_0, \sigma^2/n_0) \tag{I.21}$$

$$\sigma^2 \sim IG(\nu_0/2, \nu_0\sigma_0^2/2) \tag{I.22}$$

Proposition I.2.4. *(voir [11], chapitre2)*
Soit $y_i \sim^{iid} N(\mu, \sigma^2), i = 1, ..., n$ *et soit* $y = (y_1, ..., y_n)'$. *Si* $\theta = (\mu, \sigma^2)'$ *a une densité a priori Normale/Inverse-Gamma avec les paramètres* μ_0, n_0, ν_0 *et* σ_0^2, *alors la densité a posteriori de* θ *est aussi une Normale/Inverse-Gamma avec les paramètres* μ_1, n_1, ν_1 *et* σ_1^2, *où*

$$\mu_1 = \frac{n_0\mu_0 + n\bar{y}}{n_0 + n}$$

$$n_1 = n_0 + n$$

$$\nu_1 = \nu_0 + n$$

$$\nu_1\sigma_1^2 = \nu_0\sigma_0^2 + S + \frac{n_0 n}{n_0 + n}(\mu_0 - \bar{y})^2$$

où $S = \sum_{i=1}^{n}(y_i - \bar{y})^2$.

C'est à dire

$$\mu/\sigma^2, y \sim N(\mu_1, \sigma^2/n_1)$$

$$\sigma^2/y \sim IG(\frac{\nu_1}{2}, \frac{\nu_1\sigma_1^2}{2})$$

Lois a priori non informatives

Dans beaucoup de situations, on ne dispose pas d'informations a priori (ou très peu) sur le paramètre d'intêret. Il n'est alors pas possible de spécifier une loi a priori raisonnable et fondée. Intuitivement, la loi uniforme apparait comme celle qui traduit le mieux l'état d'indifférence, ou d'équiprobabilité, à l'égard des différentes valeurs possibles du paramètre. Historiquement, c'est d'ailleurs la première loi proposée pour ce cas. Son utilisation pose cependant plusieurs problèmes, parmi lesquels :

- elle n'est pas définie sur des ensembles non bornés :R_+, R, R^2 etc....
- elle n'est pas invariante par reparamétrisation.

Supposons que $\Theta = [0,1]$ et que $\theta \rightsquigarrow U[0,1]$. Déterminons la loi de θ^2. On a $F_{\theta^2}(x) = P[\theta^2 < x] = P[\theta < \sqrt{x}] = F_\theta(\sqrt{x}) \Longrightarrow f_{\theta^2}(x) = \frac{1}{2\sqrt{x}} f_\theta(\sqrt{x}) = \frac{1}{2\sqrt{x}}\mathbf{1}_{[0,1]}(x)$ Ainsi θ^2 ne suit pas une loi uniforme, sa loi est dans un certain sens informative car elle discrimine entre les différentes valeurs qu'elle peut prendre. On a par exemple $P[\theta^2 < \frac{1}{4}] = \frac{1}{2} \neq P[\frac{3}{4} < \theta^2 < 1] = 1 - \frac{\sqrt{3}}{2}$ ce qui n'est pas logique.

La recherche de lois pouvant traduire la situation de non-information, ou de vague information, a constitué l'un des sujets de recherche majeurs en statistique bayésienne et a aboutit à plusieurs :

- utilisation de lois impropres.

– utilisation du principe d'invariance pour certaines transformations.
– lois non informatives au sens de Jeffreys, etc...

Lois impropres

L'utilisation de telles lois comme lois a priori est justifié par le fait que dans beaucoup de cas les lois a posteriori, auxquelles elles conduisent sont propres.

Définition I.2.3. *On appelle loi impropre une mesure non finie. Si elle admet une densité f, on a*

$$\int_{\Theta} f(\theta)d\theta = +\infty \tag{I.23}$$

`Exemples:`
– $\Theta = R, \pi(\theta) = c$, c : constante réelle fixée.
– $\Theta = R_+, \pi(\theta) = \dfrac{1}{\theta}$

Exemple I.2.5. ***Modèle gaussien***
a) θ inconnu, σ^2 connu.

$$\begin{aligned}
\pi(\theta) = c \Longrightarrow \pi(\theta/X = x) &\propto c\frac{1}{(\sqrt{2\pi}\sigma)^n}e^{-\frac{1}{2}\frac{\sum_{i=1}^n (x_i-\theta)^2}{\sigma^2}} \\
&= \frac{1}{(\sqrt{2\pi}\sigma)^n}e^{-\frac{1}{2}\frac{(n-1)S^2+n(\theta-\bar{x})^2}{\sigma^2}} \\
&\Longrightarrow \pi(\theta/X = x) \propto c\frac{1}{(\sqrt{2\pi}\sigma)^n}e^{-\frac{1}{2}\frac{\sum_{i=1}^n n(\theta-\bar{x})^2}{\sigma^2}} \\
&\Longrightarrow \pi(\theta/X = x) \sim N(\bar{x}, \frac{\sigma^2}{n})
\end{aligned}$$

La loi a posteriori est donc une mesure de probabilité.
b) θ connu, σ^2 inconnu.

$$\begin{aligned}
\pi(\sigma^2) \quad = \quad \frac{1}{\sigma^2} \Longrightarrow \pi(\sigma^2/X = x) &= \frac{1}{\sigma^2}\frac{1}{(\sqrt{2\pi}\sigma)^n}e^{-\frac{1}{2}\frac{\sum_{i=1}^n (x_i-\theta)^2}{\sigma^2}} \\
\Longrightarrow \quad \pi(\sigma^2/X = x) &\sim IG(\frac{n}{2}, \frac{\sum_{i=1}^n (x_i-\theta)^2}{2})
\end{aligned}$$

La loi a posteriori est ici aussi une mesure de probabilité.

Lois non informatives de Jeffreys

Définition I.2.4. *Soit $\Theta \subset R$, on appelle loi non informative au sens de Jeffreys une loi π définie par :*

$$\pi(\theta) \propto \sqrt{I(\theta)} \tag{I.24}$$

où $I(\theta)$ est la quantité d'information de Fisher ramenée par une observation de Y sur θ qui est donnée par :

$$I(\theta) = E_\theta(\frac{\partial \log f(y/\theta)}{\partial \theta})^2 \tag{I.25}$$

Cette loi a priori vérifie effectivement l'invariance par reparametrisation, puisque

$$I(\theta) = I(h(\theta)h'(\theta))^2 \tag{I.26}$$

Le choix d'une telle loi se justifie par le fait que $I(\theta)$ est acceptée comme une mesure de la quantité d'information apportée par le modèle sur θ. Il est donc normal que les valeurs de θ pour lesquelles $I(\theta)$ est plus grande soient plus probables pour la loi a priori.
La loi a priori de Jeffreys est non informative car le choix de π ne fait pas intervenir d'autre information que celle apportée par les observations (au travers de $f(y/\theta)$. En ce sens, cette loi est non informative.

Exemple I.2.6. *Si $Y \sim B(n, \theta)$, on a*

$$f(y/\theta) = C_n^y \theta^y (1-\theta)^{n-y} \tag{I.27}$$

Dans ce cas l'information de Fisher est

$$I(\theta) = n[\frac{1}{\theta} + \frac{1}{1-\theta}] = \frac{n}{\theta(1-\theta)} \tag{I.28}$$

Donc la loi de Jeffreys pour ce problème est

$$\pi(\theta) \propto [\theta(1-\theta)]^{-1/2} \tag{I.29}$$

qui est une loi Beta $Be(1/2, 1/2)$

Pour les différents cas de la loi normale les lois de Jeffreys sont les lois obtenues précédemment.

I.2.2 Distributions a posteriori

La distribution a posteriori $\pi(\theta/y)$ résume l'information de θ apportée par les données et les considérations a priori. Elle est donnée par le théorème de Bayes qui est essentiellement une conséquence triviale de la définition des probabilités conditionnelles

$$\pi(\theta/y) = \frac{f(y/\theta)\pi(\theta)}{f(y)} \tag{I.30}$$

$$\propto f(y/\theta)\pi(\theta) \tag{I.31}$$

c'est a dire la densité a posteriori est proportionnelle au produit de la vraisemblance et la densité a priori. Sur la base de cette densité, toute caractéristique concernant θ peut être obtenue par des manipulations de base de la théorie des probabilités. Aucune approximation statistique ni asymptotique n'est nécessaire. Toutefois, pour des modèles complexes, les approximations numériques deviennent souvent nécessaires.

I.2.3 Estimation ponctuelle bayésienne

Soit un problème de décision avec Θ espace des états de la nature et A un espace d'actions.

Définition I.2.5. *(Fonction perte)*
On appelle fonction perte toute application

$$l : \Theta \times A \to R \tag{I.32}$$

$\forall\theta \in \Theta$ et $\forall a \in A$, $l(\theta, a)$ représente la perte subie lorsque θ est le vrai état de la nature et que c'est l'action a qui a été choisie.
Rappelons que dans l'approche bayésienne, θ est représenté par une fonction de densité a priori $\pi(\theta)$, et en particulier, après l'observation des données y, θ est représenté par une densité a posteriori $\pi(\theta/y)$.

Définition I.2.6. *(Perte bayésienne moyenne)*
Si $\pi(\theta)$ est la densité a priori de $\theta \in \Theta$, alors la perte bayésienne moyenne d'une action a est :

$$\varrho(\pi, a) = E[l(\theta, a)] = \int_{\Theta} l(\theta, a)\pi(\theta)d\theta \tag{I.33}$$

Un cas spécial est quand la densité π de la définition I.2.6 est une densité a posteriori :

Définition I.2.7. *(Perte a posteriori moyenne)*
Etant donné la densité a posteriori de θ, $\pi(\theta/y)$, la perte a posteriori moyenne d'une action a est

$$\varrho(\pi(/y), a) = \int_{\Theta} l(\theta, a)\pi(\theta/y)d\theta \tag{I.34}$$

Une règle de décision est dite bayésienne si elle conduit à choisir une action a parmi toutes les actions A qui minimise la perte a posteriori moyenne. Dans le contexte spécifique de l'estimation ponctuelle, on a $A = \Theta$ et la règle de décision s'appelle alors estimateur bayésien.
Plusieurs fonctions pertes peuvent être utilisées.

Exemples de fonctions perte
Perte quadratique
Si $\theta \in \Theta \subset R$ est un paramètre d'intérêt, et $\tilde{\theta}$ est un estimateur de θ, alors $l(\theta, \tilde{\theta}) = (\theta - \tilde{\theta})^2$ est la perte quadratique.

Proposition I.2.5. *Sous la perte quadratique l'estimateur bayésien de θ est la moyenne de la densité a posteriori, c'est à dire*

$$\tilde{\theta} = E(\theta/y) = \int_{\Theta} \theta\pi(\theta/y)d\theta \tag{I.35}$$

Démonstration. La perte quadratique implique que la perte a posteriori moyenne est :

$$\varrho(\theta, \tilde{\theta}) = \int_{\Theta} (\theta - \tilde{\theta})^2\pi(\theta/y)d\theta. \tag{I.36}$$

et on cherche à minimiser cette expression par rapport à $\tilde{\theta}$. En développant le carré on obtient :

$$\begin{aligned}\varrho(\theta, \tilde{\theta}) &= \int_{\Theta} \theta^2\pi(\theta/y)d\theta + \tilde{\theta}^2 \int_{\Theta} \pi(\theta/y)d\theta - 2\tilde{\theta} \int_{\Theta} \theta\pi(\theta/y)d\theta \\ &= \int_{\Theta} \theta^2\pi(\theta/y)d\theta + \tilde{\theta}^2 - 2\tilde{\theta}E(\theta/y)\end{aligned}$$

En dérivant par rapport à $\tilde{\theta}$, et en annulant la dérivée le résultat est établi. □

Ce résultat est aussi vérifié dans le cas où θ est vectoriel $\theta = (\theta_1, ..., \theta_k)'$. Dans ce cas plus général, on définit une fonction perte multidimentionnelle comme suit :

Perte quadratique multidimensionnelle
Si $\theta \in R^k$ est un paramètre, et $\tilde{\theta}$ est un estimateur de θ, alors la perte quadratique multidimensionnelle est $l(\theta, \tilde{\theta}) = (\theta - \tilde{\theta})' Q(\theta - \tilde{\theta})$ où Q est une matrice définie positive.

Proposition I.2.6. *Sous la perte quadratique multidimensionnelle la moyenne a posteriori* $E(\theta/y) = \int_\Theta \theta\pi(\theta/y)d\theta$ *est l'estimateur bayésien de* θ.

Démonstration. La perte a posteriori moyenne

$$\varrho(\theta, \tilde{\theta}) = \int_\Theta (\theta - \tilde{\theta})' Q(\theta - \tilde{\theta})\pi(\theta/y)d\theta$$

En différentiant par rapport à $\tilde{\theta}$ on obtient

$$2Q \int_\Theta (\theta - \tilde{\theta})\pi(\theta/y)d\theta$$

En posant la dérivée égale à zéro on obtient

$$\int_\Theta (\theta - \tilde{\theta})\pi(\theta/y)d\theta = 0 \tag{I.37}$$

ou

$$\int_\Theta \theta\pi(\theta/y)d\theta = \int_\Theta \tilde{\theta}\pi(\theta/y)d\theta \tag{I.38}$$

Le terme gauche de cette expression n'est que la moyenne de la densité a posteriori, $E(\theta/y)$, et donc

$$E(\theta/y) = \int_\Theta \tilde{\theta}\pi(\theta/y)d\theta = \tilde{\theta} \int_\Theta \pi(\theta/y)d\theta = \tilde{\theta}.$$

□

Perte linéaire
Si $\theta \in \Theta$ est un paramètre et $\tilde{\theta}$ est un estimateur de θ, alors la fonction perte linéaire est définie par :

$$l(\theta, \tilde{\theta}) = \begin{cases} k_0(\theta - \tilde{\theta}) & \text{si } \tilde{\theta} < \theta \\ k_1(\tilde{\theta} - \theta) & \text{si } \theta \leq \tilde{\theta} \end{cases}$$

Un cas spécial est la classe des fonctions pertes linéaires symétriques (i.e. $k_0 = k_1$). Une perte linéaire asymétrique résulte quand $k_0 \neq k_1$.

Proposition I.2.7. *Sous la perte linéaire l'estimateur de Bayes de* θ *est le quantile d'ordre* $k_0/(k_0+k_1)$ *de* $\pi(\theta/y)$, $\tilde{\theta}$ *est tel que* $\pi(\theta \leq \tilde{\theta}) = k_0/(k_0+k_1)$.

Démonstration. On cherche $\tilde{\theta}$ qui minimise

$$\begin{aligned}\varrho(\theta,\tilde{\theta}) &= \int_{\Theta} l(\theta,\tilde{\theta})\pi(\theta/y)d\theta \\ &= k_0 \int_{\tilde{\theta}<\theta} (\theta-\tilde{\theta})\pi(\theta/y)d\theta + k_1 \int_{\tilde{\theta}\leq\theta} (\tilde{\theta}-\theta)\pi(\theta/y)d\theta\end{aligned}$$

En dérivant cette expréssion par rapport à $\tilde{\theta}$ et en mettant le résultat égal à zéro on obtient ;

$$k_0 \int_{\tilde{\theta}<\theta} \pi(\theta/y)d\theta = k_1 \int_{\tilde{\theta}\leq\theta} \pi(\theta/y)d\theta$$

En ajoutant $k_0 \int_{\tilde{\theta}\leq\theta} \pi(\theta/y)d\theta$ des deux cotés de l'équation précédente on obtient $k_0 = (k_0+k_1)\int_{\tilde{\theta}\leq\theta} \pi(\theta/y)d\theta$ et donc on obtient $\int_{\tilde{\theta}\leq\theta} \pi(\theta/y)d\theta = k_0/(k_0+k_1)$ □

Notons qu'avec une perte linéaire symétrique, on obtient la médiane de la densité a posteriori comme estimateur de Bayes. Des fonctions perte asymétriques impliquent l'utilisation des quantiles autre que la médiane.

I.3 Conclusion

Dans ce chapitre, nous avons rappelé les principes de l'analyse statistique bayésienne et les principaux résultats concernant l'estimation en développant quelques exemples que nous utiliserons par la suite. Les distributions a posteriori sont souvent de formes non standards et les caractéristiques de leurs lois sont difficiles à calculer notamment lorsque le nombre de paramètres est supérieur à deux ou à trois. Les exemples développés dans la section 2 ont été choisis pour leur importance ainsi que pour la simplicité du calcul de leurs lois a posteriori. Les lois a priori choisies étaient des lois conjuguées, ou des lois non informatives simples, conduisant à des lois a posteriori ne posant aucune difficulté calculatoire. Mais dès qu'on sort de ce cadre, y compris en considérant des modèles classiques et naturels les lois a posteriori et leurs caractéristiques ne sont plus accessibles analytiquement d'où la nécessité de faire appel à des méthodes de simulation numériques pour calculer les différentes caractéristiques de la loi a posteriori. Dans le chapitre suivant nous introduisons les méthodes classiques de Monte Carlo et les méthodes de Monte Carlo par chaines de Markov.

Chapitre II

Méthodes MCMC

Quelque soit le problème d'inférence statistique, nous sommes souvent amenés à considérer des solutions numériques. Deux grandes classes de problèmes numériques se posent dans l'inférence statistique : les problèmes d'optimisation (maximisation de la vraisemblance) et les problèmes d'intégration (statistique bayésienne). Dans de nombreux cas il n'est pas toujours possible de calculer analytiquement les estimateurs associés quelque soit la méthode utilisée (maximum vraisemblance, méthode des moments, estimation bayésienne, etc). Une méthode générale est d'utiliser la simulation pour calculer les quantités d'intérêt. Dans le contexte de la théorie de décision, classique ou bayésienne, cette solution est naturelle puisque les risques et les estimateurs de Bayes entrainent des intégrales par rapport aux distributions de probabilité.
L'inférence statistique bayésienne est basée sur la détermination de la loi a posteriori et de ses caractéristiques (moyenne, mode, médiane, variance, quantiles, etc...). La majorité des caractéristiques d'intérêt s'écrivent sous la forme :

$$\int_{\Theta} g(\theta)\pi(\theta/y)$$

où $g(\theta)$ est une fonction de θ qui varie selon la caractérisque de la loi a posteriori souhaitée, par exemple $g(\theta) = \theta$ pour la moyenne de la loi a posteriori. Bien que pour le mode, il s'agit d'un problème d'optimisation, dans la majorité des autres cas il s'agit d'un calcul d'intégrales. Pour le faire numériquement plusieurs types de méthodes existent, basées sur la simulation : les méthodes classiques de Monte Carlo et les méthodes de Monte Carlo par chaines de Markov.

Les méthodes de Monte Carlo

Bien que cette méthode converge lentement, elle a comme intérêt d'être insensible à la dimension des problèmes étudiés (contrairement à des méthodes classiques d'analyse numérique qui ne sont performantes qu'en petite dimension).
Considérons par exemple le problème d'intégration numérique. Il s'agit d'approcher

$$I = \int_0^1 g(x)dx$$

la méthode de Monte Carlo consiste à écrire cette intégrale sous la forme :

$$I = E(g(U))$$

où U est une variable aléatoire suivant une loi uniforme sur $[0, 1]$. En utilisant la loi des grands nombres : si $(U_i)_{i\in\{1,...,n\}}$ est une suite de variables aléatoires indépendantes de loi uniforme sur $[0, 1]$ alors

$$\frac{1}{n}\sum_{i=1}^{n} g(U_i) \longrightarrow^{ps} E(g(U))$$

En d'autres termes, si $u_1, u_2, ..., u_n$ sont des nombres tirés au hasard dans $[0, 1]$ alors :

$\frac{1}{n}\sum_{i=1}^{n} g(u_i)$ est une approximation de $I = \int_0^1 g(x)dx$.

Prenons un exemple plus général d'une intégrale du type :

$$I = \int_0^1 g(x)f(x)dx$$

alors $I = E[g(X)]$ où X est une suite de variables aléatoires de densité f. Toujours par la loi des grands nombres, si $(X_i)_{i\in\{1,...,n\}}$ est une suite de variables aléatoires indépendantes de densité f,

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \longrightarrow^{ps} E(g(X))$$

et donc si $(x_1, x_2, ..., x_n)$ est une réalisation de $(X_1, X_2, ..., X_n)$, $\bar{g_n} = \frac{1}{n}\sum_{i=1}^{n} g(x_i)$ converge presque surement vers I.

Par ailleurs lorsque $g^2(X)$ a une espérance finie sous f, la variance asymptotique de cette approximation est

$$var(\bar{g}_n) = \frac{1}{n}\int_{\chi}(g(x) - E[g(X)])^2 f(x)dx \tag{II.1}$$

qui peut être estimée à partir de l'échantillon $(X_1, ..., X_n)$ par

$$v_n = \frac{1}{n^2}\sum_{i=1}^{n}[g(x_i) - \bar{g}_n]^2 \tag{II.2}$$

Ces méthodes nécessitent des réalisations de variables aléatoires indépendantes. Il existe des techniques classiques de génération de variables aléatoires :

Méthode d'inversion

Si X est une variable aléatoire de densité f, la fonction de répartition associée F vérifie

$$F(x) = \int_{-\infty}^{x} f(t)dt \tag{II.3}$$

donc, si on pose $U = F(X)$, la variable U est alors une variable aléatoire de loi uniforme sur $[0, 1]$ (voir [18]). En effet,

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

où F^{-1} est la fonction définie par :

$$F^{-1}(y) = inf\{x \in R/F(x) \leq y\}$$

donc on a :

1. $F(X) \rightsquigarrow U_{[0,1]}$
2. si $U \rightsquigarrow U_{[0,1]}$ alors $F^{-1}(U)$ suit la même loi que X.

Méthode d'acceptation-rejet

La méthode d'acceptation-rejet est utilisée lorsque la distribution d'intérêt $f(x)$ ne peut pas être échantillonnée facilement (voir [18]). Cette méthode exige la connaissance de la forme fonctionnelle de la densité d'intérêt f (densité cible) à une constante multiplicative près. On utilise en parallèle une densité g plus simple à échantillonner appelée densité instrumentale ou candidate. Les seules conditions sur cette densité sont :

– f et g ont des supports compatibles.
– Il existe une constante M telle que $f(x)/g(x) \leq M$ pour tout x.

Sous ces conditions, X de loi f peut être simulée comme suit : premièrement, on génère $y \sim g$ et on génère $u \sim U_{[0,1]}$.
Si $u \leq \dfrac{f(y)}{Mg(y)}$ alors on pose $x = y$. Sinon, on rejette y et u, puis on recommence la génération des valeurs candidates.

Algorithme

1. Générer $y \sim g$.
2. Générer $u \sim U_{[0,1]}$.
3. Accepter $x = y$ si $u \leq f(y)/Mg(y)$.
4. Sinon revenir à 1.

L'efficacité de cette méthode dépend de la fonction envellope g. Il y a deux but majeurs, premièrement de trouver une fonction qui satisfait $f(x) < Mg(x) \forall x$, et deuxièmement de trouver une fonction g assez similiaire à f pour avoir un taux d'acceptation important.
Cependant, on ne peut pas toujours avoir recours à ces techniques classiques. Quand les lois ne peuvent pas être simulées directement, une solution est de produire une chaîne de Markov $X_1, X_2, ..., X_N$ qui converge en distribution vers la loi objectif $f(.)$.

II.1 Les méthodes MCMC

Au cours de ces dernières années les méthodes statistiques de calcul ont été enrichies par une méthode de génération de réalisations de variables aléatoires de distribution donnée basée sur les chaines de Markov dont la distribution stationnaire est la distribution de probabilité d'intérêt. Cette classe de méthodes, appelée Monte Carlo Markov Chain ou simplement MCMC, ont été influentes dans la pratique moderne de la statistique bayésienne, où ces méthodes sont utilisées pour simuler les distributions a posteriori qui proviennent de l'analyse bayésienne.
Une méthode de Monte Carlo par chaîne de Markov (MCMC) est une méthode de simulation qui produit une chaîne de Markov $(\theta_n)_{n \in N}$ qui converge en distribution vers la distribution a posteriori (la loi objectif $f(.)$) de θ, sans calculer la constante de proportionnalité. A partir de cette chaine, différentes caractéristiques (moyenne, variance, quantiles, etc) de la distribution a posteriori peuvent être calculées.

La loi objectif de θ sachant que $X = x$, est de densité :

$$\pi(\theta/x) = \frac{\pi(\theta)f(x/\theta)}{\int \pi(\theta)f(x/\theta)d\theta} \tag{II.4}$$

Très souvent l'intégrale qui apparaît au dénominateur dans l'équation II.4 n'est pas calculable, de sorte que la densité de la loi a posteriori, $\pi(\theta/x)$, est définie à un facteur multiplicatif près :

$$\pi(\theta/x) \propto \pi(\theta)f(x/\theta) \tag{II.5}$$

Les méthodes MCMC peuvent aussi être utilisées pour l'intégration $I = \int f(x)\pi(x)dx$, ces méthodes sont adaptées à l'intégration dans des espaces de grande dimension (typiquement, R^N avec N supérieur à 3 ou 4). En effet, dans ce cas, les méthodes MCMC présentent un certain nombre d'avantages par rapport aux méthodes déterministes (méthodes d'intégration par quadrature comme la méthode des trapèzes).
La manière dont fonctionnent les méthodes Monte Carlo par Chaine de Markov est simple à décrire. Etant donnée une densité cible f, on construit un noyau de transition q utilisé pour générer une chaine de Markov $X^{(t)}$, de telle sorte que la loi limite de $X^{(t)}$ soit f.
A priori, la difficulté devrait provenir de la construction du noyau q associé à une densité arbitraire f. Mais il existe des méthodes universelles permettant de construire ces noyaux, au sens où elles sont valides quelque soit f. L'algorithme de Metropolis Hastings est un exemple de ces méthodes universelles.

II.1.1 Echantillonnage de Metropolis

L'algorithme de Metropolis est un algorithme utilisé pour simuler une variable aléatoire θ définie par sa densité de probabilité $\pi(.)$. Comme l'algorithme d'acceptation rejet, l'algorithme de Metropolis réutilise, par un mécanisme d'acceptation rejet, des variables aléatoires Y produites selon une loi instrumentale.

L'idée est de générer des valeurs de θ, le paramètre d'interêt à partir d'une distribution candidate et corriger ces valeurs de telle sorte que :
La distribution candidate est généralement dépendante de la dernière valeur de θ générée mais indépendante de toutes les autres valeurs précédentes de θ pour obéir à la propriété de Markov. La méthode consiste à générer de nouvelles valeurs à chaque instant à partir de la distribution candidate mais les accepter uniquement lorsqu'elles satisfont un certain critère.
La distribution candidate q doit être symétrique :

$$q(y/x) = q(x/y)$$

Algorithme

On simule une suite $\theta_1, \theta_2, ...$ de la façon suivante :

1. Séléctionner une valeur initiale pour θ notée θ_0.
2. Générer une valeur $\theta_* \rightarrow q(\theta/\theta_t)$.
3. Calculer le ratio de Metropolis :

$$\alpha(\theta_t, \theta_*) = min\left\{1, \frac{\pi(\theta_*)}{\pi(\theta_t)}\right\} \tag{II.6}$$

4. Simuler $u_t \sim U[0,1]$. La valeur obtenue est notée u_t. Si $u_t < \alpha(\theta_t, \theta_*)$ alors $\theta_{t+1} = \theta_*$, sinon $\theta_{t+1} = \theta_t$.
5. Retouner en 2.

Le principe de l'algorithme de Metropolis est de perturber de façon stochastique la valeur courante θ_t, d'accepter aléatoirement le résultat θ_* de cette perturbation aléatoire.
Supposons que la perturbation aléatoire appliquée à θ_t ait conduit à une valeur θ_* telle que $\pi(\theta_*) > \pi(\theta_t)$. Dans ce cas $\alpha(\theta_t, \theta_*) = 1$ et θ_* est accepté avec probabilité 1.
Si, au contraire, la perturbation aléatoire a conduit à une valeur θ_* telle que $\pi(\theta_*) < \pi(\theta_t)$ on ne refuse pas systématiquement. Le ratio $\alpha(\theta_t, \theta_*) = \dfrac{\pi(\theta_*)}{\pi(\theta_t)}$ est strictement inférieur à 1 mais non nul ; on peut accepter θ_* avec une probabilité d'autant plus forte que la dégradation de $\pi(.)$ a été peu importante.

II.1.2 Echantillonnage de Metropolis-Hastings

Hastings (1970) a généralisé l'algorithme de Metropolis en utilisant des distributions candidates non symétriques.

Algorithme

On simule une suite $\theta_1, \theta_2, ...$ de la façon suivante :

1. Séléctionner une valeur initiale pour θ notée θ_0.
2. Générer une valeur $\theta_* \rightarrow q(\theta/\theta_t)$.
3. Calculer le ratio de Metropolis-Hastings :

$$\alpha(\theta_t, \theta_*) = min\left\{1, \frac{\pi(\theta_*)q(\theta_t/\theta_*)}{\pi(\theta_t)q(\theta_*/\theta_t)}\right\} \tag{II.7}$$

4. Simuler $u_t \sim U[0,1]$. La valeur obtenue est notée u_t. Si $u_t < \alpha(\theta_t, \theta_*)$ alors $\theta_{t+1} = \theta_*$, sinon $\theta_{t+1} = \theta_t$.

5. Retouner en 2.

Remarque : le point 4 s'énonce, de façon équivalente :
Accepter $\theta_{t+1} = \theta_*$ avec probabilité $\alpha(\theta_t, \theta_*)$, sinon poser $\theta_{t+1} = \theta_t$.

On peut remarquer que dans l'algorithme de Metropolis-Hastings (MH), la loi objectif $\pi(.)$ n'intervient que dans le calcul du ratio de Metropolis-Hastings. Il résulte de son expression que la densité $\pi(.)$ de la loi objectif peut être définie à un facteur près. En effet, $\pi(.)$ apparaît à la fois au numérateur et au dénominateur dans II.7. L'algorithme de Metropolis-Hastings est donc bien adapté aux problèmes d'estimation bayésienne.
Dans les modèles à plusieurs paramètres cet algorithme peut être utilisé de plusieurs manières.

- Premièrement, θ peut être considéré comme un vecteur contenant tous les paramètres d'intérêt et une distribution candidate multivariée doit être utilisée.
- Deuxièmement, l'algorithme précédent peut être utilisé séparément pour chaqu'un des paramètres inconnus θ_i. C'est à dire à l'instant t, générer un nouveau θ_{1i}, ensuite θ_{2i} jusqu'a ce que tous les paramètres soient mis à jours et faire de même à l'étape $t + 1$.
- Troisièmement, une méthode combinée où les paramètres sont mis à jours en blocs certain en multivarié et d'autres en univarié.

Cas particuliers

Algorithme de Metropolis-Hastings indépendant
L'algorithme de Metropolis-Hastings utilise une distribution q qui dépend uniquement de l'état présent de la chaîne. Si on a besoin d'une distribution q qui soit indépendante de l'état précédent de la chaîne ($q(y/x) = g(y)$), on fait appel a un cas particulier de l'algorithme qui est l'algorithme de Metropolis-Hastings indépendant.
On simule une suite $\theta_1, \theta_2, \theta_3, ...$ de la façon suivante :

1. Générer $\theta_* \sim g(y)$.
2. Calculer le ratio de Metropolis-Hastings :

$$\alpha(\theta_t, \theta_*) = min\left\{1, \frac{\pi(\theta_*)g(\theta_t)}{\pi(\theta_t)g(\theta_*)}\right\} \tag{II.8}$$

3. Simuler $u_t \sim U[0, 1]$. Si $u_t < \alpha(\theta_t, \theta_*)$ alors $\theta_{t+1} = \theta_*$, sinon $\theta_{t+1} = \theta_t$.

Dans l'algorithme de Metropolis-Hastings indépendant θ_* est généré indépendamment de θ_t ; cependant θ_t et θ_{t+1} ne sont pas indépendants car la valeur θ_t apparaît dans le ratio de Metropolis-Hastings II.8. Cette méthode apparaît

comme une généralisation de la méthode acceptation rejet avec comme loi instrumentale g .
L'algorithme de Metropolis-Hastings indépendant et la méthode d'acceptation rejet sont deux méthodes de simulation qui ont des ressemblances entre elles qui permettent une comparaison :
-L'échantillon de la méthode d'acceptation rejet est iid tandis que celui de l'algorithme de Metropolis-Hastings ne l'est pas. Bien que les θ_* sont générés indépendamment, l'échantillon résultant n'est pas iid car la probabilité d'acceptation de θ_* dépend de θ_t.

L'algorithme de Metropolis-Hastings à marche aléatoire
Une approche plus naturelle pour la construction d'une proposition de Metropolis-Hastings vise donc à prendre en compte la valeur simulée précédemment pour générer la valeur suivante. La mise en oeuvre de cette idée est de simuler θ_* a partir de :

$$\theta_* = \theta_t + \varepsilon_t \tag{II.9}$$

où ε_t est une perturbation aléatoire de la distribution g indépendante de θ_t, par exemple une distribution uniforme ou normale. En termes de l'algorithme général de Metropolis-Hastings la densité $q(y/x)$ est dans ce cas de la forme $g(y-x)$. La chaîne de Markov associée est une marche aléatoire, si g est symétrique autour de zéro on tombe sur le cas de l'algorithme de Metropolis-Hastings symétrique mais dans ce cas la chaîne de Markov n'est pas une marche aléatoire.

Algorithme
Étant donné θ_t

1. Générer $\theta_* \rightarrow g(\theta - \theta_t)$
2. Prendre : $\theta_{t+1} = \theta_*$ avec une probabilité $\alpha(\theta_t, \theta_*) = min\left\{1, \dfrac{\pi(\theta_*)}{\pi(\theta_t)}\right\}$ et $\theta_{t+1} = \theta_t$ ailleurs.

Algorithme de Metropolis Hastings multi-blocs En pratique quand la dimension de θ est grande, il peut être difficile de construire un algorithme de M-H avec un seul bloc qui converge rapidement vers la distribution d'intérêt. Dans tels cas, il est utile de diviser la variable en des blocs de plus petites tailles et de construire une chaine de Markov avec ces petits blocs.
Supposons, pour illustrer celà, que θ est divisé en deux blocs (θ_1, θ_2). Pour chaque bloc soit $q_1(\theta_1, \theta_{*1}/\theta_2); q_2(\theta_2, \theta_{*2}/\theta_1)$ les densités candidates correspondantes. On définit aussi (par analogie avec le cas uni-bloc)

$$\alpha(\theta_1, \theta_{*1}/\theta_2) = min\{1, \frac{\pi(\theta_{*1}/\theta_2)q_1(\theta_{*1}, \theta_1/\theta_2)}{\pi(\theta_1/\theta_2)q_1(\theta_1, \theta_{*1}/\theta_2)}\} \tag{II.10}$$

et

$$\alpha(\theta_2, \theta_{*2}/y, \theta_1) = min\{1, \frac{\pi(\theta_{*2}/\theta_1)q_2(\theta_{*2}, \theta_2/\theta_1)}{\pi(\theta_2/\theta_1)q_2(\theta_2, \theta_{*2}/\theta_1)}\} \tag{II.11}$$

comme le ratio de M-H pour chacun des deux blocs conditionnellement à l'autre bloc. Les densités $\pi(\theta_1/\theta_2)$ et $\pi(\theta_1/\theta_2)$ sont les densités conditionnelles.

Algorithme

1. Spécifier une valeur initiale $\theta^0 = (\theta_1^{(0)}, \theta_2^{(0)})$
2. Répéter pour $j = 1, 2, ..., M$
   Répéter pour $k = 1, 2$
   i. Proposer une valeur

$$\theta_{*k} \sim q_k(\theta_k^{(j-1)}, ./\theta_{-k}) \tag{II.12}$$

   ii. Calculer le ration de M-H

$$\alpha((\theta_k^{(j-1)}, \theta_{*k}/y, \theta_{-k}) = min\{1, \frac{\pi(\theta_{*k}/\theta_{-k})q_k(\theta_{*k}, \theta_k^{(j-1)}, /\theta_{-k})}{h(\theta_k^{(j-1)}/\theta_{-k})q_k(\theta_k^{(j-1)}, \theta_{*k}, /\theta_{-k})}\}$$

   iii. mettre à jour le kème bloc comme suit
3. Retourner les valeurs $\{(\theta^1, \theta^2, ..., \theta^M\}$

Dans certains cas la distribution candidate n'est pas facile à déterminer, donc l'algorithme de M-H ne peut pas être utilisé, donc on doit faire appel à d'autres méthodes. Nous étudions dans la suite une autre classe de méthodes MCMC, très répandue, connue sous le nom d'échantillonnage de Gibbs. Cette méthode permet de décomposer des problèmes complexes en plusieurs problèmes plus simples, comme une suite de cibles en petite dimension.

II.1.3 Echantillonnage de Gibbs

L'échantillonneur de Gibbs est un cas particulier de l'algorithme Metropolis-Hastings. L'échantillonneur de Gibbs est mieux adapté aux problèmes où les distributions marginales des paramètres d'intérêt sont difficiles à calculer, mais les distributions conditionnelles de chaque paramètre sachant tous les autres paramètres et les données ont des formes standards. Par exemple, supposons que la distribution marginale a posteriori ne peut pas être obtenue à partir de la densité a posteriori conjointe analytiquement. Cependant, supposons que les distributions a posteriori conditionnelles ont des formes connues

et faciles à échantillonner, par exemple la distribution Normale ou Gamma. L'échantillonneur de Gibbs peut donc être utilisé pour échantillonner indirectement à partir de la distribution a posteriori marginale.
Si les distributions conditionnelles des paramètres ont des formes standards, alors on peut simuler facilement. Si ce n'est pas le cas et si la distribution conditionnelle n'a pas une forme standard, alors différentes méthodes peuvent être utilisées (comme les méthodes vues précédemment). L'avantage fondamental en utilisant une structure de Gibbs est qu'elle divise un modèle complexe en un grand nombre de cibles plus petites et plus simples.

Algorithme
L'algorithme d'échantillonnage de Gibbs est utilisé pour simuler une distribution $\pi(\theta)$ quand

1. θ admet une décomposition de la forme $\theta = (\theta_1, \theta_2, ..., \theta_N)$
2. Les lois conditionnelles $\pi_i(\theta_i/\theta_{i-1})$ sont simulables aisément.

En notant $\theta_{-i} = (\theta_1, \theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_N)$ et $\pi_i(\theta_i/\theta_{-i})$ la densité de la loi θ_i sachantθ_{-i}. Le principe de l'algorithme de Gibbs est de réactualiser « composante par composante ». Un passage de l'échantillonneur de Gibbs se décompose de la façon suivante :
sachant $\theta^t = (\theta_1^t, ..., \theta_N^t)$, générer θ^{t+1} comme suit :

- $\theta_1^{t+1} \leftarrow \pi_1(\theta_1/\theta_2^t, ..., \theta_N^t)$
- $\theta_2^{t+1} \leftarrow \pi_2(\theta_2/\theta_1^{t+1}, \theta_3^t, ..., \theta_N^t)$
- $\vdots$
- $\theta_i^{t+1} \leftarrow \pi_i(\theta_i/\theta_1^{t+1}, ..., \theta_{i-1}^{t+1}, \theta_{i+1}^t, ..., \theta_N^t)$
- $\vdots$
- $\theta_N^{t+1} \leftarrow \pi_N(\theta_N/\theta_1^{t+1}, ..., \theta_{N-1}^{t+1})$

Sous certaines conditions, la chaîne de Markov $(\theta^t)_{t=1,2,...}$ qui est produite par l'algorithme converge, en distribution, vers la loi jointe $\pi(.)$ quand t, le nombre d'itérations de l'algorithme, converge vers l'infini.

L'échantillonneur de Gibbs comme un cas spécial de l'algorithme de Metropolis-Hastings

En observant l'algorithme d'échantillonnage de Gibbs comme il a été écrit précédemment il n'est pas immédiatement clair que cet échantillonneur est un cas particulier de l'algorithme de Metropolis Hastings. En appliquant ce dernier avec comme distribution candidate d'un élément θ_i du vecteur θ

$$q(\theta_*/\theta_{t-1}) = \begin{cases} \pi(\theta_{*(i)}/\theta_{t-1(-i)}, y) & \text{si } \theta_{*(-i)} = \theta_{t-1(i)} \\ 0 & \text{sinon} \end{cases}$$

où $\theta_{(-i)}$ est le vecteur θ sans l'élément i. Alors le ratio $\alpha(\theta_t, \theta_*)$

$$\begin{aligned}\alpha(\theta_t, \theta_*) &= \frac{\pi(\theta_*/y)/q(\theta_*/\theta_{t-1})}{\pi(\theta_{t-1}/y)/q(\theta_{t-1}/\theta_*)} \\ &= \frac{\pi(\theta_*/y)/\pi(\theta_{*(i)}/\theta_{t-1(-i)}, y)}{\pi(\theta_{t-1}/y)/\pi(\theta_{t-1(i)}/\theta_{t-1(-i)}, y)} \\ &= \frac{\pi(\theta_{t-1(i)}/y)}{\pi(\theta_{t-1(i)}/y)} \\ &= 1\end{aligned}$$

C'est le même algorithme de l'échantillonneur de Gibbs décrit précédemment mais écrit cette fois-ci en format Metropolis-Hastings.
Le point faible de l'algorithme de Metropolis-Hastings est le choix de la loi instrumentale. Cette loi peut être très différente de la loi cible f, et l'algorithme risque alors de simuler "grossièrement" f. Ce n'est pas le cas pour l'échantillonneur de Gibbs, dont la loi instrumentale est directement déduite de la loi de f. Cependant, la structure composée de l'échantillonneur de Gibbs peut être une faiblesse. La liberté du choix de q dans l'algorithme de Metropolis-Hastings permet parfois de remédier à ce problème.
Pour conserver les avantages des deux méthodes, l'approche hybride est considérée.

II.1.4 Algorithmes hybrides

On appelle algorithme MCMC hybride ou Gibbs hybride une méthode MCMC combinant simultanément des étapes d'un échantillonneur de Gibbs avec des étapes d'algorithmes de Metropolis-Hastings.
Un échantillonneur de Gibbs peut être facilement étendu à des contextes où certaines distributions conditionnelles ne peuvent être simulées par les générateurs de variables aléatoires standards. Si dans un ensemble de distributions conditionnelles $f_1, ..., f_p$, une certaine densité f_i est non standard, dans ce cas on peut simuler cette densité en utilisant un algorithme de Metropolis-Hastings.

II.2 Convergence des méthodes MCMC

Dans les méthodes de chaines de Markov on ne s'intéresse pas à la convergence vers une valeur donnée mais on s'intéresse plutôt à la convergence vers une distribution, à savoir en statistique bayésienne vers la distribution jointe a posteriori d'intérêt. Il y a plusieurs points à aborder lors de l'étude de convergence vers une distribution. Premièrement, repérer le moment où la chaine commence à échantillonner à partir de la distribution a posteriori (la distribution stationnaire est atteinte). Deuxièmement, la taille de l'échantillon nécessaire pour donner des estimations d'une précision donnée. Toutefois, la simulation MCMC a deux défauts :

1. La distribution des échantillons, $p(\theta^{(i)})$ converge lentement vers la distribution cible.
2. Les échantillons sont dépendants.

L'idéal est que l'algorithme de simulation retourne des échantillons *iid* pour la distribution cible (a posteriori). Il est nécessaire d'avoir des diagnostics pour savoir si la chaine a convergé. Il existe pour cela plusieurs tests : visuels et statistiques. Dans ce qui suit, nous decrivons les outils les plus utilisés.

II.2.1 Résumé statistique

Une fois que la chaine de Markov est générée, les sorties sont des séquences de valeurs, une séquence pour chaque paramètre, supposée être à partir de la distribution a posteriori. A partir de chaque séquence de valeurs, on peut décrire le paramètre qu'elle représente via certaines caractéristiques numériques. On rappelle dans ce qui suit ces principales caractéristiques.

Mesures de location

Il y a trois estimateurs principaux qui peuvent être utilisés pour le paramètre d'interêt, θ, à savoir :

- La moyenne de l'échantillon
- La médiane
- Le mode

Mesures de dispersion

Il y a deux princiaux groupes de caractéristiques de la dispersion d'un échantillon :

- La variance et l'écart-type.
- Les quantiles.

II.2.2 Convergence et burn-in

Rappelons que les méthodes MCMC sont des procédures itératives, telles que : pour l'état actuel de la chaine, $\theta^{(i)}$, l'algorithme calcule une mise à jour probabilistique $\theta^{(i+1)}$.
La mise à jour, est faite de telle sorte que la distribution $p(\theta^i) \longrightarrow \pi(\theta/y)$, la distribution cible quand $i \to \infty$, pour n'importe quelle valeur initiale $\theta^{(0)}$. Par conséquent, Les premières valeurs simulées sont influencées par la distribution initiale de $\theta^{(0)}$, qui ne sont pas simulées à partir de $\pi(\theta/y)$.
En pratique, on écarte un ensemble de valeurs simulées initiales parce que non représentatif de la distribution cible. C'est les B premières valeurs simulées, $\theta^{(0)}, \theta^{(1)}, ..., \theta^{(B)}$ qui sont ignorées. Cette partie de la chaine écartée est connue comme la phase burn-in de la chaine.
La valeur de B, la longueur du burn-in, est déterminée en utilisant plusieurs méthodes. La méthode la plus facile est de voir la trace de chacun des paramètres d'intérêt. Si un paramètre, θ est considéré quand la convergence est atteinte en θ_B, les observations θ_i, $i > B$ doivent provenir de la même distribution. Une approche equivalente est de considérer la trace de la moyenne du paramètre d'intérêt en fonction du temps. Cette trace doit être approximativement constante quand la convergence est atteinte. Comme on peut utiliser différents diagnostics de convergence qui indiquent quand $p(\theta^{B+1})$ et $\pi(\theta/y)$ sont assez proches.

II.3 Diagnostics de convergence

Il convient de souligner dès le début qu'en pratique, il n'existe pas de tests de convergence exactes.
Les diagnostics de convergence et l'analyse des échantillons simulés utilisent deux types de méthodes :

- Analyse graphique des sorties.
- Tests formels de convergence.

II.4 Diagnostics de convergence

II.4.1 Analyse graphique

La première étape dans toute analyse des sorties est de visualiser les "traces" des valeurs générées des différentes variables $\theta_j^{(1)}, \theta_j^{(2)}, ..., \theta_j^{(M)}$ c'est à dire de représenter graphiquement les valeurs générées en fonction du nombre d'itérations. Il ne doit pas y avoir de :

– dérive continu
– forte corrélation

dans la séquence des valeurs suivants le burn-in (les échantillons sont alors supposés suivre la même distribution).

Diagnostics visuels

Un moyen de voir si notre chaine a convergé est de voir comment notre chaine se comporte dans l'espace du paramètre.
Si la chaine met beaucoup de temps pour bouger dans cet espace, elle prendra beaucoup temps pour converger.
On peut voir comment se comporte notre chaine via une inspection visuelle.
Généralement, $\theta^{(0)}$ est loin du support de la densité a posteriori. Au départ la chaine va apparaitre s'éloigner de $\theta^{(0)}$ vers une région de forte probabilité a posteriori centrée autour d'un mode de $\pi(\theta/y)$.
Si le modèle a convergé, le graphe de la trace se déplace autour du mode de la distribution.
Le temps nécessaire pour s'installer dans une région d'un mode est la valeur minimale que peut prendre B.
Si le modèle a convergé, générer d'autres valeurs de la distribution a posteriori n'influent pas sur le calcul de la moyenne.

Autocorrélation et graphes d'autocorrelation

Un autre moyen d'évaluer la convergence est d'évaluer les autocorrelations entre les valeurs de notre chaine de Markov.
L'autocorrelation au pas k, ρ_k est la corrélation entre chaque valeur de la chaine et son k^{ieme} pas :

$$\rho_k = \frac{\sum_{i=1}^{n-k}(x_i - \bar{x})(x_{i+k} - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{II.13}$$

Si l'autocorrelation reste relativement grande pour de grandes valeurs de k, ceci indique un grand degré de corrélation entre les valeurs de la chaine.
Les graphes des autocorrelations montrent les séries des corrélations dans la chaine.
Certaines corrélations entre des valeurs adjacentes résultent de la nature markovienne de l'algorithme. L'augmentation du nombre d'itérations doit réduire les autocorrelations.
La présence de corrélation indique que les valeurs ne sont pas indépendantes.
Si le degré des autocorrelations est grand pour un paramètre, alors la trace est un mauvais diagnostic de convergence.

II.4.2 Le diagnostic de Raftery and Lewis

Le diagnostic de Raftery et Lewis (voir [5]) controle la convergence d'une chaine en utilisant l'estimation d'un quantile q.
Supposons qu'on veut estimer un certain quantile a posteriori q.
On définit une tolérance acceptable r pour q et une probabilité s d'être dans cette tolérance, le diagnostic de Raftery and Lewis va calculer le nombre d'itérations M et le nombre de burn-in B nécessaires pour satisfaire les conditions spécifiées.
Le diagnostic est utilisé pour déterminer M et B à partir d'une chaine pilote. En pratique, on peut aussi tester notre chaine pour voir si elle satisfait les résultats que le diagnostic suggère.

Les entrées

1. Sélectionner un quantile a posteriori qui nous intéresse (par exemple, le quantile 0.025)
2. Sélectionner une tolérance acceptable r pour ce quantile (par exemple si $r = 0.005$, alors on veut mesurer le quantile 0.025 avec une erreur de ± 0.005).
3. Sélectionner une probabilité s, qui est la probabilité désirée d'être dans $(q - r, q + r)$.
4. Exécuter une chaine de Markov de longueur minimale égale à :

$$n_{min} = \left[\Phi^{-1} \left(\frac{s+1}{2} \right) \frac{\sqrt{q(1-q)}}{r} \right]^2 \tag{II.14}$$

où $\Phi^{-1}(.)$ est la fonction de répartition inverse de la loi normale.

II.4.3 Le test de Gelman et Rubin

Gelman et Rubin (voir [9]) propose un test de convergence basé sur les sorties de deux ou plus chaines simulées. Ce test est peut être le diagnostic le plus populaire utilisé actuellement. L'approche utilise plusieures chaines simulées à partir de différentes valeurs initiales. La méthode compare les variances intra et inter chaines. Quand les chaines convergent la variance intra de chaque séquence et la variance inter des séquences de chaque variable vont être a peu près égales. Les étapes de simulations sont les suivantes :

– Simuler indépendamment $m \geq 2$ séquences (chaines), chacune de longueure $2n$, avec des valeurs initiales assez dispersées, pour diminuer l'effet de la distribution de départ. Ecarter les n premières itérations de chaque séquence, et s'intéresser aux n dernières valeurs.

– Pour chaque paramètre d'intérêt, calculer : B/n = la variance entre les moyennes des m séquences (la variance inter séquence), $\bar{\theta_{i.}}$, chacune basée sur n valeurs de θ,

$$B/n = \sum_{i=1}^{m} (\bar{\theta_{i.}} - \bar{\theta_{..}})^2/(m-1)$$

et
W = la moyenne des m variances intra séquence, s_i^2,

$$W = \sum_{i=1}^{m} s_i^2/m.$$

Si uniquement une séquence a été simulée, B ne peut pas être calculé.
– Estimer la moyenne cible par $\hat{\mu}$, la moyenne échantillonnée des mn valeurs simulées de θ, $\hat{\mu} = \bar{\theta_{..}}$.
– Estimer la variance cible par une moyenne pondérée de W et B

$$\hat{\sigma^2} = \frac{n-1}{n}W + \frac{1}{n}B,$$

qui surestime σ^2,
– Approximer l'échantillon des n dernières itérations de θ par une distribution de Student avec une moyenne $\hat{\mu}$, variance $\sqrt{\hat{V}} = \sqrt{\hat{\sigma^2} + B/mn}$ et de degrés de liberté $df = 2\hat{V}^2/v\hat{a}r(\hat{V})$,
– Contrôler la convergence de la simulation itérative en estimant le facteur par lequel la variance de la distribution actuelle de θ peut être réduite.
Ce facteur de réduction d'échelle potentiellle (pontential scale reduction) est estimé par $\sqrt{\hat{R}} = \sqrt{(\hat{V}/W)df(df-2)}$, qui tend vers 1 quand $n \to \infty$.
$\hat{R}$ est le ratio de la variance actuelle estimée, $\hat{V}$, et la variance intra séquence, W, avec un facteur pour tenir compte de la variance supplémentaire de la t distribution de Student.
Si le facteur de réduction d'échelle potentiellle est grand, donc on peut penser que la simulation d'une chaine plus longue peut améliorer l'inférence sur la distribution cible. Si on a plus d'un paramètre, il faut calculer le facteur de réduction d'échelle potentielle de chaque paramètre. Il faut exécuter des chaines assez longues de telles sorte que tous les facteurs de réduction d'échelle potentielle soient assez petits. On peut alors combiner les mn valeurs des chaines pour produire une chaine convergente.

II.4.4 Diagnostics basés sur l'analyse spectrale

Les chaines de Markov sont un cas special des processus stochastiques, donc il est naturel d'utiliser les méthodes de séries chronologiques pour essayer d'évaluer la convergence. Il existe plusieurs moyens pratiques d'évaluation de convergence fondés sur l'analyse spectrale. Ici, on introduit deux diagnostics qui utilisent des techniques standards de l'analyse spectrale.

II.4.5 Le diagnostic de Geweke

Geweke (voir [5]) recommonde l'utilisation des méthodes de l'analyse spectrale pour montrer la convergence quand le but de l'analyse est d'estimer la moyenne. Si les valeurs de $\theta^{(i)}$ sont générées, le résultat est considéré comme une série chronologique. La méthode de Geweke se base sur l'hypothèse que la nature du processus MCMC implique l'existence d'une densité spectrale $S_\theta(w)$ pour cette série chronologique qui n'est pas discontinue en 0. Geweke suppose que la longueur du burn-in B a été choisie et ces B itérations ont été écartées. La moyenne basée sur n itérations est $\bar{\theta}_n = \frac{1}{n}\sum_{i=1}^{n}\theta_i$, la variance asymptotique est $S_\theta(0)/n$, sa racine carrée peut être utilisée pour estimer l'erreur standard de la moyenne.
La statistique du diagnostic de Geweke après n itérations est calculée en prenant la différence entre les moyennes $\bar{\theta}^A$, basée sur les premières n_A itérations et $\bar{\theta}^B$ basée sur les dernières n_B itérations et en divisant par l'erreur standard asymptotique de la différence

$$Z_n = \frac{\bar{\theta}^A - \bar{\theta}^B}{(n_A^{-1}\hat{S}_\theta(0)^A + n_B^{-1}\hat{S}_\theta(0)^B)^{\frac{1}{2}}} \tag{II.15}$$

Si $n_A + n_B < n$ et en utilisant le théorème central limite Z_n peut être approximée à une loi normale centrée réduite quand $n \to \infty$. Geweke suggère l'utilisation de $n_A = 0.1n$ et $n_B = 0.5n$. Si cette statistique donne une valeur absolue assez large (supérieure à 1.96 au seuil $\alpha = 0.05$), alors la convergence n'est pas encore atteinte.

En pratique, on doit noter que ce diagnostic est utilisé pour vérifier une condition nécessaire mais pas suffisante pour la convergence. Il peut nous informer sur la non convergence, et non pas quand elle est atteinte.

II.4.6 Diagnostic de Heidelberg et Welch

Les tests de Kolmogorov Smirnov et de Cramer-von Mises, peuvent être appliqués à une réalisation unique de la chaine $(\theta^{(t)})$ pour comparer les distributions de deux moitiés (ou d'autres sous ensembles) de cet échantillon.

Comme les tests non paramétriques sont calibrés en termes d'échantillons iid, Heidelberger et Welch (voir [5]) utilisent comme alternative au test de Kolmogarov Smirnov, la statistique de Cramer-von Mises qui approche la distance entre les distributions de deux parties de la chaine en tenant compte des corrélations entre les $\theta^{(t)}$ pour accepter ou rejeter l'hypothèse nulle que la chaine de Markov est issue d'une distribution stationnaire. Le diagnostic se déroule en deux temps

Partie 1

1. Générer une chaine de N itérations et définir un niveau α.
2. Calculer la statistique du test de la chaine entière. Accepter ou rejeter l'hypothèse nulle que la chaine est d'une distribution stationnaire.
3. Si l'hypothèse nulle est rejetée, écarter le premier dixième de la chaine. Calculer la statistique du test et accepter ou rejeter
4. Si l'hypothèse nulle est rejetée, écarter le prochain dixième de la chaine et calculer la statistique du test.
5. Répéter jusqu'à ce que l'hypothèse nulle soit accepter ou la moitié de la chaine soit écarter. Si le test rejette toujours l'hypothèse nulle, alors le test rejette la chaine et il faut exécuter une autre chaine plus longue.

Partie 2

Si la chaine passe la première partie du diagnostic, donc on utilise la partie de la chaine non écartée de la première partie pour la deuxième partie du test.

Le test halfwidth calcule la moitié de l'intervalle de confiance de niveau $(1-\alpha)$ de la moyenne.

Si le ratio de halfwidth et de la moyenne est plus petit que ϵ, alors la chaine passe le test. Sinon, on doit exécuter une chaine plus longue.

II.5 Le facteur d'inefficacité

Le facteur d'inefficacité (inefficiency factor (IF)) est une mesure pour comparer les performances de différents algorithmes alternatifs. Le IF est calculé en fonction d'une mesure appelée taille de l'échantillon effective (Effective Sample Size (ESS)) qui est définie comme suit :

Taille de l'échantillon effective

Les valeurs de la chaine générées sont dépendantes, donc la précision des estimateurs est moins bonne que s'ils avaient été obtenus avec des observations indépendantes.

L'ESS (Effective Simple Size) donne pour une chaine donnée la taille d'un échantillon indépendant qui aurait donné la même précision. Si par exemple, on se base sur M valeurs générées par une chaine l'ESS sera M_0 avec $M_0 < M$ l'idéal est d'avoir $M_0 = M$.
Le facteur d'inefficacité (IF) est défini par $IF = \frac{M}{M_0}$.
IF est utilisé pour comparer les performances de deux algorithmes, le meilleur étant celui qui a un IF le plus petit.
Le IF peut être aussi calculé en utilisant la formule suivante :

$$IF = 1 + 2\frac{B}{B-1}\sum_{i=1}^{B} K_{QS}(\frac{i}{B})\hat{\rho}_i \tag{II.16}$$

où $\hat{\rho}_i$ est l'estimateur de la corrélation au pas i des itérations MCMC, K_{QS} est le noyau spectral quadratique (QS) et B est le bandwidth. Le noyau QS est défini comme

$$K_{QS}(x) = \frac{25}{12\pi^2 x^2}(\frac{\sin(6\pi x/5}{6\pi x/5} - \cos(6\pi x/5)) \tag{II.17}$$

La bandwidth B est sélectionné automatiquement par Andrews (1991), qui estime la bandwidth comme une fonction des données. Pour le noyau QS la bandwidth est définie comme :

$$\hat{B} = 1.3221(\hat{\alpha}(2)M)^{1/5} \tag{II.18}$$

où M est le nombre d'itérations dans la chaine de Markov et

$$\hat{\alpha}(2) = \frac{4\hat{\rho_a^2}\hat{\sigma_a^4}}{(1-\hat{\rho_a})^8} / \frac{\hat{\sigma_a^4}}{(1-\hat{\rho_a})^4} \tag{II.19}$$

Les termes $\hat{\rho_a}$ et $\hat{\sigma_a}$ sont estimés en approximant notre chaine par un autoregressif d'ordre 1, où $\hat{\rho_a}$ est le coefficient autoregressif et $\hat{\sigma_a}$ est l'erreur standard estimée (voir [19] pour plus de détails).
La relation entre le IF et l'erreur standard de Monte Carlo (MCSE) estimée de la moyenne a posteriori pour chaque paramètre est donnée par :

$$MCSE^2 = \frac{\sigma^2_{MCMC}}{M} IF \tag{II.20}$$

où σ^2_{MCMC} est la variance d'échantillon des itérations MCMC et M dénote le nombre d'itérations. Il est clair que IF représente le ratio de la variance de la chaine simulée (le carré de MCSE estimée) et de la variance obtenue à partir de réalisations indépendantes. Par exemple, une valeur 100 de IF implique

que , pour un nombre d'itérations donné M, la corrélation dans les itérations MCMC produit une MCSE estimée de la moyenne a posteriori qui est dix fois plus grande que celle de réalisations indépendantes.
Le but est de réduire la dépendance dans les itérations MCMC et donc d'avoir un IF le plus petit possible. On dit qu'un algorithme 1 est préféré a un autre algorithme 2 si pour un certain paramètre θ si le plus grand IF des éléments de θ calculés à partir de l'algorithme 1 est plus petit que le plus grand IF des éléments de θ calculés à partir de l'algorithme 2.

II.6 Conclusion

Dans ce chapitre, nous avons introduit d'une part la classe des méthodes MCMC en particulier l'algorithme de Metropolis Hastings et l'échantillonneur de Gibbs ainsi que leurs application en estimation bayésienne, d'autre part nous avons développé les différents diagnostics utilisés pour s'assurer de leurs convergence.

Chapitre III

Modèles à espace d'état

Récemment, il y a eu un intérêt croissant dans l'application des modèles à espace d'état dans l'analyse des séries chronologiques. Les modèles à espace d'état considèrent les séries chronologiques comme sorties (output) d'un système dynamique perturbé par des bruits aléatoires. Ils interprètent la série chronologique comme combinaison de plusieurs composantes telles qu'une tendance, une saisonnalité ou une composante régressive. En même temps ils ont une structure en offrant un cadre souple pour une très large gamme d'application. Les calculs peuvent être mis en oeuvre par des algorithmes récursifs. Les problèmes d'estimation et de prévision sont résolus par des calculs récursifs portant sur la distribution conditionnelle des quantités d'intérêt, compte tenu des informations disponibles. En ce sens, ils sont tout naturellement traité dans un cadre bayésien.
Les modèles à espace d'état peuvent être utilisé pour modéliser des séries univariées on multivariées même en présence de non stationnarité.
La grande flexibilité des modèles à espace d'état est l'une des raisons de leur vaste application dans beaucoup de problèmes appliqués.
Les modèles à espace d'état sont couramment utilisés avec les données de séries chronologiques. Les données de séries chronologiques comme leur nom l'indique sont ordonnées par le temps. Elles surviennent généralement dans les domaines de la macroéconomie (par exemple, le taux de chômage observé chaque trimestre) et de la finance (par exemple, le prix d'un actif quotidien pendant un an).

III.1 Modèles à espace d'état

Le modèle à espace d'état est composé de deux équations l'une est appelée équation de mesure l'autre est appelée équation d'état. Dans un modèle à

espace d'état le vrai état (α_t) n'est pas observé, ce qui est observé est un transformé bruité de ce dernier (y_t).
Dans leur version élémentaire, les modèles à espace d'état reposent sur un certain nombre d'hypothèses principales

- les équations de mesure et d'état sont linéaires.
- les bruits d'observation et d'innovation sont des bruits blancs.
- les variables d'état suivent à l'instant initial une loi gaussienne.

À ces dernières, se sont ajoutées des hypothèses secondaires permettant de simplifier les procédures d'estimation.
Nous allons commencer par le modèle univarié le plus simple appelé modèle à niveau local.

III.1.1 Modèle à niveau local

Soit $(y_t)_{t\in[1,T]}$ une série d'observations, on appelle modèle à niveau local, le modèle :

$$y_t = \alpha_t + \varepsilon_t \tag{III.1}$$

$$\alpha_{t+1} = \alpha_t + \eta_t \tag{III.2}$$

où

- ε_t est i.i.d $N(0, H)$.
- η_t est i.i.d $N(0, Q)$.
- ε_t et η_s sont indépendantes pour tout t et s.

Notons que dans (III.1) t varie de 1 jusqu'à T tandis que dans (III.2) il varie de 1 jusqu'à $T-1$. L'équation (III.2) ne fournit pas une expression de α_1, qui est appelée condition initiale.
L'équation (III.1) est appelée équation de mesure tandis que l'équation (III.2) est appelée équation d'état. α_t est dite variable d'état ou variable latente.

III.1.2 Modèle à espace d'état linéaire gaussien

La classe la plus importante des modèles à espace d'état est donnée par les modèles à espace d'état linéaires gaussiens, appelés aussi les modèles dynamiques linéaires (Dynamic Linear Models (DLM)). Ces modèles sont définis par le système d'équations suivant :

$$y_t = Z_t\alpha_t + \varepsilon_t,\ \varepsilon_t \sim N(0, H_t) \tag{III.3}$$

$$\alpha_t = T_t\alpha_{t-1} + R_t\eta_t,\ \eta_t \sim N(0, Q_t) \tag{III.4}$$

pour $t = 1, ..., T$. Ici y_t est un $(p\times 1)$ vecteur d'observations, α_t est un $(m\times 1)$ vecteur d'états, R_t est une matrice composée de r colonnes de la matrice

identité, et les matrices variances H_t et Q_t sont non singulières. Les composantes des vecteurs ε_t et η_t sont indépendantes et ces deux vecteurs sont indépendants entre eux. Les matrices H_t, Q_t, Z_t et T_t sont supposées connues dépendant d'un vecteur de paramètres θ. Un modèle DLM est spécifié par une distribution normale pour le vecteur d'état en $t = 0$

$$\alpha_0 \sim N(m_0, C_0) \tag{III.5}$$

III.1.3 Modèle à espace d'état linéaire non gaussien

Le modèle à espace d'état non gaussien est défini par les équations de mesure et d'état (voir [6]), telle que $y_t/\alpha_1, ..., \alpha_t,\ y_1, ..., y_{t-1} \sim p(y_t/Z_t\alpha_t)$ et $\eta_t \sim p(\eta_t)$ pour $t = 1, ..., T$, les η_t sont iid.
Le modèle est dit non gaussien si $p(y_t/Z_t\alpha_t)$ et/ou $p(\eta_t)$ sont non gaussiennes, en particulier lorsque $p(y_t/Z_t\alpha_t)$ appartient à la famille des lois exponentielles ou lorsque les observations sont générées par la relation

$$y_t = Z_t\alpha_t + \varepsilon_t,\ \varepsilon_t \sim p(\varepsilon_t) \tag{III.6}$$

où les ε_t sont indépendants de lois non gaussiennes.

III.1.4 Modèle à espace d'état général

Un modèle à espace d'état général peut être décrit par le système d'équations suivants :

$$Y_t = h_t(\alpha_t, \varepsilon_t) \tag{III.7}$$

$$\alpha_t = g_t(\alpha_{t-1}, \eta_t) \tag{III.8}$$

où h_t et g_t sont deux fonctions arbitraires. Pour les modèles à espace d'état linéaires h_t et g_t sont des fonctions linéaires et pour les modèles linéaires gaussiens on ajoute l'hypothèse de normalité des distributions.

Remarque Les quantités Z_t, T_t, H_t et Q_t sont connues mais peuvent dépendre d'un paramètre θ univarié ou multivarié.

III.2 Estimation dans les modèles à espace d'état

L'estimation dans les modèles à espace d'état consiste à estimer les paramètres du modèle ainsi que le vecteur d'état.

III.2.1 Estimation des variables latentes

Pour un modèle à espace d'état donné, la tâche principale est de faire l'inférence des états non observés ou de prédire les observations futur. L'estimation et la prévision sont résolues en calculant les distributions conditionnelles des quantités qui nous intéressent.
Le problème consiste à estimer le vecteur d'état α_t compte tenu des informations disponibles à l'instant n, postérieur, antérieur ou identique à l'instant t. Pour cela on calcule les densités conditionnelles $p(\alpha_n/y_{1:t})$. On peut considérer les cas suivants :

1. $t > n$ il s'agit dans ce cas d'estimer la valeur d'état à un instant t dans le futur. C'est ce qu'on appelle la prédiction.
2. $t = n$ il s'agit dans ce cas de déterminer une estimée de l'état, compte tenu de toutes les mesures disponibles jusqu'à instant considéré n. C'est le cas du filtrage.
3. $t < n$ en utilisant les informations disponibles jusqu'à l'instant n, on essaye d'estimer l'état à l'instant t antérieur à n. On fait alors un lissage.

Les formules récursives dites de Kalman ont pour objectif de trouver des estimateurs linéaires optimaux (au sens du minimum de l'erreur quadratique), du vecteur d'état α_t en fonction des observations $y_{1:t}$. On notera $P_t(\alpha_k)$ le meilleur prédicteur linéaire de α_k en fonction de $y_{1:t}$.
En conclusion, les problèmes suivants sont considérés :

- prédiction : estimer α_t en fonction de $y_{1:t-1}$
- filtrage : estimer α_t en fonction de $y_{1:t}$
- lissage : estimer α_t en fonction de $y_{1:n}$ avec $n > t$

Dans les trois cas, la récursion porte sur la prise en compte d'une observation y supplémentaire.
Dans les cas de la prédiction et du filtrage, l'indice t du α_t estimé augmente également à chaque itération ; dans le cas du lissage, t est fixé alors que n augmente.
Il existe différentes façons d'écrire les équations récursives de Kalman selon le but poursuivi.
Nous décrirons ici seulement le système d'équations le plus couramment utilisé (filtrage de Kalman), qui inclut lui même une phase de prédiction.
Soit $P_t(\alpha_k)$ le meilleur prédicteur linéaire de α_k en fonction de $y_{1:t}$.
Le filtre de Kalman a donc pour objectif de calculer à chaque instant t la valeur de $P_t(\alpha_k)$, ainsi que la matrice de covariance de l'erreur commise sur cet estimateur.
Pour cela, le filtre opère à chaque itération t en plusieurs étapes :

- prédiction de l'état : calcul de $P_{t-1}(\alpha_t)$ et covariance associée $C_{t/t-1}$ de l'erreur de prédiction
- prédiction de la mesure : calcul de $P_{t-1}(y_t)$
- calcul du terme d'innovation, égal à l'écart entre la mesure y_t et sa valeur prédite, et de sa covariance S_t
- calcul du gain du filtrage K_t
- filtrage : calcul de $P_t(\alpha_t)$ comme combinaison linéaire de la prédiction et de l'innovation, et covariance associée $C_{t/t}$ de l'erreur d'estimation

Il est nécessaire également d'initialiser ce processus.

Equations du filtre de Kalman

Initialisation

Consiste à donner une estimée de α à l'instant 0, soit $P_0(\alpha_0)$ ainsi que la matrice de covariance $C_{0/0}$. Ceci se fait à partir des informations a priori dont on dispose sur l'état à l'instant 0 ; en absence d'information a priori, $C_{0/0}$ sera pris très grand, ce qui a pour effet d'accorder un poids négligeable à la valeur initiale.

1. Prédiction de l'état

$$P_{t-1}(\alpha_t) = T_{t-1} P_{t-1}(\alpha_{t-1}) \tag{III.9}$$

$$C_{t/t-1} = T_{t-1} C_{t-1/t-1} T'_{t-1} + Q_{t-1} \tag{III.10}$$

2. Prédiction de la mesure

$$P_{t-1}(y_t) = Z_t P_{t-1}(\alpha_t) \tag{III.11}$$

3. Calcul de l'innovation à l'instant t et de sa covariance S_t

$$\upsilon_t = y_t - P_{t-1}(y_t) \tag{III.12}$$

$$S_t = Z_t C_{t/t-1} Z'_t + H_t \tag{III.13}$$

4. Calcul du gain K_t

$$K_t = C_{t/t-1} Z'_t S_t^{-1} \tag{III.14}$$

5. Estimation d'état courant, comme combinaison linéaire de la valeur prédite et de l'innovation

$$P_t(\alpha_t) = P_{t-1}(\alpha_t) + K_t \upsilon_t \tag{III.15}$$

$$C_{t/t} = C_{t/t-1} - K_t S_t K'_t \tag{III.16}$$

Remarques

1. La dernière équation :

$$C_{t/t} = C_{t/t-1} - K_t S_t K_t' \tag{III.17}$$

fait apparaitre explicitement la diminution de la covariance de l'erreur d'estimation apportée par une nouvelle mesure. On peut l'écrire de façon équivalente :

$$C_{t/t} = (I - K_t Z_t) C_{t/t-1} (I - K_t Z_t)' + K_t H_t K_t' \tag{III.18}$$

issue de l'équation :

$$P_t(\alpha_t) = (I - K_t Z_t) P_t(\alpha_t) + K_t y_t \tag{III.19}$$

2. Une mesure y_t de mauvaise qualité ($H_t \to \infty$) conduit à $P_t(\alpha_t) = P_{t-1}(\alpha_t)$ (la mesure y_t n'est pas prise en compte) ; une mesure y_t sans erreur ($H_t = 0$)conduit à $Z_t P_t(\alpha_t) = y_t$.

Le filtre de Kalman pour les modèles dynamiques linéaires (DLM)

Les résultats précédents résolvent les problèmes du filtrage et de prévision, en général le calcul des distributions conditionnelles n'est pas toujours facile. Les DLM sont un cas particulier où les récursions générales sont simplifiées considérablement. Dans ce cas, en utilisant les propriétés de la distribution gaussienne multivariée, il est facile de prouver que le vecteur aléatoire $(\alpha_0, \alpha_1, ..., \alpha_t, y_1, ..., y_t)$ a une distribution gaussienne pour tout $t \geq 1$. Il découle que les densités marginales et conditionnelles sont aussi gaussiennes. Du fait que toutes ces distributions sont gaussiennes, elles sont complètement déterminées par leurs moyennes et matrices de variances. La solution du problème du filtre des DLM est donnée par le célèbre filtre de Kalman.

Proposition III.2.1. *(voir [14])*
Considérons le DLM spécifié en III.3 et III.4. Soit

$$\alpha_{t-1}/y_{1,t-1} \sim N(m_{t-1}, V_{t-1}) \tag{III.20}$$

donc les affirmations suivantes sont vérifiées :
i) La distribution prédictive de α_t sachant $y_{1:t-1}$ est gaussienne de paramètres :

$$a_t = E(\alpha_t/y_{1:t-1}) = T_t m_{t-1} \tag{III.21}$$

$$R_t = var(\alpha_t/y_{1:t-1}) = T_t V_{t-1} T_t' + Q_t \tag{III.22}$$

ii) La distribution prédictive de Y_t *sachant* $y_{1:t-1}$ *est gaussienne de paramètres :*

$$f_t = E(Y_t/y_{1:t-1}) = Z_t a_t \tag{III.23}$$

$$W_t = var(Y_t/y_{1:t-1}) = Z_t R_t Z_t' + H_t \tag{III.24}$$

iii) La distribution filtrée de α_t *sachant* $y_{1:t}$ *est gaussienne de paramètres*

$$m_t = E(\alpha_t/y_{1:t}) = a_t + R_t Z_t' W_t^{-1} e_t \tag{III.25}$$

$$V_t = var(\alpha_t/y_{1:t}) = R_t - R_t Z_t' W_t^{-1} Z_t R_t \tag{III.26}$$

où $e_t = Y_t - f_t$ *l'erreur de prévision.*

Le filtre de Kalman nous permet de calculer les distributions prédictives et du filtrage respectivement, en commençant à partir de $\alpha_0 \sim N(m_0, V_0)$, puis de calculer $\pi(\alpha_1/y_1)$, et de procéder récursivement au fur et à mesure de nouvelles données deviennent disponibles.
La distribution conditionnelle de $\alpha_t/y_{1:t}$ résout le problème du filtrage. Cependant, dans plusieurs cas on s'intéresse à l'estimation ponctuelle. L'estimateur bayésien de α_t compte tenu des informations $y_{1:t}$, par rapport à la fonction de perte quadratique $l(\alpha_t, a) = (\alpha_t - a) = (\alpha_t - a)'H(\alpha_t - a)$ est l'espérance conditionnelle $m_t = E(\alpha_t/y_{1:t})$. C'est l'estimation optimale car elle minimise la perte conditionnelle espérée $E((\alpha_t - a)'H(\theta_t - a)/y_{1:t-1}$ par rapport à a. Si $H = I_p$, la perte espérée minimale est la matrice variance conditionnelle $var(\theta_t/y_{1:t})$. L'expression de m_t a la forme intuitive estimation-correction la moyenne du filtre (filtrée) est égale à la moyenne de prédiction a_t plus une correction en fonction de la différence entre la nouvelle observation et sa prédiction. Le poids du terme de la correction est données par la matrice gain

$$K_t = R_t Z_t' W_t^{-1} \tag{III.27}$$

Ainsi, le poids de la donnée actuelle y_t dépend de la variance de l'observation $\mathsf{V_t}$ (par W_t) et de $R_t = var(\alpha_t/y_{1:t}) = T_t V_{t-1} T_t' + Q_t$.

Le lissage de Kalman

Le lissage de Kalman nous permet de calculer les distributions $\alpha_t/y_{1:T}$, en commençant à partir de $t = T - 1$, dans ce cas $\alpha_T/y_{1:T} \sim N(s_T = m_T, S_T = V_T)$, et ensuite procéder en arrière pour calculer les distributions de $\alpha_t/y_{1:t}$ pour $t = T - 2,\ t = T - 3$,etc. Notons que la récursion du lissage dépend des données uniquement par le filtrage et par les moments prédictifs obtenus en utilisant le filtre de Kalman.

Proposition III.2.2. *(voir [14])*

Pour un modèle DLM défini en III.3 et III.4, si $\alpha_{t+1}/y_{1:T} \sim N(s_{t+1}, S_{t+1})$ *alors*

$$\alpha_t/y_{1:T} \sim N(s_t, S_t) \tag{III.28}$$

où

$$s_t = m_t + V_t T'_{t+1} R^{-1}_{t+1}(s_{t+1} - a_{t+1}) \tag{III.29}$$

et

$$S_t = V_t - V_t T'_{t+1} R^{-1}_{t+1}(R_{t+1} - S_{t+1}) R^{-1}_{t+1} T_{t+1} V_t \tag{III.30}$$

III.2.2 Estimation des paramètres

Dans la partie précédente, le paramètre θ ainsi que les paramètres du vecteur d'état initial sont supposés connus. En pratique, ces paramètres sont inconnus et doivent être estimés. Il existe plusieurs méthodes d'estimation, on peut citer l'algorithme EM et les méthodes bayésiennes qui sont appropriées à ce genre de modèles.

Algorithme EM

L'algorithme EM est couramment utilisé pour déterminer les estimateurs du maximum de vraisemblance (EMV) des paramètres d'un modèle à espace d'état. Cet algorithme itératif a le mérite d'être simple, même s'il est relativement lent à converger par rapport à des algorithmes plus sophistiqués. Il consiste à estimer le maximum de vraisemblance de modèles stochastiques à variables latentes (cachées).

Pour procéder à une estimation par maximum de vraisemblance des paramètres d'un modèle à espace d'état, il est nécessaire d'avoir l'expression de la fonction de vraisemblance. Pour chaque jeu de paramètres θ, la log-vraisemblance associée à un échantillon $y_1, ..., y_T$ doit être calculée.

L'algorithme EM est un algorithme itératif qui génère une séquence d'estimations $(\theta_i)_{i\geq 1}$ à partir d'une valeur initiale θ_0. Chaque itération se décompose en deux étapes qui s'écrivent :

- La première étape E "Expectation" calcule la vraisemblance d'un modèle à espace d'état. Ces formules mobilisent en particulier l'application d'un filtre de Kalman pour connaître l'espérance conditionnelle de l'état $P_{t-1}(\alpha_t)$ et de sa covariance sachant θ_i et les observations $y_{0:T}$ fixés.
- La seconde étape M "Maximisation", consiste à rechercher un jeu de paramètres maximisant la vraisemblance estimée dans l'étape E. Cette maximisation peut-être analytique ou numérique selon la complexité du problème.

Après un cycle « Étape E / Étape M», on obtient θ_{i+1} et on peut montrer que $L(y_t/\theta_{i+1}) > L(y_t/\theta_i)$.
En itérant ces étapes E et M, les paramètres estimés par l'algorithme convergent généralement vers le maximum de vraisemblance.
Méthodes bayésiennes

Parmi toutes les méthodes d'estimation, on choisit l'approche bayésienne pour estimer les paramètres d'un modèle à espace d'état parce qu'elle a plusieurs avantages. Premièrement, le passage d'un système déterministe à un système stochastique c'est à dire l'incertitude qui est due à l'oubli de variables, erreurs de mesure ou imperfections est décrite à travers des probabilités. En conséquence, l'estimation des quantités d'intérêt (en particulier, l'état du système à l'instant t) est résolue en calculant leurs distributions conditionnelles, sachant l'information disponible. Comme les modèles à espace d'état sont basés sur l'idée de décrire les sorties du système comme fonction des états inobservés affectés par erreurs aléatoires, cette façon de modéliser la dépendance temporelle dans les données en conditionnant par rapport aux variables latentes est simple est extrêmement puissante et assez naturelle dans une approche bayésienne. Un autre avantage est que les calculs peuvent être fait de façon récursive : les distributions conditionnelles d'intérêt peuvent être mises à jour en intégrant les nouvelles données sans la nécessité de stocker toutes les valeurs du passé. Ceci est très avantageux lorsque les données arrivent séquentiellement dans le temps et la réduction de la capacité de stockage devient plus importante pour les grands ensembles de données.

III.3 Conclusion

Les modèles à espace d'état sont très avantageux car ils permettent de modéliser une large gamme de séries chronologiques notamment celles où on considère que ce ne sont pas les vrais états qui sont observés mais une composonte bruitée. D'autre part ce type de modèles s'applique bien aux séries financières et actuarielles où les données arrivent séquentiellement dans le temps. Le chapitre suivant a pour objet deux modèles à espace particuliers modélisant des phénomènes financiers. Ainsi qu'une extension aux modèles de diffusion qui sont des modèles à temps continu qui sont utilisés en théorie actuarielle, l'accent sera mis sur le modèle SV en temps continu.

Chapitre IV

Les modèles SCD et SV

Dans ce chapitre, nous allons définir deux modèles très utilisés en théories financière et actuarielle à savoir le modèle à durée conditionnelle stochastique (Stochastic Conditional Duration (SCD)) et le modèle à volatilité stochastique (Stochastic Volatility (SV)). Ce sont des modèles à espace d'état non linéaires et non gaussiens. Ces deux modèles appartiennent à la classe des modèles à temps discret. Nous introduisons aussi les processus de diffusion qui sont des modèles à temps continu. Dans cette partie nous allons aussi considérer l'estimation de ces deux classes de modèles qui consiste à estimer les variables d'états (latentes) ainsi que les paramètres inconnus.
L'estimation des variables d'état peut être effectuée à l'aide de la méthode pas à pas appelée aussi single-move sampler qui consiste à mettre à jour le vecteur des paramètres élément par élément.
Lorsque les paramètres sont corrélés, particulièrement dans les modèles de séries chronologiques la procédure single-move converge très lentement. Une alternative d'estimation plus efficace est constituée par les algorithmes nommés multi-move sampler qui consiste à mettre a jour plusieurs paramètres à la fois.

IV.1 Le modèle à durée conditionnelle stochastique (SCD)

La large disponibilité des données au niveau des transactions financières a permis de chercher à modéliser la microstructure des marchés financiers. De nouveaux modèles et nouvelles méthodes ont été développées pour permettre d'analyser les modèles "intraday".
Le modèle à durée conditionnelle stochastique (Stochastic Conditional Duration) est un modèle qui est utilisé pour modéliser une séquence de durées

entre des transactions boursières par exemple. C'est un modèle qui a été introduit en contraste du modèle à durée conditionnelle autoregressive (Autoregressive Conditional Duration (ACD)), dans lequel la moyenne conditionnelle des durées est considérée comme une fonction conditionnellement déterministe des informations du passé, tandis que le modèle SCD traite la moyenne conditionnelle des durées en fonction d'un processus latent avec une distribution conditionnelle ayant un support positif.

Soit $y = (y_1, y_2, ..., y_T)^{'}$ un vecteur T-dimensionnel de durées, le modèle SCD est défini par les deux équations d'état et de mesure suivantes :

$$y_t = \exp(\alpha_t)\varepsilon_t \tag{IV.1}$$

$$\alpha_t = W_{t-1}^{'}\delta + \phi\alpha_{t-1} + \sigma_\eta\eta_t \tag{IV.2}$$

- $\eta_t \sim^{iid} N(0.1)$
- ε_t a une distribution à support positif.
- Pour $t = 1, 2, ..., T$, y_t conditionnellement à α_t est supposée indépendante de η_t pour tout t.
- Le processus de variables latentes (IV.2) est supposé avoir une variance finie, avec $\mid \phi \mid < 1$.
- Le $(1 \times k)$ vecteur W_t contient la $t^{ème}$ ligne de la matrice de régresseurs observés W, et δ un vecteur de coefficients.
- L'état initial a une distribution marginale donnée par :

$$\alpha_1 \sim N\left(\frac{W_1^{'}\delta}{1-\phi}, \frac{\sigma_\eta^2}{1-\phi^2}\right) \tag{IV.3}$$

Les y_t conditionnellement à α_t sont suposées indépendantes entre elles c'est à dire :

$$p(y/\alpha) = \prod_{t=1}^{T} p(y_t/\alpha_t) \tag{IV.4}$$

La fonction densité de probabilité définie en (IV.4) est supposée avoir un support positif, telles que Exponentielle, Weibull ou Gamma.
Dans notre cas on choisit une densité Exponentielle, on a alors :

$$\varepsilon_t \sim Exp(1)\ iid$$

IV.2 Modèle à volatilité stochastique (SV)

Au cours de la dernière décennie, on s'est intéressé à la modélisation de l'évolution dynamique de la volatilité des séries à haute fréquence en finance

en utilisant les modèles à volatilité stochastique (Stochastic Volatility(SV)). Le modèle SV offre une alternative aux modèles du type ARCH. La principale caractéristique de ces modèles est que la volatilité est modélisée par une variable latente non observée. Les modèles SV sont intéressants car ils sont proches des modèles souvent utilisés dans la théorie financière pour représenter le comportement des prix des actifs financiers, il est plus réaliste et fléxible pour la modélisation des séries temporelles financières que les modèles du type ARCH
Le modèle s'écrit :

$$y_t = e^{h_t/2}\varepsilon_t \quad t \geq 1 \tag{IV.5}$$

$$h_{t+1} = \mu + \phi(h_t - \mu) + \sigma_\eta \eta_t \tag{IV.6}$$

$$h_1 \sim N\left(\mu, \frac{\sigma^2}{1-\phi^2}\right) \tag{IV.7}$$

- h_t est la log volatilité au temps t qui est supposée être un processus stationnaire ($|\phi| < 1$).
- ε_t et η_t sont des normales non corrélées entr'elles et non corrélées avec h_1 où ε_t est une normale standard.
- ϕ est la persistance de la volatilité.
- σ_η est la volatilité de la log volatilité.

IV.3 Modèles à volatilité stochastique en temps continu

IV.3.1 Processus de diffusion

Les processus de diffusion sont un outil important pour modéliser les modèles financiers et actuariels. Mais l'estimation est problématique parce que ces modèles sont formulés en temps continu tandis que les données sont disponibles à des fréquences discètes, ceci implique que les estimations obtenues par la discrétisation des processus de diffusion peuvent être sujet à des biais de discrétisation. On utilise les méthodes MCMC pour estimer les paramètres dans les processus de diffusion.
Puisque les données ne sont disponibles qu'à des moments discrets, et peut être peu fréquemment, on introduit un ensemble de $(m-1)$ données latentes auxiliaires entre chaque paire d'observations. Les méthodes MCMC sont particulièrement adaptées à ce type de problèmes. En effet ces méthodes sont très utiles pour résoudre des problèmes dans lesquels la fonction de vraisemblance implique le calcul d'une intégrale de dimension élevée (par exemple modèles à espace d'état, le modèle SV,...etc).

Dans ce qui suit nous décrirons les étapes élémentaires pour la simulation des données d'un processus de diffusion en général et la simulation des paramètres d'un modèle SV en particulier.

IV.3.2 Mouvement Brownien ou processus de Wiener

Le mouvement brownien doit son nom au botaniste Robert Brown, ce dernier observait le pollen au microscope et a constaté la présence de très petites particules bougeant dans tous les sens. C'est de là que vient le nom. Une manière d'expliquer et de définir ce mouvement est la suivante :
supposons une marche aléatoire symétrique telle que :

$$\begin{cases} 1 & \text{avec la probabilité } \frac{1}{2} \\ -1 & \text{avec la probabilité } \frac{1}{2} \end{cases}$$

ensuite accélérons ce processus en prenant des pas de plus en plus petits (i.e $\Delta Y \longrightarrow 0$) et en prenant des intervalles de temps de plus en plus petits (i.e $\Delta t \longrightarrow 0$), en passant à la limite, on obtient le mouvement brownien. Ce mouvement a été étudié par de nombreux mathématiciens. Norbert Wiener, notamment, a complètement défini le processus aléatoire qui modélise ce phénomène, c'est pour cela qu'on entend parfois parler du processus de Wiener lorsqu'on parle du mouvement brownien.
Un processus stochastique à temps continu $\{W_t, 0 \leq t \leq T\}$ est appelé mouvement Brownien sur $[0, T)$ s'il possède les propriétés suivantes

1. $W_0(\omega) = 0$ presque sûrement
2. Les incréments de W_t sont indépendants : pour toute suite finie de temps $0 \leq t_1 < t_2 < ... < t_n$, les variables aléatoires $W_{t_2} - W_{t_1}, ...W_{t_n} - W_{t_{n-1}}$ sont indépendantes.
3. Pour tout $0 \leq s \leq t \leq T$, l'incrément $W_t - W_s$ suit une loi normale de moyenne nulle et de variance $t - s$
4. Pour tout ω sur un ensemble de probabilité 1, $W_t(\omega)$ est une fonction continue de t.
5. $W_{t+s} - W_t$ ne dépend pas de t.

IV.3.3 Introduction aux équations différentielles stochastiques

Les équations différentielles stochastiques (EDS) sont des équations différentielles qui contiennent un ou plusieurs termes qui sont des processus stochastiques. La solution de ces équations est un processus stochastique aussi.

Les EDS incorporent un bruit blanc qui peut être vu comme la dérivée d'un mouvement brownien (qui sera introduit plus tard). Une équation différentielle s'écrit :

$$dY_t = \mu(Y_t;\theta)dt + \sigma(Y_t;\theta)dW_t,\ Y_0 = y_0 \tag{IV.8}$$

où W_t est un mouvement brownien de dimension d, μ et σ sont les coefficients de l'équation (ce sont des fonctions de $R^d \times \Theta$).
Le coefficient μ est appelé coefficient de dérive, tandis que σ est dit coefficient de diffusion.
On peut aussi exprimer la même équation sous forme d'intégrale :

$$Y_{t+s} - Y_t = \int_t^{t+s} \mu(Y_u;\theta)du + \int_t^{t+s} \sigma(Y_u;\theta)dW_u \tag{IV.9}$$

Un processus solution de l'équation (IV.8) est appelé processus de diffusion ou plus simplement une diffusion.

Remarque Le mouvement brownien est un processus de diffusion de coefficient de dérive nul et de coefficient de diffusion égal à 1.

Parmi les modèles de diffusion utilisés en théorie financière et actuarielle, les modèles à elasticité constante de variance (CEV) qui sont des modèles à un facteur tels que le modèle de Cox (racine carré), le modèle d'Orstein-Uhlenbeck, etc... . Une généralisation de ces modèles est constituée des modèles à volatilité stochastique (SV). Dans cette partie nous étudions le modèles SV à temps continu ainsi que l'estimation de ses paramètres en utilisant les méthodes MCMC.

IV.3.4 Le modèle SV

Dans ce chapitre nous allons étudier l'estimation des paramètres du modèle à volatilité stochastique (SV), à temps continu, écrit sous forme d'un processus de diffusion comme suit :

$$dr_t = (\theta_r + \kappa_r r_t)dt + \exp(\frac{1}{2}Z_t)r_t^{\beta}dW_{1,t} \tag{IV.10}$$

$$dZ_t = (\theta_z + \kappa_z Z_t)dt + \sigma_z dW_{2,t} \tag{IV.11}$$

où r_t est supposé être observé, $t = 0, 1, ..., T$, et Z_t est un processus inobservé.

IV.4 Estimation

IV.4.1 Introduction

Dans cette partie on étudie l'estimation des paramètres des deux modèles SCD et SV, modèles à espace d'état non linéaires et non gaussiens. Pour ce faire on utilise les méthodes MCMC qui sont très appropriées pour la simulation de ce type de modèles. La représentation non gaussienne du modèle à espace d'état est approximée par un modèle linéaire gaussien, ce modèle approximatif définit la distribution candidate à partir de laquelle le vecteur d'état va être simulé via l'application du filtre et du lissage de Kalman. L'avantage des méthodes MCMC par rapport aux autres méthodes de simulation est la possibilité d'une part d'utiliser des algorithmes hybrides c'est à dire d'utiliser un algorithme de Metropolis-Hastings (MH) dans les étapes d'un Gibbs sampler et d'autre part puisqu'on a un vecteur d'état de grande dimension de le diviser en plusieurs blocs (multi-move) de telle sorte que la simulation d'un des blocs soit conditionnelle aux blocs précedents.

IV.4.2 Le modèle SCD

En définissant le vecteur des paramètres inconnus du processus latent comme $\theta = (\delta, \phi, \sigma_\eta)^{'}$, la densité a posteriori jointe des paramètres inconnus dans le modèle SCD est donnée par :

$$p(\alpha, \theta/y, W) \propto p(y/\alpha) \times p(\alpha/W, \theta) \times p(\theta) \tag{IV.12}$$

où $p(\alpha/W, \theta)$ dénote la densité jointe de α conditionnellement à W et θ, et $p(\theta)$ est la densité a priori de θ. La densité $p(y/\alpha)$ est comme définie précédemment.

$$p(\alpha/W, \theta) = \{\prod_{i=2}^{T} p(\alpha_i/W_{i-1}, \alpha_{i-1}, \theta)\} \times p(\alpha_1/W_1, \theta) \tag{IV.13}$$

où

$$p(\alpha_i/W_{i-1}, \alpha_{i-1}, \theta) \propto \exp\{\frac{-1}{2\sigma_\eta^2}[\alpha_i - W_{i-1}^{'}\delta - \phi\alpha_{i-1}]^2\}$$

pour $i = 2, \cdots, T$, et $p(\alpha_1/W_1, \theta)$ comme définit en (IV.3).
En utilisant l'algorithme de Gibbs, les distributions conditionnelles des états inobservés de chaque bloc doivent être obtenues. Dans la simulation multi move adoptée ici tous les états latents sont simulés dans des blocs de taille superieure à 1. Comme le modèle SCD est un modèle à espace d'état non gaussien, la difficulté est de trouver une bonne densité candidate des états

de chaque bloc. Une des approches possibles est d'approximer l'équation de mesure par une équation de densité gaussienne avec un mode égal c'est à dire que le mode de la densité gaussienne approximative utilisée soit égal à celui de la densité non gaussienne. Cette approximation est performée via le filtre de Kalman.
Les étapes de l'algorithme de Gibbs sont données par :

1. Initialiser α
2. Générer $\theta/y, W, \alpha$
3. Générer $\alpha/y, W, \theta$ où θ est divisé en blocs de taille supérieure ou égale à 1.
4. Répéter les étapes 2 et 3 jusqu'à convergence.

Simulation de θ

Etant donné (IV.12), la densité conditionnelle de θ est donnée par :

$$p(\theta/y, \alpha, W) \propto p(\alpha/W, \theta) \times p(\theta) \tag{IV.14}$$

où $p(\alpha/W, \theta)$ est définie en (IV.13) et $p(\theta) = p(\delta) \times p(\phi) \times p(\sigma_\eta)$ on suppose l'indépendance entre les lois a priori des éléments de θ. Pour σ_η on adopte une densité a priori de loi Inverse Gamma telle que :

$$\sigma_\eta \sim IG\left(\frac{\sigma_r}{2}, \frac{S_\sigma}{2}\right)$$

où σ_r et S_σ sont les hyperparamètres.
Les lois a priori de ϕ et de δ sont respectivement des lois uniforme sur l'intervalle $(-1, 1)$ et uniforme sur R .
Avec le choix de ces lois a priori et la densité initiale de α définie en (IV.3), $p(\theta/y, \alpha, W)$ n'a pas une forme standard à partir de laquelle θ peut être simulé directement.
Donc, on utilise un algorithme de MH pour simuler θ indirectement, via une densité candidate spécifiée comme suit (voir [19]). Définir :

$$Z = \begin{pmatrix} \omega_{1,1} & \cdots & \omega_{1,k} & \alpha_1 \\ \vdots & & \vdots & \vdots \\ \omega_{T-1,1} & \cdots & \omega_{T-1,k} & \alpha_{T-1} \end{pmatrix}$$

où $w_{i,j}$ est le $(i,j)^{ème}$ élément de W. En définissant $\beta = (\delta', \phi)'$, on exprime (IV.2) comme :

$$\alpha = Z\beta + u$$

où $\alpha = (\alpha_2, ..., \alpha_T)$ et $u \sim N(0, \sigma_\eta^2 I_{T-1})$. En adoptant une loi a priori standard non informative pour $(\beta', \sigma_\eta)'$ de la forme $p(\beta', \sigma_\eta) \propto \frac{1}{\sigma_\eta}$, une candidate pour θ est donnée par une distribution Normale Gamma.
Les étapes de l'algorithme de MH de la $j^{ème}$ itération de l'échantillonneur de Gibbs sont les suivantes :

1. Initialiser θ^{j-1}.
2. Simuler θ^* à partir de la distribution candidate Normale Gamma.
3. Accepter $\theta^j = \theta^*$, avec une probabilité égale à $\min\left(1, \frac{w(\theta^*/.)}{w(\theta^{j-1}/.)}\right)$, où $w(\theta/.) = \frac{p(\theta/.)}{q(\theta/.)}$, avec $p(\theta/)$ dénote la loi a postériori conditionnelle dans (IV.14) et $q(\theta/.)$ est la densité candidate.
4. Accepter $\theta^j = \theta^{j-1}$ sinon.

Simulation de alpha

Le schéma en blocs de alpha est défini comme suit :

$$\alpha = (\alpha_1 \cdots \alpha_{k_1}, \alpha_{k_1+1} \cdots \alpha_{k_2}, \alpha_{k_2+1} \cdots, \cdots \alpha_{k_K}, \alpha_{k_K+1} \cdots \alpha_T),$$

où $k_1, k_2, \cdots, k_K$ sont les K points séparants les $(K+1)$ blocs.
Définir $\alpha_{B_l} = (\alpha_{k_{l-1}+1} \cdots \alpha_{k_l})', l = 1, \cdots, K+1$ avec $k_0 = 0$ et $\alpha_{k_0+1} = \alpha_1$, les étapes de simulation de alpha, insérées à l'itération j de l'échantillonneur de Gibbs sont :

1. Générer $\alpha_{B_1}^{(j)}/y, W, \alpha_{B_2}^{(j-1)}, \theta^{(j)}$.
2. Générer $\alpha_{B_l}^{(j)}/y, W, \alpha_{B_{l-1}}^{(j)}, \alpha_{B_{l+1}}^{(j-1)}, \theta^{(j)}$. pour $l = 2, 3, \cdots, K$
3. Générer $\alpha_{B_{K+1}}^{(j)}/y, W, \alpha_{B_K}^{(j)}, \theta^{(j)}$.

Pour chaque bloc α_{B_l} une approximation linéaire gaussienne du modèle à espace d'états non gaussien défini en (IV.1) et (IV.2) est produite. Ce modèle approximatif est défini par l'équation (IV.2) en combinaison avec l'équation de mesure approximative définie par :

$$\tilde{y}_i = \alpha_i + \tilde{\varepsilon}_i \tag{IV.15}$$

pour $i = 1, 2, \cdots, T$ où $\tilde{\varepsilon}_i \sim N(0, \tilde{H}_i)$, $\tilde{y}_i$ et$\tilde{H}_i$ sont des fonctions particulières de y_i et α_i.
L'approximation linéaire gaussienne est donc un modèle candidat à partir duquel α_{B_l} est généré.
En bref, le modèle approximatif est déterminé de telle manière que le mode

de la densité candidate associée à ce modèle $q(\alpha_{B_l}/y, W, \alpha_{B_{l+1}}, \theta)$, soit égal au mode de la densité conditionnelle a posteriori de α_{B_l} qui est basé sur le modèle non gaussien, $p(\alpha_{B_l}/y, W, \alpha_{B_{l+1}}, \theta)$. Générer à partir de $q(\alpha_{B_l}/y, W, \alpha_{B_{l+1}}, \theta)$ est implémenté en utilisant le filtre et le lissage de Kalman.
Le mode de la densité candidate $q(\alpha/y, W, \theta)$ est la solution de l'équation $\frac{\partial \ln q(\alpha/y, W, \theta)}{\partial \alpha} = 0$, de manière equivalente c'est la solution de l'équation $\frac{\partial \ln q(\alpha, y/W, \theta)}{\partial \alpha} = 0$. En considérant le modèle linéaire gaussien donné par (IV.15), l'hypothèse de la densité en (IV.3) de α_1 et l'équation d'état linéaire gaussienne, on obtient

$$\begin{aligned}
\ln q(\alpha, y/W, \theta) &= \text{constante} - \frac{1}{2}(\frac{1-\phi^2}{\sigma_\eta^2})(\alpha_1 - W_1\delta)^2 \\
&\quad -\frac{1}{2\sigma_\eta^2}\sum_{i=2}^{T}(\alpha_i - W_{i-1}\delta - \phi\alpha_{i-1})^2 \\
&\quad -\frac{1}{2}\sum_{i=1}^{T}\frac{(\tilde{y}_i - \alpha_i)^2}{\tilde{H}_i}
\end{aligned}$$

En dérivant par rapport à α_i, en ignorant le premier et le dernier état et en mettant le résultat égal à zéro, on obtient :

$$\begin{aligned}
\frac{\partial \ln q(\alpha, y/W, \theta)}{\partial \alpha_i} &= -\left(\frac{1}{\sigma_\eta^2}\right)(\alpha_i - W_{i-1}\delta - \phi\alpha_{i-1}) \\
&\quad +\left(\frac{\phi}{\sigma_\eta^2}\right)(\alpha_{i+1} - W_i\delta - \phi\alpha_i) \\
&\quad +\frac{(\tilde{y}_i - \alpha_i)}{\tilde{H}_i} \\
&= 0
\end{aligned}$$

avec $i = 2, ..., T-1$. Comme $q(\alpha/y, W, \theta)$ est gaussienne, donc la solution de l'équation précédente est la moyenne de $q(\alpha/y, W, \theta)$ qui peut obtenue en appliquant le filtre et le lissage de Kalman au modèle défini par les équations (IV.15) et (IV.2). De même, pour le modèle non gaussien, le mode de $p(\alpha/y, W, \theta)$ est la solution de $\frac{\partial \ln q(\alpha/y, W, \theta)}{\partial \alpha} = 0$ donc de l'équation $\frac{\partial \ln p(\alpha, y/W, \theta)}{\partial \alpha} = 0$. Sachant le modèle donné par (IV.1) et (IV.2) et la

distribution de l'état initial (IV.3) :

$$\begin{aligned}\ln q(\alpha, y/W, \theta) &= \text{constante} - \frac{1}{2}(\frac{1-\phi^2}{\sigma_\eta^2})(\alpha_1 - W_1\delta)^2 \\ &\quad -\frac{1}{2\sigma_\eta^2}\sum_{i=2}^{T}(\alpha_i - W_{i-1}\delta - \phi\alpha_{i-1})^2 \\ &\quad -\sum_{i=1}^{T} h(y_i/\alpha_i)\end{aligned}$$

où $h(y_i/\alpha_i) = -\ln p(y_i/\alpha_i)$. De la même manière, on dérive par rapport à α_i et on met le résultat égal à zéro :

$$\begin{aligned}\frac{\partial \ln p(\alpha, y/W, \theta)}{\partial \alpha_i} &= -\left(\frac{1}{\sigma_\eta^2}\right)(\alpha_i - W_{i-1}\delta - \phi\alpha_{i-1}) \\ &\quad +\left(\frac{\phi}{\sigma_\eta^2}\right)(\alpha_{i+1} - W_i\delta - \phi\alpha_i) \\ &\quad -\frac{\partial h(y_i/\alpha_i)}{\partial \alpha_i} \\ &= 0\end{aligned}$$

pour $i = 2, ..., T-1$. Le modèle approximatif (IV.15) est choisi de telle sorte que les deux équations précédentes aient la même solution. Pour le modèle SCD de densité conditionnelle exponentielle on trouve (pour plus de détails voir [19]) :

$$\tilde{H}_i = y_i^{-1}\exp(\alpha_i) \quad \text{(IV.16)}$$

$$\tilde{y}_i = \alpha_i - y_i^{-1}\exp(\alpha_i) + 1 \quad \text{(IV.17)}$$

La procédure itérative est basée sur les étapes suivantes :

1. Initialiser $\tilde{H}_i$ et $\tilde{y}_i$ à partir d'une valeur initiale de α_i.
2. Exécuter le filtre et le lissage de Kalman basé sur (IV.15) et (IV.2) pour produire le mode de $q(\alpha/y, W, \theta)$.
3. Substituer le mode de $q(\alpha/y, W, \theta)$ dans l'équation précédente et vérifier la condition du premier ordre.
4. Si la la condition du premier ordre n'est pas satisfaite recalculer $\tilde{H}_i$ et $\tilde{y}_i$ en utilisant les formules précédentes
5. Répéter à partir de l'étape 2 jusqu'à ce que la condition du premier ordre soit satisfaite.

6. Utiliser $\tilde{H}_i$ et $\tilde{y}_i$ ainsi calculés pour définir le modèle candidat pour α.

Les étapes de simulation de $\alpha_{B_l}, l = 1, 2, \cdots, K+1$, à l'itération j de l'échantillonneur de Gibbs sont :

1. Initialiser α_{B_l}
2. Implémenter la procédure itérative pour produire $\tilde{H}$ et $\tilde{y}$.
3. Définir l'équation de mesure comme (IV.15), pour le $(k_l - k_{l-1})_{\grave{e}m}$ élément dans le block α_{B_l}.
4. Générer une candidate $\alpha^*_{B_l}$ à partir de $q(\alpha_{B_l}/y, W, \alpha_{B_{l+1}}, \theta)$ en utilisant le filre et le lissage de Kalman.
5. Accepter $\alpha^{(j)}_{B_l} = \alpha^*_{B_l}$, avec une probabilité égale a $\min\left(1, \dfrac{\omega(y_{B_l}/\alpha^*_{B_l})}{\omega(y_{B_l}/\alpha^{(j-1)}_{B_l})}\right)$ où $\omega(y_{B_l}/\alpha_{B_l}) = \dfrac{p(y_{B_l}/\alpha_{B_l})}{q(y_{B_l}/\alpha_{B_l})}$ et y_{B_l} dénote le bloc de y correspondant à α_{B_l}. Notant que :

$$\frac{p(\alpha_{B_l}/y, W, \alpha_{B_{l-1}}, \alpha_{B_{l+1}}, \theta)}{q(\alpha_{B_l}/y, W, \alpha_{B_{l-1}}, \alpha_{B_{l+1}}, \theta)} \propto \frac{p(y_{B_l}/\alpha_{B_l})p(\alpha_{B_l}/W, \alpha_{B_{l-1}}, \alpha_{B_{l+1}}, \theta)}{q(y_{B_l}/\alpha_{B_l})q(\alpha_{B_l}/W, \alpha_{B_{l-1}}, \alpha_{B_{l+1}}, \theta)} \quad \text{(IV.18)}$$

$$\propto \frac{p(y_{B_l}/\alpha_{B_l})}{q(y_{B_l}/\alpha_{B_l})} \quad \text{(IV.19)}$$

car $p(\alpha_{B_l}/W, \alpha_{B_{l-1}}, \alpha_{B_{l+1}}, \theta) = q(\alpha_{B_l}/W, \alpha_{B_{l-1}}, \alpha_{B_{l+1}}, \theta)$

6. Accepter $\alpha^{(j)}_{B_l} = \alpha^{(j-1)}_{B_l}$ ailleurs.

IV.4.3 Le modèle SV

La simulation des états non observés dans le modèle SV est basée sur l'algorithme multimove qui est approprié pour les modèles de forme espace d'état. Cette simulation est effectuée en utilisant le filtre de Kalman en approximant l'équation de mesure non gaussienne par une mixture de lois gaussiennes. Les algorithmes de simulation associés avec ce dernier peuvent être intégrés facilement dans les échantillonneurs MCMC.
L'estimation du vecteur de paramètres θ du modèle SV se fait élément par élément contrairement au modèle SCD où on a estimé les paramètres en un seul bloc.

Simulation de h

Dans le but d'estimer le vecteur des variables latentes, l'expression (IV.5) (l'équation de mesure) doit être transformée, de telle sotre que celle ci soit

linéaire et gaussienne.
En élévant au carré et en prennant le logarithme naturel des deux cotés de l'équation, on obtient :

$$y_t^* = h_t + z_t, \tag{IV.20}$$

où $y_t^* = log(y_t^2 + c)$ et $z_t = log\varepsilon_t^2$
z_t est une distribution $log\chi^2$ qui n'a pas une forme standard. Kim, Shephard et Chib (1998) remarquent que z_t peut être approximée par une mixture discrete de densités normales.

$$f(z_t) = \sum_{i=1}^{K} q_i f_N(z_t/m_i - 1.2704, v_i^2),$$

est une mixture de K densités normales f_N avec des probabilités q_i, moyennes $m_i - 1.2704$ et variances v_i^2. Les constantes q_i, m_i, v_i^2 sont séléctionnées pour approximer la densité exacte de $log\varepsilon_t^2$.
Kim, Shephard et Chib (1998) choisissent une mixture avec sept composantes, car ils remarquent qu'au dela de K=7 il n'y a pas des changements importants dans les résultats.
Il faut noter que la densité mixture peut aussi être écrite en fonction des composantes d'une variable discète s telle que :

$$z_t/s_t = i \sim N(m_i - 1.2704, v_i^2) \tag{IV.21}$$

$$Pr(s_t = i) = q_i.$$

Cette représentation va être utilisée dans l'algorithme MCMC.
Les différents poids, moyennes et variances sont donnés dans la table suivante :

ω	$Pr(\omega = i)$	m_i	σ_i^2
1	0.00730	−10.12999	5.79596
2	0.10556	−3.97281	2.61369
3	0.00002	−8.56686	5.17950
4	0.04395	2.77786	0.16735
5	0.34001	0.61942	0.64009
6	0.24566	1.79518	0.34023
7	0.25750	−1.08819	1.26261

Dans le contexte MCMC, les modèles à mixture sont mieux simulés par l'expoitation de la représentation donnée en (IV.21). La densité a posteriori qui nous interesse est $\pi(s, h, \phi, \sigma_\eta^2, \mu/y^*)$, où $s = (s_1, \cdots, s_T)$. Dans ce cas, h et s peuvent être simulés séparement dans un bloc.
Les étapes de l'échantillonneur de Gibbs sont les suivantes :

1. Initialiser s, ϕ, σ_η^2 et μ.
2. Générer h à partir de $h/y^*, s, \phi, \sigma_\eta^2, \mu$
3. Générer s à partir de $s/y^*, h$.
4. Générer ϕ, σ_η^2, μ.
5. Revenir en (2) jusqu'à convergence.

$y^*/s, \phi, \sigma_\eta, \mu$ est une série chronologique gaussienne. La littérature des séries chronologiques nomme ces modèles modèles partiellement non gaussiens ou conditionnellement gaussiens. Cette structure particulière de ce modèle implique qu'on peut générer $h/y^*, s, \phi, \sigma_\eta, \mu$ en utilisant le filtre de Kalman et le lissage.
La génération de s à partir de $s/y^*, h$, se fait indépendamment en générant chaque s_t en utilisant la fonction de masse de probabilité :

$$Pr(s_t = i/y^*, h_t) \propto q_i f_N(y_t^*/h_t + m_i - 1.2704, v_i^2)$$

Simulation de sigma

Si on suppose une loi a priori conjuguée $\sigma_\eta^2/\phi, \mu \sim IG(\sigma_r/2, S_\sigma/2)$, alors σ_η^2 est généré à partir de :

$$IG\left(\frac{n+\sigma_r}{2}, \frac{S_\sigma + (h_t-\mu)^2(1-\phi^2) + \sum_{t=1}^{n-1}((h_{t+1}-\mu)-\phi(h_t-\mu))^2}{2}\right) \tag{IV.22}$$

où IG dénote la distribution Inverse-Gamma.

Simulation de phi

Soit $\phi = 2\phi^* - 1$ où ϕ^* est une Beta de paramètres $(\phi^{(1)}, \phi^{(2)})$, notre loi a priori de ϕ est :

$$\pi(\phi) \propto \{\frac{(1+\phi)}{2}\}^{\phi^{(1)}-1}\{\frac{(1-\phi)}{2}\}^{\phi^{(2)}-1}, \phi^{(1)}, \phi^{(2)} > \frac{1}{2}$$

et a un support sur l'intervalle $(-1, 1)$ avec une moyenne a priori égale à $2\phi^{(1)}/(\phi^{(1)} + \phi^{(2)}) - 1$.
Sous la loi a priori spécifiée, la densité a postériori conditionnelle de ϕ est proportionnelle à :

$$\pi(\phi) f(h/\mu, \phi, \sigma_\eta^2)$$

où

$$log(f(h/\mu,\phi,\sigma_\eta^2) \propto \; - \; \frac{(h_1-\mu)^2(1-\phi^2)}{2\sigma_\eta^2} + \frac{1}{2}log(1-\phi^2) \quad \text{(IV.23)}$$

$$- \; \frac{\sum_{t=1}^{n-1}(h_{t+1}-\mu)-\phi(h_t-\mu)^2}{2\sigma_\eta^2} \quad \text{(IV.24)}$$

Comme la densité log posteriori n'a pas une forme standard, une approche pour simuler la loi a postériori est d'utiliser l'algorithme de MH. Kim,Shephard et Chib (1998) utilisent une densité cadidate normale de paramètre :

$$\hat{\phi} = \sum_{t=1}^{T-1}(h_{t+1}-\mu)(h_t-\mu)/\sum_{t=1}^{T-1}(h_t-\mu)^2$$

et de variance :

$$V_\phi = \sigma_\eta^2\{\sum_{t=1}^{T-1}(h_t-\mu)^2\}^{-1}$$

Accepter la valeur ϕ^* simulé à partir de la candidate normale avec une probabilité égale à :

$$\exp\{g(\phi^*) - g(\phi^{(j-1)})\}$$

où

$$g(\phi) = log\pi(\phi) - \frac{(h_t-\mu)^2(1-\phi^2)}{2\sigma_\eta^2} + \frac{1}{2}log(1-\phi^2)$$

Si ϕ^* est rejété, on prend $\phi^{(j)}\phi^{(j-1)}$.

Simulation de mu

μ est généré à partir de la densité conditionnelle :

$$\mu/h,\phi,\sigma_\eta^2 \sim N(\hat{\mu},\sigma_\mu^2)$$

où

$$\hat{\mu} = \sigma_\mu^2\left\{\frac{(1-\phi^2)}{\sigma_\eta^2}h_1 + \frac{(1-\phi)}{\sigma_\eta^2}\sum_{t=1}^{n-1}(h_{t+1}-\phi h_t)\right\}$$

et

$$\sigma_\mu^2 = \sigma_\eta^2\{(n-1)(1-\phi)^2 + (1-\phi)^2\}^{-1}$$

IV.4.4 Méthodes MCMC pour les processus de diffusion

Dans ce qui suit, nous allons décrire un plan d'échantillonnage MCMC qui permet de simuler à partir de la distribution a posteriori les paramètres entrant dans les fonctions de dérive et de diffusion. Le processus d'intérêt est :

$$dY_t = \mu(Y_t;\theta)dt + \sigma(Y_t;\theta)dW_t \tag{IV.25}$$

où $\mu : R^d \times \Theta \to R^d$ et $\sigma : R^d \times \Theta \to R^{d\times d}$ pour $\Theta \subseteq R^k$.
On suppose que le processus génère des observations en temps discret, notre objectif est de faire l'inférence du vecteur de paramètres, θ. Le processus de diffusion Y_t et le mouvement Brownien sont supposés de dimension d. On suppose aussi :

1. Les fonctions de dérive et de diffusion satisfont la condition de Lipschitz et sont bornées linéairement

$$\|\mu(x;\theta) - \mu(y;\theta)\| + \|\sigma(x;\theta) - \sigma(y;\theta)\| \leq C\|x - y\| \tag{IV.26}$$

et

$$\|\mu(x;\theta)\|^2 + \|\sigma(x;\theta)\|^2 \leq C^2(1 + \|x\|^2) \tag{IV.27}$$

où C est une constante positive et $\|.\|$ représente la norme Euclidienne.

2. $\sigma(x;\theta)\sigma(x;\theta)'$ est définie positive pour tout $x \in R^d$, où $'$ représente la matrice transposée.
3. On a un échantillon d'observations $Y_{t,j}, t = 0, 1, 2, ..., T,\ j = 1, 2, ..., d_1$ à utiliser dans l'estimation où $d_1 \leq d$.

L'hypothèse 1 est suffisante pour assurer l'existence d'une forte (unique) solution $Y_t, t \in [0, T]$ de (IV.25). Il suffit que les conditions soient vérifiées avec probabilité 1. L'hypothèse 2 est nécessaire pour avoir une fonction de vraisemblance bien définie. Conformément à l'hypothèse 3, le processus Y_t peut contenir des composantes observées et inobservées. Si Y_t contient des composantes inobservées, le processus (IV.25) définit un modèle à espace d'état non linéaire à temps continu. Un exemple de modèles avec des composantes inobservées est le modèle SV. Donc il est utile d'écrire $Y_t = (X_t, Z_t)$, où X définie la partie observable du système (donc $X = r$ selon les notations précedentes) et Z (appelée aussi variable d'état) désigne la partie non observée. X_t et Z_t sont supposés de dimensions d_1 et d_2 respectivement, alors $d = d_1 + d_2$. On travaille avec la version discrétisée de (IV.25) donnée par :

$$\Delta Y_t = \mu(Y_t;\theta)\Delta t + \sigma(Y_t;\theta)\Delta W_t \tag{IV.28}$$

où ΔW_t est un vecteur aléatoire iid $N(0, I_d\Delta t)$ de dimension d, où I_d représente la matrice identité de dimension d.
On suppose que les observations sont disponibles en temps entiers, mais supposer $\Delta t = 1$ n'est pas suffisant donc on propose de mettre $\Delta t = 1/m$ pour un certain m positif (voir [8]). Ceci assure que le biais introduit par la discritisation peut être rendue arbitrairement petit. D'autre part, ceci introduit un problème de valeurs manquantes. On divise l'intervalle de temps $[0, T]$ en $n = mT$ points equidistants $0 = t_0 < t_1 < ... < t_{n-1} < t_n = T$. Ceci implique que X_{t_i} est une observation pour tout i entier multiple de m. Pour les variables non observées Z, les m observations sont manquantes dans laps de temps $(t, t + 1]$. Au total, cette configuration implique que $d_1T(m - 1)$ données sont manquantes dans la partie observable du système, tandis que d_2Tm points de la partie non observées sont manquantes.
L'idée essentielle est de substituer les données manquantes, Y_{t_i} avec la simulation, $\hat{Y}_{t_i}$. Soit $\hat{Y}$ une $d \times n$ matrice contenant toutes les données (observées et simulées appelées aussi données augmentées).

$$\hat{Y} = \begin{bmatrix} X_{1,t_0} & \hat{X}_{1,t_1} & \cdots & X_{1,t_m} & \hat{X}_{1,t_{m+1}} & \cdots & X_{1,t_n} \\ X_{2,t_0} & \hat{X}_{2,t_1} & & X_{2,t_m} & \hat{X}_{2,t_{m+1}} & \cdots & X_{2,t_n} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ X_{d_1,t_0} & \hat{X}_{d_1,t_1} & \cdots & X_{d_1,t_m} & \hat{X}_{d_1,t_{m+1}} & \cdots & X_{d_1,t_n} \\ \hat{Z}_{1,t_0} & \hat{Z}_{1,t_1} & \cdots & \hat{Z}_{1,t_m} & \hat{Z}_{1,t_{m+1}} & \cdots & \hat{Z}_{1,t_n} \\ \vdots & \vdots & & \vdots & \vdots & \ddots & \vdots \\ \hat{Z}_{d_2,t_0} & \hat{Z}_{d_2,t_1} & \cdots & \hat{Z}_{d_2,t_m} & \hat{Z}_{d_2,t_{m+1}} & \cdots & \hat{Z}_{d_2,t_n} \end{bmatrix}$$

$\hat{Y}_i$ représente la $i^{ème}$ colonne de $\hat{Y}$.
La densité a posteriori conjointe est donnée par :

$$\pi(\hat{Y}, \theta) \propto \prod_{i=1}^{n} p(\hat{Y}_i/\theta)p(\theta) \tag{IV.29}$$

où $p(\theta)$ est la densité a priori du paramètre et

$$p(\hat{Y}_i/\theta) = |\sigma_{i-1}^{-2}|^{1/2} \exp\{-\frac{1}{2}\|(\Delta\hat{Y}_i - \mu_{i-1}\Delta t)\sigma_{i-1}^{-1}(\Delta t)^{-1/2}\|^2\} \tag{IV.30}$$

On définit $\Delta\hat{Y}_i = \hat{Y}_i - \hat{Y}_{i-1}$ et on utilise les notations abrégées $\mu_i = \mu(\hat{Y}_i; \theta)$, $\sigma_i = \sigma(\hat{Y}_i; \theta)$ et $\sigma_{i-1}^{-2} = (\sigma_{i-1}\sigma'_{i-1})^{-1}$, où $|\sigma_{i-1}^{-2}|$ définit le déterminant de σ_{i-1}^{-2}. Notons que toutes les densités a posteriori d'intérêt (en particulier $\pi(\theta/\hat{Y}_i)$) sont proportionnelles à (IV.29).

L'objectif principal de cette analyse est d'obtenir une séquence $\{\theta^{(h)}\}_{h=1}^{M}$ du vecteur de paramètres à partir de la densité a posteriori marginale en utilisant l'échantillonneur de Gibbs.
L'inclusion des données non observées dans la configuration du modèle doit être traitée spécifiquement. Une façon de traiter ces données manquantes est de formuler la densité conjointe des paramètres aussi bien que pour les données observées que pour les données non observées

La densité a posteriori conditionnelle des données manquantes

Simuler toutes les observations manquantes en une opération est impossible en raison de la dimension de la densité associée. Ainsi il est nécessaire de procéder avec des blocs de dimensions plus petites. Plusieurs manières sont possibles pour choisir les blocs, on prend par exemple le cas où l'échantillonneur de Gibbs met à jour $\hat{Y}$ colonne par colonne. Pour ce faire on doit utiliser la densité a posteriori conditionnelle de la $i^{ème}$ colonne de $\hat{Y}$. Cette densité est définie par la relation suivante :

$$\pi(\hat{Y}_i/\hat{Y}_{-i},\theta) \propto p(\hat{Y}_i/\hat{Y}_{i-1},\hat{Y}_{i+1},\theta) \tag{IV.31}$$

où $\hat{Y}_i/\hat{Y}_{i-1},\hat{Y}_{i+1},\theta$ a pour densité :

$$\left|\sigma_{i-1}^{-2}\right|^{1/2}\left|\sigma_i^{-2}\right|^{1/2} \times \exp\left\{-\frac{1}{2}\left\|(\Delta\hat{Y}_i-\mu_{i-1}\Delta t)\sigma_{i-1}^{-1}(\Delta t)^{-1/2}\right\|^2\right. \tag{IV.32}$$

$$\left.-\frac{1}{2}\left\|(\Delta\hat{Y}_{i+1}-\mu_i\Delta t)\sigma_i^{-1/2}(\Delta t)^{-1}\right\|^2\right\} \tag{IV.33}$$

où $\hat{Y}_{-i}$ représente tous les éléments de $\hat{Y}$ sauf la $i^{ème}$ colonne. L'équation précédente résulte de (IV.29) parce que tous les termes du produit de (IV.29) où $\hat{Y}_i$ n'intervient pas peuvent être absorbés par la constante de proportionnalité.
Il est clair que la densité (IV.33) peut avoir diverses formes en fonction de la forme des fonctions de dérive et de diffusion.
Il est important de distinguer entre la densité conditionnelle $p(\hat{Y}_i/\hat{Y}_{i-1},\hat{Y}_{i+1},\theta)$ et la densité de transition $p(\hat{Y}_i/\hat{Y}_{i-1},\theta)$. La dernière définie la densité conditionnelle de $\hat{Y}_i$ par rapport à $\hat{Y}_{i-1}$ tandis que la première est conditionnée par rapport à la valeur précédente $\hat{Y}_{i-1}$ et la valeur suivante $\hat{Y}_{i+1}$. A l'itération h de l'échantillonneur de Gibbs, on simule :

$$\hat{Y}_i^{(h)} \sim \pi(\hat{Y}_i/\hat{Y}_{i-1}^{(h)},\hat{Y}_{i+1}^{(h-1)},\theta) \tag{IV.34}$$

pour tout $i = 0,1,...,n$. Naturellement, chaque fois que i est un multiple de m, on simule uniquement les d_2 éléments correspondant aux états non

observés Z. Pour ce faire il suffit de connaitre $\pi(\hat{Y}_i/\hat{Y}_{i-1}, \hat{Y}_{i+1}, \theta)$ puisque

$$p(Z_{1,i}, ..., Z_{d_2,i}/\hat{Y}_{i-1}, \hat{Y}_{i+1}, X_{1,i}, X_{2,i}, ..., X_{d_1,i}, \theta) \propto p(\hat{Y}_i/\hat{Y}_{i-1}, \hat{Y}_{i+1}, \theta) \quad \text{(IV.35)}$$

par le théorème de Bayes. Par conséquent, on utilise la même expression pour la densité cible pour simuler Z_i pour X_i fixé en une valeur observée.

La prochaine étape de la mise en oeuvre de l'échantillonneur de Gibbs est de concevoir une méthode pour simuler $\hat{Y}_i/\hat{Y}_{i-1}, \hat{Y}_{i+1}, \theta$. On commence par le cas d'un processus unidimentionnel $Y_t \in R$ avec les fonctions de dérive et de diffusion constantes, $\mu(x) = \mu$, $\sigma(x) = \sigma$ pour tout $x \in R$. Donc le vecteur de paramètres est $\theta = (\mu, \sigma)$, on obtient le résultat suivant :

Proposition IV.4.1. *(voir [8])*
Pour un processus scalaire, $Y_t \in R$ avec des fonctions de dérive et de diffusion constantes, on a :

$$\hat{Y}_i/\hat{Y}_{i-1}, \hat{Y}_{i+1}, \theta \sim N(\frac{1}{2}(\hat{Y}_{i-1} + \hat{Y}_{i+1}), \frac{1}{2}\sigma^2\Delta t) \quad \text{(IV.36)}$$

Passons maintenant au cas genéral où les fonctions de dérive et de diffusion sont arbitraires. Dans ce cas on utilise l'algorithme hybride acceptation/rejet Metropolis-Hastings (AR-MH) (voir annexe B). Cet algorithme fonctionne mieux si pour une densité proposale q, il existe une constante positive C, telle que Cq forme une bonne approximation de la densité cible p. On propose comme densité proposale, $q(x/Y_{i-1}, Y_{i+1}, \theta)$ une densité normale de moyenne $m_i = 1/2(\hat{Y}_{i-1} + \hat{Y}_{i+1})$ et de matrice de variance covariance $1/2\sigma(\hat{Y}_{i-1})\sigma(\hat{Y}_{i-1})'\Delta t$.

Proposition IV.4.2. *(voir [8])*
Pour des fonctions de dérive et de diffusion qui satisfont les hypothèses 1-3 et $\Delta t \to 0$, on a :

$$\Delta t^{-1/2}(\hat{Y}_i - \frac{1}{2}(\hat{Y}_{i-1} + \hat{Y}_{i+1})) \Rightarrow N(0, \frac{1}{2}\sigma^2_{i-1}) \quad \text{(IV.37)}$$

En vertu du résultat précédent, on peut dire que lorsque Δt est assez petit, on a approximativement :

$$\hat{Y}_i/\hat{Y}_{i-1}, \hat{Y}_{i+1} \sim N(\frac{1}{2}(\hat{Y}_{i-1} + \hat{Y}_{i+1}), \frac{1}{2}\sigma^2_{i-1}\Delta t)$$

une autre condition pour la mise en oeuvre efficace de l'algorithme AR-MH est de trouver une constante C telle que Cq soit aussi proche de la distribution cible p que possible, on suggère de prendre C comme le ratio des deux

distributions évaluées en m_i :

$$\begin{aligned} C &= \frac{p(m_i/\hat{Y}_{i-1}, \hat{Y}_{i+1}, \theta)}{q(m_i/\hat{Y}_{i-1}, \hat{Y}_{i+1}, \theta)} \\ &= p(m_i/\hat{Y}_{i-1}, \hat{Y}_{i+1}, \theta) \left| (2\pi)^d \frac{1}{2} \sigma(\hat{Y}_{i-1}) \sigma(\hat{Y}_{i-1})' \Delta t \right|^{-1/2} \end{aligned}$$

Ce choix est motivé par le fait que si p a approximativement la même forme que q, alors le rapport des deux densités est constant pour tout x et donc pour $x = m_i$

La densité a posteriori conditionnelle des paramètres

La seconde étape majeure est de considérer $\hat{Y}^{(h)}$ constant et simuler $\theta \sim \pi(\theta/\hat{Y}^{(h)})$ qui est déduite de (IV.29). Dans l'exemple suivant on supposera des lois a priori non informatives.
Pour obtenir des échantillons de θ conditionnellement à $\hat{Y}$, on utilise deux blocs dans l'échantillonneur de Gibbs contenant les paramètres $\{\theta_r, \kappa_r\}$ et les paramètres du processus log volatilité $\{\theta_z, \kappa_z\}$ et σ_z. Donc les coefficients de dérive et de diffusion sont respectivement :

$$\mu(Y_i; \theta) = \begin{bmatrix} \theta_r + \kappa_r X_i \\ \theta_z + \kappa_z Z_i \end{bmatrix}$$

$$\sigma(Y_i; \theta) = \begin{bmatrix} \exp(\frac{1}{2} Z_i) X_i^{\beta} & 0 \\ 0 & \sigma_z \end{bmatrix}$$

$\pi(\theta/\hat{Y})$ peut être écrit comme le produit des deux densités :

$$\pi(\theta/\hat{Y}) \propto p(\theta_r, \kappa_r/\hat{Y}) p(\theta_z, \kappa_z, \sigma_z/\hat{Y}) \tag{IV.38}$$

où

$$p(\theta_r, \kappa_r/\hat{Y}) = \prod_{i=1}^{n} \exp(-\frac{1}{2} Z_i) r_i^{-1/2} \Delta t^{-1/2} \times \exp\{-\frac{(\Delta r_i - (\theta_r + \kappa_r r_i)\Delta t)^2}{2 \exp(Z_i) r_i \Delta t}\} \tag{IV.39}$$

et

$$p(\theta_z, \kappa_z, \sigma_z/\hat{Y}) = \prod_{i=1}^{n} \sigma_z^{-1} \Delta t^{-1/2} \times \exp\{-\frac{(\Delta Z_i - (\theta_z + \kappa_z Z_i)\Delta t)^2}{2\sigma_z^2 \Delta t}\} \tag{IV.40}$$

Pour le modèle SV donné par (IV.11), les échantillons MCMC simulés pour chacun des paramètres sont très corrélés donc la convergence est très lente.

Dans ce qui suit nous présentons une reparametrisation du modèle pour réduire les dépendances et le temps d'execution des chaines.
On peut écrire :

$$\begin{aligned} dr_t &= \kappa_r(\theta_r - r_t)dt + \exp(\frac{1}{2}k)\exp(\frac{1}{2}z_t - \frac{1}{2}k)r_t^{\beta}dW_{1,t} \\ &= \kappa_r(\theta_r - r_t)dt + \sigma_r \exp(\frac{1}{2}Z_t^*)r_t^{\beta}dW_{1,t} \end{aligned}$$

où $\sigma_r = \exp(\frac{1}{2}k)$ où k est une constante.
Donc on peut considérer la parametrisation :

$$\begin{aligned} dr_t &= \kappa_r(\theta_r - r_t)dt + \sigma_r \exp(\frac{1}{2}z_t^*)r_t^{\beta}dW_{1,t} \\ dZ_t^* &= \kappa_z Z_t^* dt + \sigma_z dW_{2,t} \end{aligned}$$

mais cette parametrisation a certains incovénients (voir Eraker (2001)), pour les éliminer on considère le système suivant :

$$\begin{aligned} dr_t &= \kappa_r(\theta_r - r_t)dt + \sigma_r \exp(\frac{1}{2}Z_t - \frac{1}{2}k)r_t^{\beta}dW_{1,t} \\ dZ_t &= \kappa_z(k - Z_t)dt + \sigma_z dW_{2,t} \end{aligned}$$

Notons que la distribution conjointe a posteriori de tous les paramètres ne changera pas par l'introdution d'une constante k.
On peut mettre $k = \bar{Z}^{(t)}$ où $\bar{Z}^{(t)}$ est la moyenne du processus latent à l'itération t, donc l'algorithme de simulation à l'itération t est le suivant :

1. Simuler les variables latentes $Z_i^{(t)}$ pour $i = 0, 1, ..., n$ et $r_i^{(t)}$ pour $i = 1, ..., m-1, m+1, ..., 2(m-1), 2(m+1), ..., n-1$ en utilisant l'algorithme hybride AR-MH.
2. Simuler $\kappa_r^{(t)}$ et , $\theta_r^{(t)}$ à partir d'une distribution normale et $\sigma_r^{2(t)}$ à partir d'une distribution inverse gamma.
3. Simuler $\beta^{(t)}$ en utilisant l'algorithme de MH avec une densité normale comme proposale.
4. Prendre $Z_i^{(t)} \leftarrow Z_i^{(t)} - \bar{Z}^{(t)}$ et $\sigma_r^{(t)} \leftarrow \sigma_r^{(t)} \exp(\frac{1}{2}\bar{Z}^{(t)})$
5. Simuler κ_z et σ_z^2 à partir des distributions normale et gamma.

IV.5 Conclusion

Dans ce chapitre, nous avons introduit deux modèles a espace d'état en temps discret à savoir le modèle SCD et SV. Nous avons aussi introduit le modèle SV en temps continu. L'estimation des paramètres de ces modèles a été étudiée dans un cadre bayésien en utilisant les méthodes MCMC.

Chapitre V

Application

V.1 Introduction

Dans ce chapitre nous allons étudier l'efficience des méthodes bayésiennes dans l'estimation des modèles à espace d'états en utilisant les deux modèles non gaussiens particuliers : SCD et SV. Nous nous intéréssons particulièrement au comportement de la convergence de ces méthodes en faisant varier la taille des données T afin de voir si l'augmentation de l'information disponible au départ (le nombre d'observations T) accelère la convergence. Les critères de comparaison utilisés sont d'une part le facteur d'inefficacité et d'autres part les différents diagnostics utilisés dans les méthodes MCMC (diagnostics de Gelman, Geweke, Raftery et Heidelberg).
Cette étude a été réalisée en utilisant des données artificielles (simulées). Ces données ont été simulées en utilisant le logiciel libre R 2.8.0. Les différents diagnostics ont été implémentés via le package CODA.

V.2 Le modèle à durée conditionnelle stochastique (SCD)

Pour le modèle SCD les données artificielles sont simulées pour les valeurs $\mu = -0.05, \phi = 0.95, \sigma^2 = 0.01$. Nous avons générées trois chaines avec différentes valeurs initiales qui sont respectivement pour les trois chaines $(\mu = 0, \phi = 0.6, \sigma^2 = 0.09)$, $(\mu = 0.5, \phi = 0.8, \sigma^2 = 0.04)$ et $(\mu = -0.5, \phi = 0.9, \sigma^2 = 0.0225)$. Les différentes chaines ont été générées pour les valeurs de $T = (500, 1000, 2000, 3000)$ et pour les longueurs de $M = (15000, 25000, 35000, 45000, 55000)$. Afin de réduire l'influence des valeurs initiales nous avons pris une période de burn-in de 5000 itérations pour toutes les chaines.

V.2.1 Diagnostics visuels (graphiques)

Pour l'étude de la convergence des chaines vers les vraies valeurs des paramètres on commence par une analyse graphique à savoir la trace et le graphe des autocorrélations afin de voir l'évolution des valeurs de la chaine ainsi que l'évolution des autocorrélations.

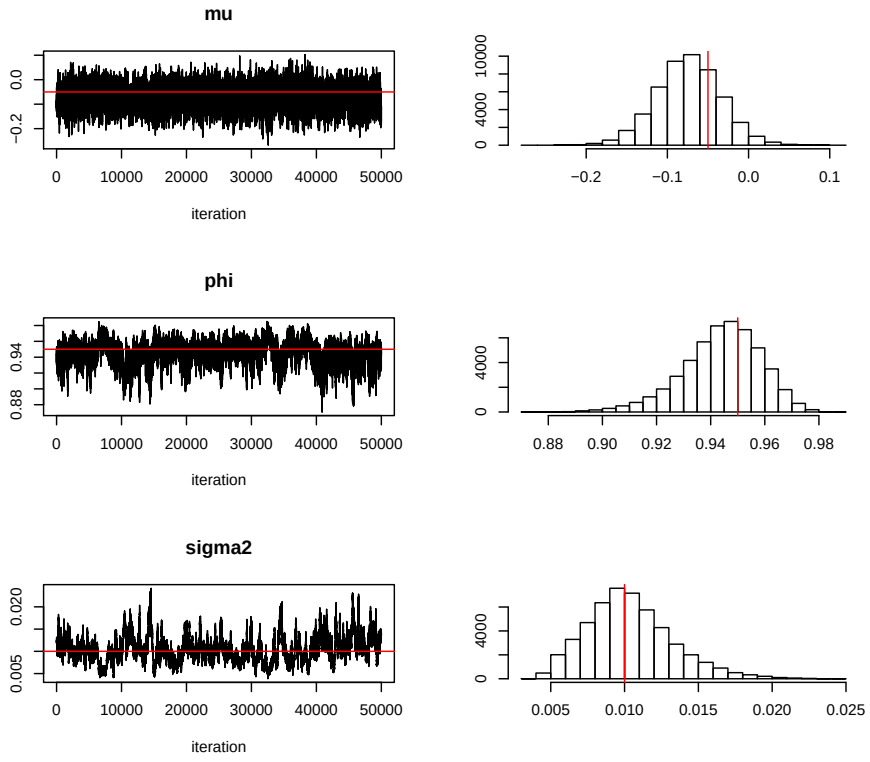
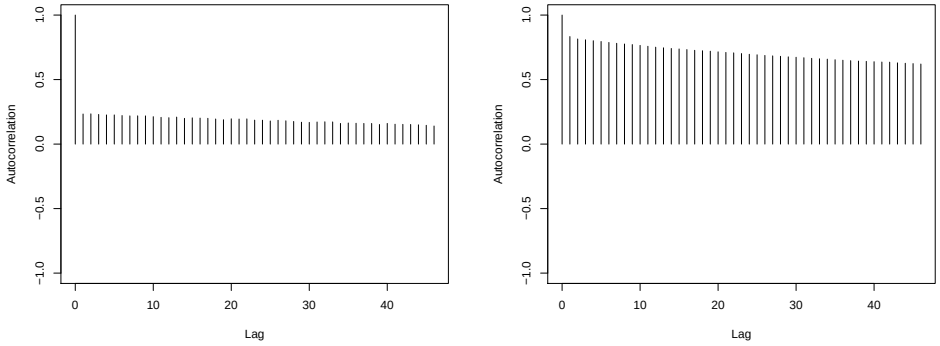


FIGURE V.1 – Traces et histogrammes des trois paramètres.

La figure (V.1), montre que les valeurs des différentes réalisations varient autour des vraies valeurs (indiquées en rouge dans la figure) pour les trois paramètres aussi bien que dans les traces que dans les histogrammes.

Les figures (V.2) et (V.3) des autocorrélations indiquent de fortes autocorrélations pour les paramètres ϕ et σ^2 et plus faibles pour μ. Donc les réalisations restent encore dépendantes surtout celles de ϕ et σ^2.



(a) (b)

FIGURE V.2 – Corrélogrammes de mu et phi.

Tableau des moyennes

Les estimateurs bayésiens (moyennes) sont calculés à partir des distributions a posteriori marginales.
Le tableau V.1 porte les valeurs moyennes de chaque paramètre calculées à partir des trois chaines générées pour les différentes valeurs de M et de T.

	M	10^3	2×10^3	3×10^3	4×10^3	5×10^3
$T = 500$	μ	-0.047	-0.047	-0.046	-0.045	-0.045
	ϕ	0.917	0.916	0.915	0.916	0.916
	σ^2	0.016	0.017	0.016	0.016	0.016
$T = 1000$	μ	-0.031	-0.031	-0.031	-0.030	-0.029
	ϕ	0.932	0.931	0.933	0.933	0.934
	σ^2	0.014	0.014	0.014	0.014	0.013
$T = 2000$	μ	-0.051	-0.051	-0.051	-0.051	-0.051
	ϕ	0.953	0.954	0.954	0.954	0.954
	σ^2	0.011	0.011	0.011	0.011	0.011
$T = 3000$	μ	-0.020	-0.020	-0.020	-0.020	-0.021
	ϕ	0.946	0.945	0.946	0.947	0.946
	σ^2	0.009	0.012	0.011	0.010	0.010

TABLE V.1 – Tableau des moyennes pour le modèle SCD

Les figures (V.4), (V.5), (V.6) et (V.7) montrent l'évolution des valeurs moyennes par rapport aux nombres d'itérations M et par rapport à T.

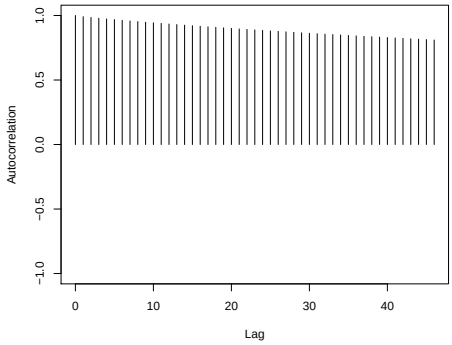
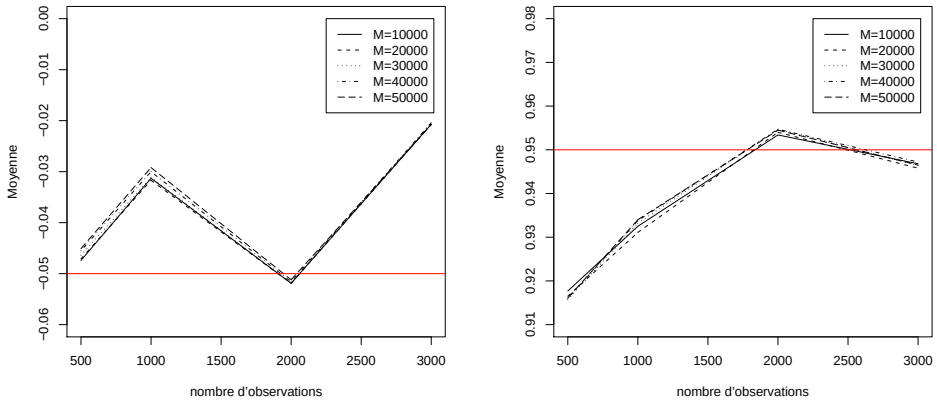


FIGURE V.3 – Corrélogramme de sigma.



(a) (b)

FIGURE V.4 – Evolution de la moyenne par rapport à T pour mu et phi.

D'après ces figures on constate que les valeurs moyennes restent stationnaires par rapport à M tandis qu'on remarque une amélioration des estimateurs lorsque T augmente, cette amélioration est plus importante pour ϕ et σ^2 que pour μ.

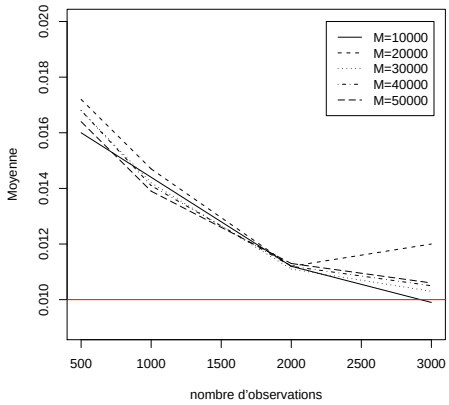
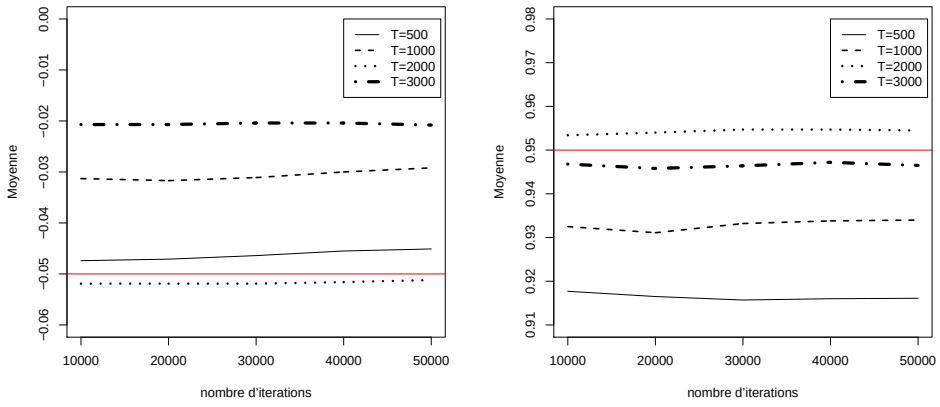


FIGURE V.5 – Evolution de la moyenne par rapport à T pour sigma.



(a) (b)

FIGURE V.6 – Evolution de la moyenne par rapport à M pour mu et phi.

V.2.2 Diagnostics de convergence

Facteur d'inefficacité

Le tableau V.2 comporte les valeurs moyennes du facteur d'inefficacité calculés à partir des trois chaines.

Les figurent (V.8), (V.9), (V.10) et (V.11) donnent l'évolution du IF par

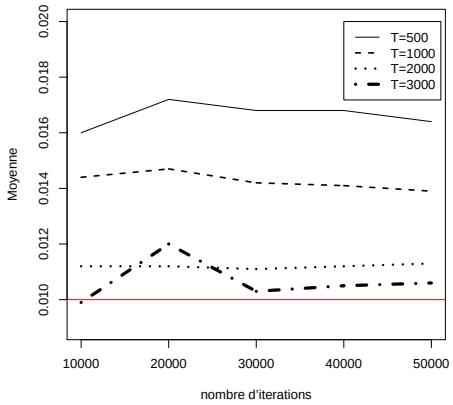


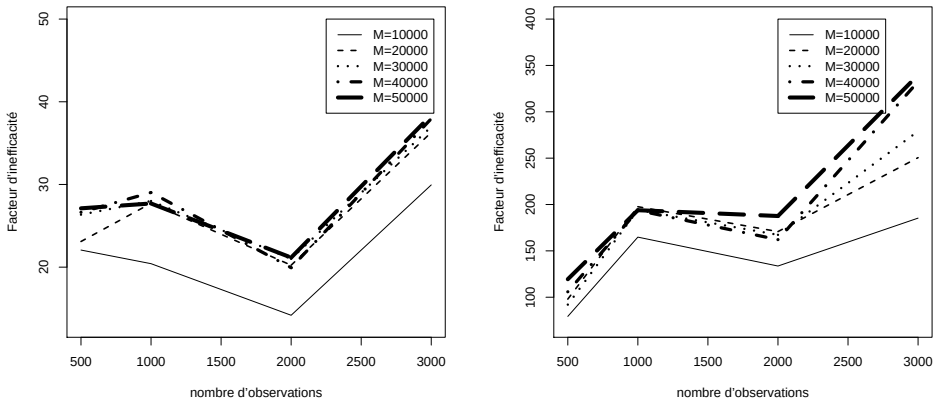
FIGURE V.7 – Evolution de la moyenne par rapport à M pour sigma.

	M	10^3	2×10^3	3×10^3	4×10^3	5×10^3
$T = 500$	μ	22.06	23.10	26.34	26.63	27.11
	ϕ	79.29	97.95	92.05	105.8	119.4
	σ^2	180.1	193.5	191.7	196.3	210.6
$T = 1000$	μ	20.42	27.72	27.99	29.02	27.69
	ϕ	164.8	101.2	194.1	193.6	193.8
	σ^2	324.1	361.7	356.8	358.7	350.6
$T = 2000$	μ	14.18	20.23	21.01	19.93	21.16
	ϕ	133.6	170.7	167.2	162.1	187.6
	σ^2	276.7	341.6	329.6	312.7	353.8
$T = 3000$	μ	29.94	36.28	36.97	37.95	38.38
	ϕ	185.3	250.5	279.1	332.5	339.6
	σ^2	360.5	432.3	464.2	495.2	497.5

TABLE V.2 – Evolution du facteur d'inefficacité pour le modèle SCD

rapport à M et par rapport à T pour les différents paramètres.

D'après ces résultats on remarque que :
- les valeurs les plus faibles du IF sont obtenues par μ, plus importantes pour ϕ et les plus grandes valeurs sont atteintes pour σ^2.
- le IF varie de la même manière et dans le même sens pour les trois paramètres.
- le IF ne varie pas d'une manière importante par rapport à M pour tous les paramètres (pratiquement stationnaire).



(a) (b)

FIGURE V.8 – Evolution du facteur d'inefficacité par rapport à T pour mu et phi.

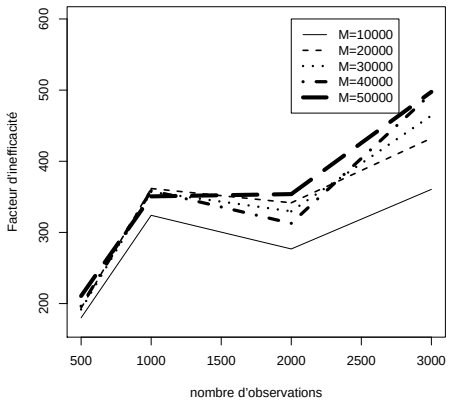
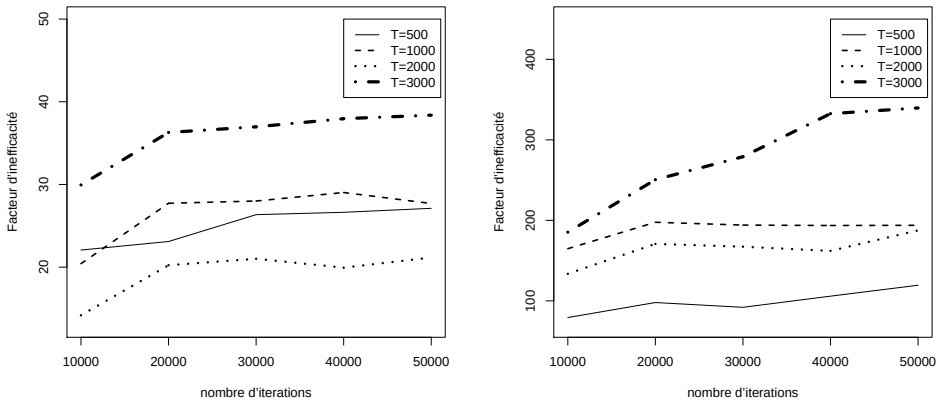


FIGURE V.9 – Evolution du facteur d'inefficacité par rapport à T pour sigma.

– le IF augmente lorsque T augmente faiblement pour μ et très fortement pour ϕ et σ^2.

Les valeurs imoptantes du IF indiquent que les valeurs des différentes chaines restent très corrélées entre elles. Donc plus T est grand plus il faut plus d'itérations.



(a) (b)

FIGURE V.10 – Evolution du facteur d'inefficacité par rapport à M pour mu et phi.

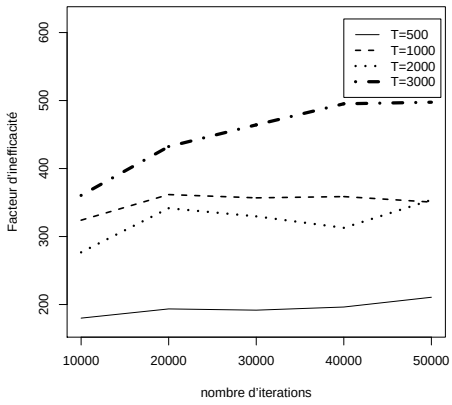


FIGURE V.11 – Evolution du facteur d'inefficacité par rapport à M pour sigma.

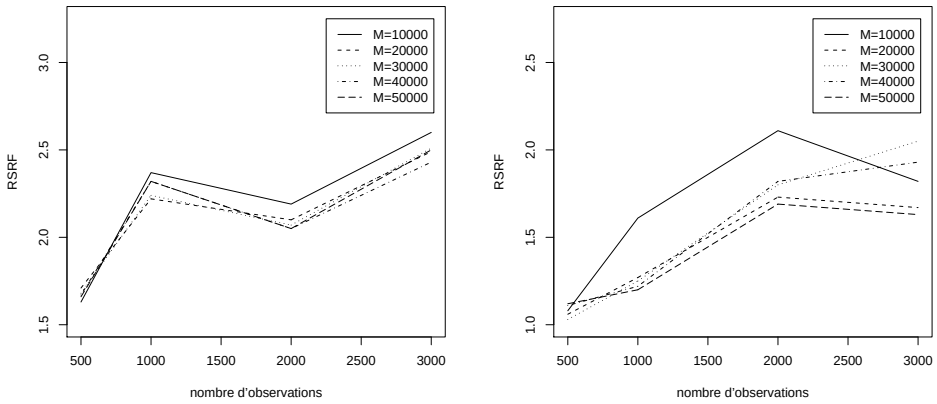
Diagnostic de Gelman

Le diagnostic de Gelman nous fournit le facteur de réduction d'échelle potentielle (Potential Scale Reduction Factor (PSRF)) noté R. Pour que la chaine converge il faut que R soit assez petit (proche de 1).

Le tableau V.3 donne les valeurs de R obtenues en variant T et M simultanément. Les figures (V.12), (V.13), (V.14) et (V.15) montrent l'évolution de R par rapport à T pour les différentes valeurs de M ainsi que son évolution par rapport à M pour toutes les valeurs de T.

	M	10^3	2×10^3	3×10^3	4×10^3	5×10^3
$T = 500$	μ	1.63	1.71	1.68	1.67	1.66
	ϕ	1.08	1.06	1.03	1.11	1.12
	σ^2	1.06	1.06	1.02	1.12	1.13
$T = 1000$	μ	2.37	2.22	2.24	2.32	2.32
	ϕ	1.61	1.27	1.25	1.22	1.20
	σ^2	1.22	1.05	1.01	1.03	1.03
$T = 2000$	μ	2.19	2.10	2.07	2.05	2.05
	ϕ	2.11	1.73	1.80	1.82	1.69
	σ^2	2.13	2.10	2.16	2.20	1.98
$T = 3000$	μ	2.60	2.49	2.51	2.43	2.50
	ϕ	1.82	1.67	2.05	1.93	1.63
	σ^2	1.67	1.58	1.92	1.77	1.58

TABLE V.3 – Evolution de R du diagnostic de Gelman pour le modèle SCD



(a) (b)

FIGURE V.12 – Evolution du R par rapport à T pour mu et phi.

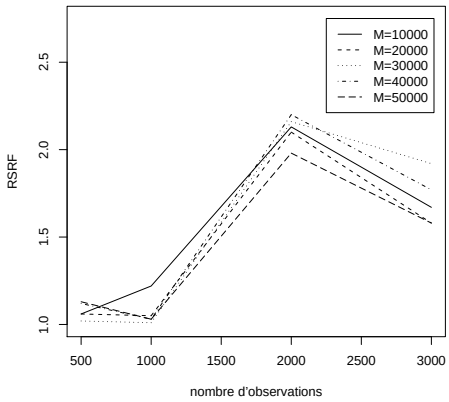
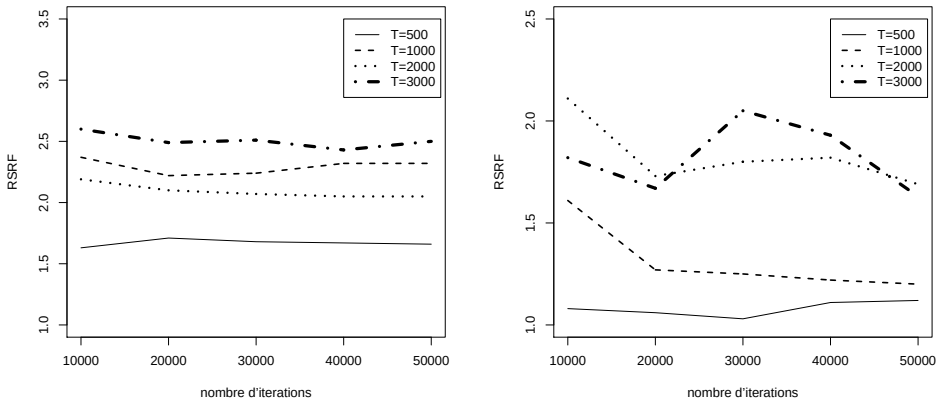


FIGURE V.13 – Evolution du R par rapport à T pour sigma.



(a) (b)

FIGURE V.14 – Evolution du R par rapport à M pour mu et phi.

Ces résultats montrent que R ne varie pas d'une manière importante lorsque M augmente notamment pour μ où il est pratiquement constant pour toutes les valeurs de T, alors qu'il augmente lorsque T augmente et s'éloigne de la valeur 1. Ce diagnostic confirme les conclusions obtenues avec le IF.

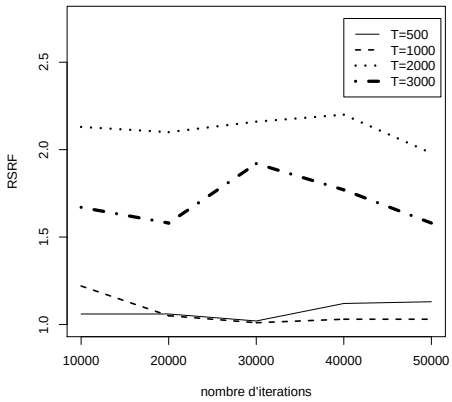


FIGURE V.15 – Evolution du R par rapport à M pour sigma.

Diagnostic de Geweke

Ce diagnostic nous fournit la valeur de la statistique du test qui doit être en valeur absolue inférieure à 1.96 pour atteindre la stationnarité. Les différentes valeurs de la statistique sont mentionnées dans le tableau V.4 pour les différentes valeurs de M et de T.

	M	10^3	2×10^3	3×10^3	4×10^3	5×10^3
$T = 500$	μ	4.82	2.77	1.68	0.29	0.54
	ϕ	-2.75	-0.58	1.58	-0.23	-0.35
	σ^2	-1.39	-3.09	-4.51	-0.34	0.08
$T = 1000$	μ	-2.10	0.01	-0.07	-0.25	-1.22
	ϕ	-5.64	-0.30	0.01	0.37	0.98
	σ^2	8.73	1.11	1.12	0.59	0.02
$T = 2000$	μ	3.40	0.63	0.13	0.68	0.57
	ϕ	2.73	2.28	0.67	1.01	2.37
	σ^2	-2.43	-2.47	-1.12	-1.39	-2.47
$T = 3000$	μ	-2.16	-2.72	-2.19	-1.99	0.06
	ϕ	-5.66	-0.81	-2.38	-2.74	-0.56
	σ^2	7.91	2.01	3.30	2.85	0.51

TABLE V.4 – Valeurs de la statistique de Geweke pour le modèle SCD

D'après le tableau V.4 on constate une amélioration lorsque M augmente c'est à dire que les chaines se stationnarise lorsque M augmente. Mais lorsque

T augmente les conclusions ne sont pas claires.

Diagnostic de Heidelberg et Welch

Le diagnostic de Heidelberg et Welch qui teste la stationnarité de la chaine simulée est constituée de deux tests, donc pour que la chaine soit stationnaire il faut qu'elle passe les deux tests.
Le tableau V.5 donne les résultats de ce diagnostic, où P indique que le test est positif (test passed) et F indique que le test est négatif (test failled).

	M	10^3	2×10^3	3×10^3	4×10^3	5×10^3
$T = 500$	μ	P-P	P-P	P-P	P-P	P-P
	ϕ	P-P	P-P	P-P	P-P	P-P
	σ^2	P-P	P-P	P-P	P-P	P-P
$T = 1000$	μ	P-F	P-F	P-F	P-F	P-F
	ϕ	F-F	P-P	P-P	P-P	P-P
	σ^2	F-F	P-F	P-F	P-F	P-F
$T = 2000$	μ	P-P	P-P	P-P	P-P	P-P
	ϕ	P-P	P-P	P-P	P-P	P-P
	σ^2	P-P	P-P	P-P	P-P	P-P
$T = 3000$	μ	P-P	P-P	P-P	P-P	P-P
	ϕ	P-P	P-P	P-P	P-P	P-P
	σ^2	P-P	P-P	P-P	P-P	P-F

TABLE V.5 – Résultats des deux tests de Heidelberg et Welch pour le modèle SCD

Les résultats du diagnostic de Heidelberg et Welch montrent que pratiquement toutes les chaines sont stationnaires pour tous les paramètres et toutes les valeurs de M et T.

Diagnostic de Raftery et Lewis

Ce diagnostic nous fournit le nombre d'itérations M nécessaires pour satisfaire les conditions spécifiées lors de l'application du diagnostic ainsi qu'un estimateur de I (facteur de dépendance) tel que $I = M/M_{min}$ avec M_{min} est le nombre d'itérations minimal pour effectué le diagnostic. Une valeur de I supérieure à 5 indique des autocorrélations élévées. Le tableau V.6 donne les valeurs de I.

	M	10^3	2×10^3	3×10^3	4×10^3	5×10^3
$T = 500$	μ	2.60	3.03	3.74	3.41	4.92
	ϕ	17.61	26.56	27.16	30.23	31.7
	σ^2	26.1	35.5	35.4	43.1	48.46
$T = 1000$	μ	1.54	2.60	3.50	3.37	4.48
	ϕ	32.79	47.76	45.23	50.23	51.23
	σ^2	27.96	37.6	57.63	64.96	63.93
$T = 2000$	μ	1.11	1.41	1.45	1.85	2.19
	ϕ	12.03	26.06	29.13	29.03	35.16
	σ^2	33.13	51.3	60.53	63.96	79.26
$T = 3000$	μ	1.62	2.17	2.64	3.13	4.83
	ϕ	17.01	23.6	38.0	38.06	90.3
	σ^2	28.8	48.56	74.43	87.8	79.9

TABLE V.6 – Evolution du facteur de dépendance du diagnostic de Raftery et Lewis pour le modèle SCD

On remarque que les valeurs de I (indiquées dans le tableau V.6) sont toutes inférieures à 5 pour μ et supérieures à 5 pour ϕ et σ^2. On remarque aussi que ces valeurs augmentent lorsque T augmente. Donc on peut dire que le nombre d'itérations est suffisant pour atteindre la convergence pour le paramètre μ tandis qu'il ne l'est pas pour ϕ et σ^2 donc il faut exécuter des chaines plus longues.

Conclusion

L'application des différents diagnostics de convergence sur le modèle SCD nous montre que les valeurs des chaines tournent autour des vraies valeurs pour les trois paramètres. On constate aussi que les chaines sont stationnaires dans la majorité des cas (différentes valeurs de M et T) comme le montre le diagnostic de Geweke ainsi que celui de Heibelberg et Welch. Le facteur d'inefficacité ainsi que le diagnostic de Gelman concluent que les résultats restent pratiquement stationnaires lorsque le nombre d'itérations augmente. Tandis que lorsque T augmente ces diagnostics indiquent de fortes autocorrélations entre les valeurs des chaines notamment pour ϕ et σ^2, ces autocorrélations sont beaucoup moins importantes pour μ. La dépendance entre les réalisations des chaines est une preuve de non convergence et pour réduire la dépendance il faut executer des chaines de plus grandes tailles. Ces résultats sont confirmés par le diagnostic de Raftery qui conclut que le nombre d'itérations est suffisant pour μ tandis qu'il faut plus d'itérations pour les paramètres ϕ et σ^2. On remarque aussi que les résultats sont plus mauvais lorsque T augmente. Donc plus T augmente plus il faut d'itérations.

V.3 Le modèle à volatilité stochastique (SV)

Dans le but de voir l'influence du nombre d'observations (T) sur la convergence des trois paramètres (μ, ϕ, σ^2) et pour répondre au problème de savoir comment influe le nombre d'observations disponibles T sur la convergence des chaines générées, nous avons généré trois chaines pour $T = 500, 1000, 2000, 3000$ avec un nombre d'itérations $M = 15000, 25000, 35000, 45000, 55000$ avec un burn in $M_0 = 5000$.
Les valeurs initiales des trois paramètres (μ, ϕ, σ^2) pour chacune des chaines sont respectivement $(0, 0.8, 0.01)$, $(-0.5, 0.85, 0.0225)$ et $(-1, 0.9, 0.09)$. Les vraies valeurs des paramètres sont $(-1.17, 0.96, 0.04)$.

V.3.1 Diagnostics visuels (graphiques)

Pour l'étude de la convergence des chaines vers les vraies valeurs des paramètres on commence par une analyse graphique à savoir la trace et le graphe des autocorrélations afin de voir l'évolution des valeurs de la chaine ainsi que l'évolution des autocorrélations.

D'après la figure (V.16) on remarque que les valeurs des chaines varient autour des vraies valeurs pour les trois paramètres (dont les valeurs sont indiquées en lignes rouges) aussi bien que dans les traces que dans les histogrammes.

Les figures (V.17) et (V.18) indiquent de très importantes autocorrélations pour les paramètres ϕ et σ^2 et beaucoup moins importantes pour μ, .

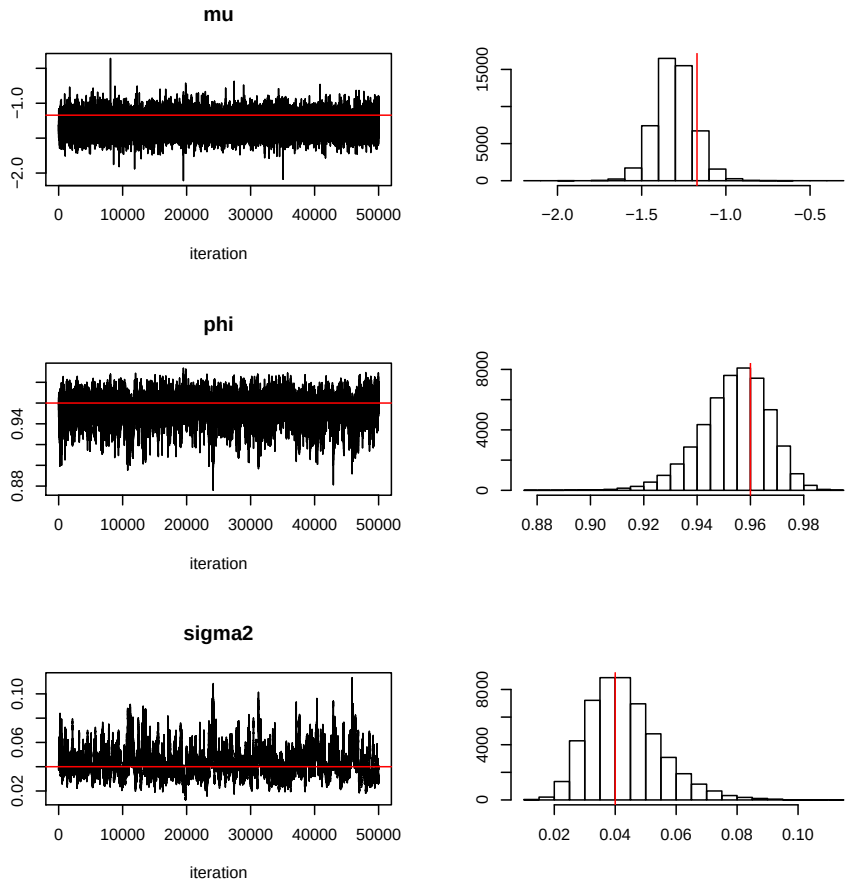


FIGURE V.16 – Traces et histogrammes des trois paramètres.

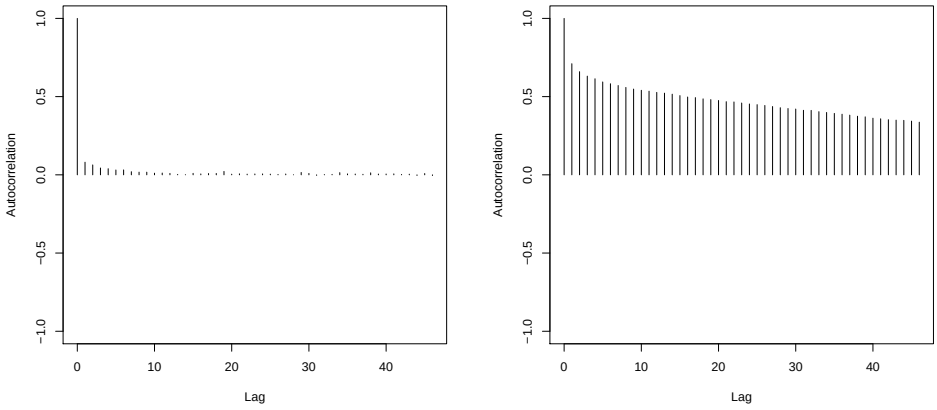
Estimations obtenues

Les estimations portées dans le tableau V.7 sont les moyennes des estimations obtenues par trois chaines.

Tableau des moyennes

D'après le tableau V.7 on remarque qu'en augmentant M les estimations restent stationnaires tandis qu'en faisant varier T on constante une légère amélioration notamment pour μ et ϕ.

Les figures (V.19), (V.20), (V.21) et (V.22) montrent l'évolution des valeurs moyennes par rapport au nombre d'itérations M et par rapport à T.



(a) (b)

FIGURE V.17 – Corrélogrammes de mu et phi.

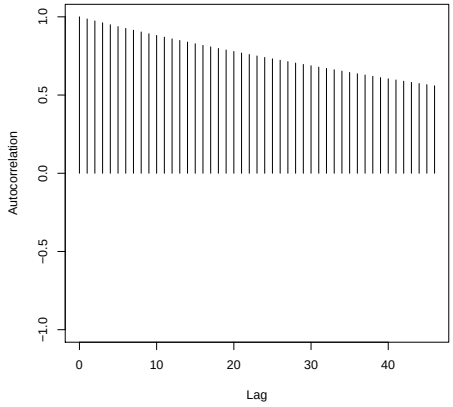


FIGURE V.18 – Corrélogramme de sigma.

D'après ces figures on constate que les valeurs moyennes restent stationnaires par rapport à M tandis qu'on remarque une amélioration des estimateurs en faisant varier T qui est plus importante pour ϕ et σ^2.

	M	10^3	2×10^3	3×10^3	4×10^3	5×10^3
$T = 500$	μ	-0.832	-0.915	-0.915	-0.916	-0.914
	ϕ	0.961	0.960	0.954	0.960	0.961
	σ^2	0.033	0.032	0.032	0.032	0.032
$T = 1000$	μ	-1.206	-1.205	-1.206	-1.206	-1.206
	ϕ	0.945	0.947	0.947	0.947	0.947
	σ^2	0.041	0.039	0.039	0.039	0.039
$T = 2000$	μ	-1.277	-1.278	-1.278	-1.278	-1.278
	ϕ	0.950	0.950	0.950	0.950	0.950
	σ^2	0.049	0.049	0.049	0.049	0.049
$T = 3000$	μ	-1.133	-1.133	-1.133	-1.133	-1.133
	ϕ	0.964	0.964	0.964	0.964	0.964
	σ^2	0.034	0.035	0.035	0.035	0.035

TABLE V.7 – Tableau des moyennes pour le modèle SV

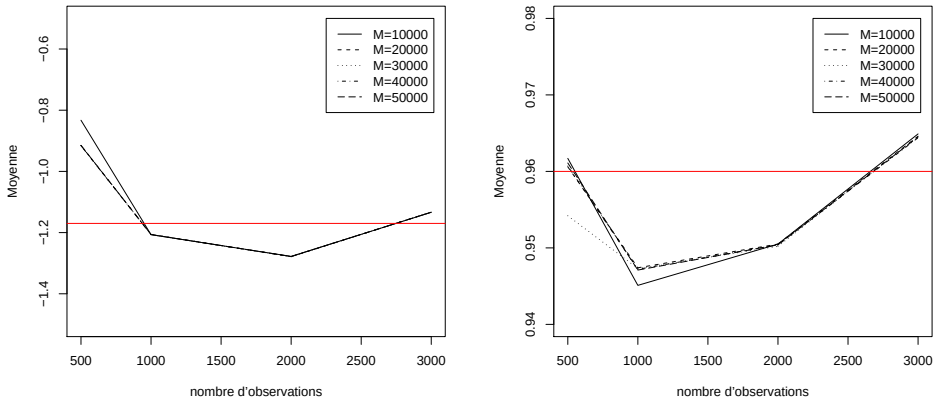


FIGURE V.19 – Evolution de la moyenne par rapport à T pour mu et phi.

V.3.2 Diagnostics de convergence

Facteur d'inefficacité

Le tableau V.8 donne les valeurs moyennes du IF des trois chaines générées pour chacun des paramètres

Les figures (V.23), (V.24), (V.25) et (V.26) montrent l'évoltion du IF par rapport à T et par rapport à M.

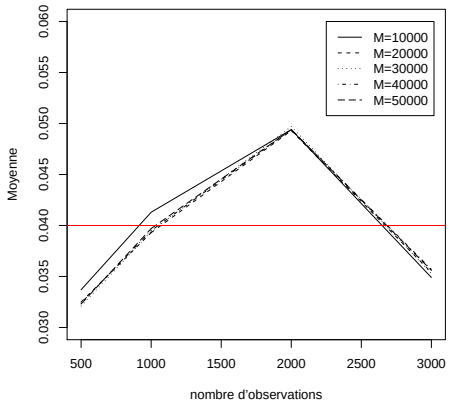
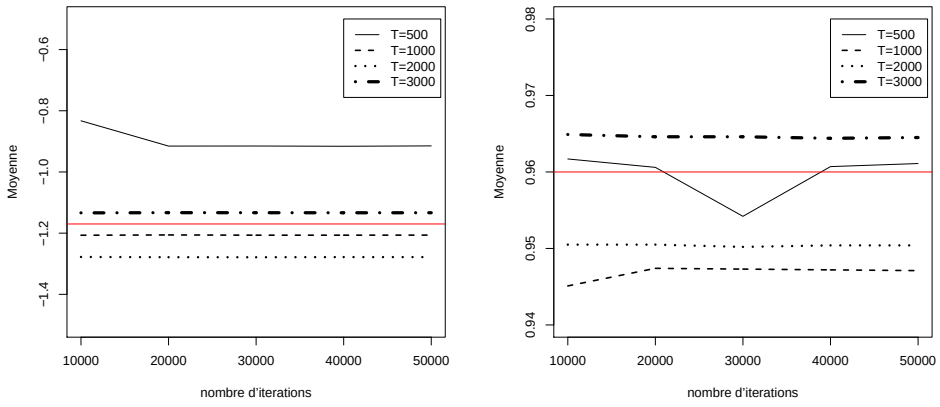


FIGURE V.20 – Evolution de la moyenne par rapport à T pour sigma.



(a) (b)

FIGURE V.21 – Evolution de la moyenne par rapport à M pour mu et phi.

D'après ces figures on remarque que le IF reste stationnaire quand M augmente alors qu'il augmente lorsque T augmente très faiblement pour μ et de manière plus importante pour ϕ et σ^2. Donc plus T augmente plus les valeurs des chaines deviennent de plus en plus dépendantes.

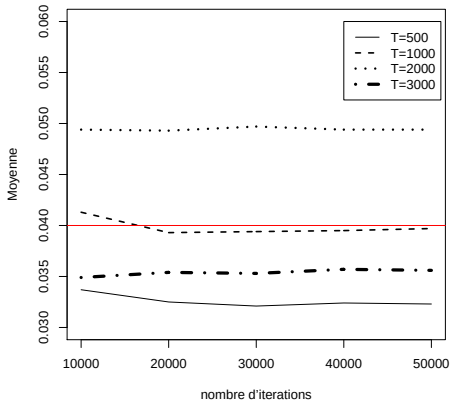


FIGURE V.22 – Evolution de la moyenne par rapport à M pour sigma.

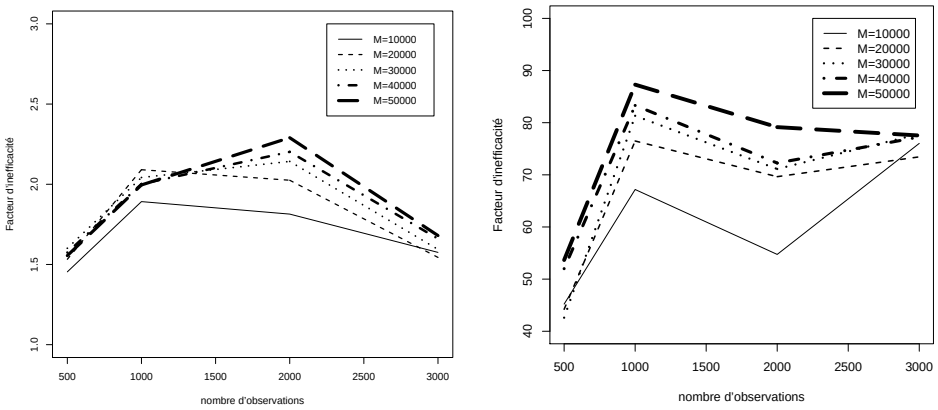
	M	10^3	2×10^3	3×10^3	4×10^3	5×10^3
$T = 500$	μ	1.45	1.53	1.60	1.57	1.55
	ϕ	45.20	44.20	42.60	52.0	53.66
	σ^2	97.32	111.3	114	119.7	122.2
$T = 1000$	μ	1.89	2.09	2.04	2.00	1.99
	ϕ	67.18	76.48	81.32	83.30	87.29
	σ^2	152.1	149.5	153.4	153.6	158.1
$T = 2000$	μ	1.81	2.02	2.14	2.20	2.29
	ϕ	54.73	69.64	71.12	72.26	79.14
	σ^2	126.07	135.7	133.4	138.7	143.9
$T = 3000$	μ	1.57	1.54	1.59	1.66	1.68
	ϕ	76.02	73.44	77.84	77.24	77.54
	σ^2	150.9	143.8	145.3	144.8	145.9

TABLE V.8 – Evolution du facteur d'inefficacité pour le modèle SV

Diagnostic de Gelman

Le tableau (V.9) donne les valeurs de R du diagnostic de Gelman. Quand R dépasse 1, il faut executer une chaine de plus grande taille pour atteindre la convergence.

Les figures (V.27) et (V.28) représentent l'évolution de R par rapport à T, chaque courbe représente la variation de R pour une valeur donnée de M qui représente le nombre d'itérations. Les figures (V.29) et (V.30) représentent l'évolution de R par rapport à M, chaque courbe représente l'évolution de ce



(a) (b)

FIGURE V.23 – Evolution du facteur d'inefficacité par rapport à T pour mu et phi.

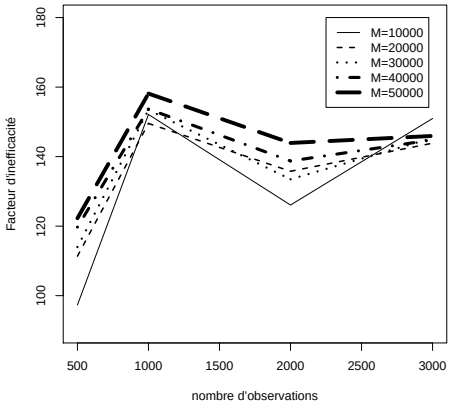
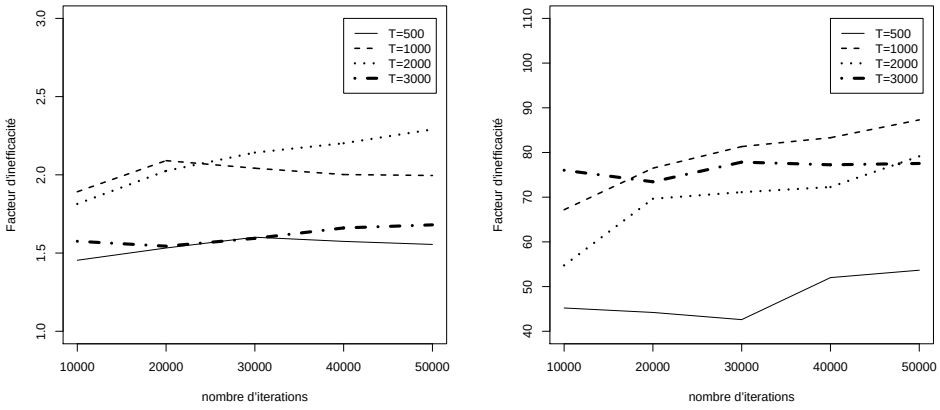


FIGURE V.24 – Evolution du facteur d'inefficacité par rapport à T pour sigma.

dernier pour une valeur donnée de T.
D'après les deux premières figures on remarque que l'évolution de R a le même comportement pour les trois paramères et pour n'importe quelle valeur de M, donc pas un changement important.
Les figures (V.29) et (V.30) montrent une augmentation des valeurs de R



(a) (b)

FIGURE V.25 – Evolution du facteur d'inefficacité par rapport à M pour mu et phi.

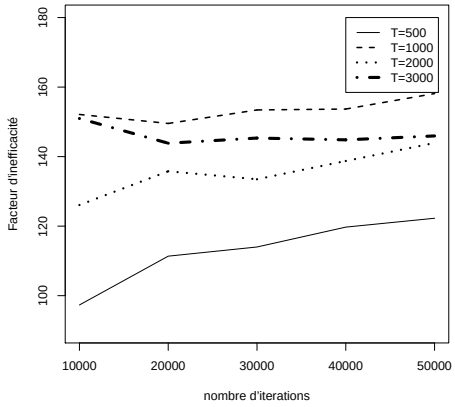
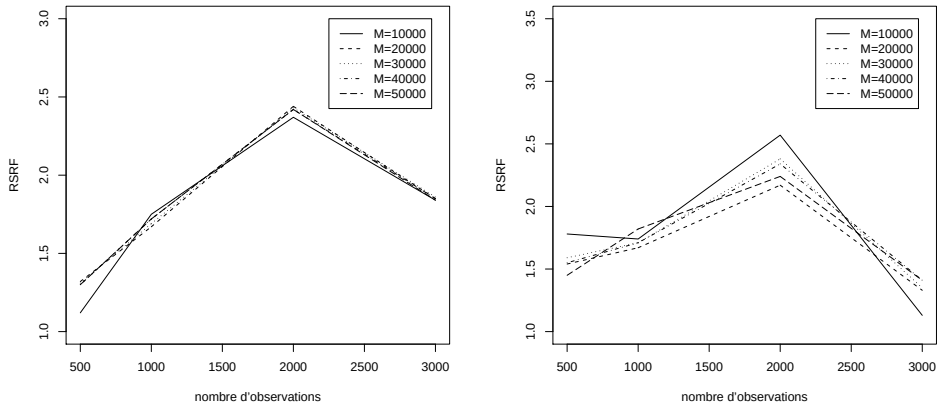


FIGURE V.26 – Evolution du facteur d'inefficacité par rapport à M pour sigma.

quand T augmente alors qu'il reste pratiquement constant quand le M augmente. Une valeur de R superieure à 1 indique que la variance de la distribution stationnaire et la variance de la chaine simulée ne sont pas (.....) pour réduire l'écart entre ces deux variables il faut simuler une chaine plus longue.

	M	10^3	2×10^3	3×10^3	4×10^3	5×10^3
$T = 500$	μ	1.12	1.32	1.31	1.30	1.30
	ϕ	1.78	1.54	1.59	1.55	1.45
	σ^2	1.05	1.11	1.16	1.14	1.10
$T = 1000$	μ	1.75	1.67	1.69	1.72	1.72
	ϕ	1.74	1.67	1.71	1.71	1.82
	σ^2	1.23	1.16	1.18	1.17	1.26
$T = 2000$	μ	2.37	2.44	2.43	2.42	2.42
	ϕ	2.57	2.17	2.38	2.34	2.24
	σ^2	2.90	2.37	2.69	2.63	2.46
$T = 3000$	μ	1.84	1.85	1.86	1.85	1.84
	ϕ	1.13	1.33	1.35	1.41	1.41
	σ^2	1.51	1.88	1.89	2.09	2.07

TABLE V.9 – Evolution de R du diagnostic de Gelman pour le modèle SV



(a) (b)

FIGURE V.27 – Evolution du R par rapport à T pour mu et phi.

Diagnostic de Geweke

Le tableau V.10 comporte les valeurs de la statistique du test de Geweke pour toutes les valeurs de M et de T.

D'après ce tableau on remarque que pratiquement toutes les valeurs de la statistique tu test sont inférieures à 1.96 en valeurs absolues pour les différentes valeurs de M et de T donc on peut conclure que les chaines des

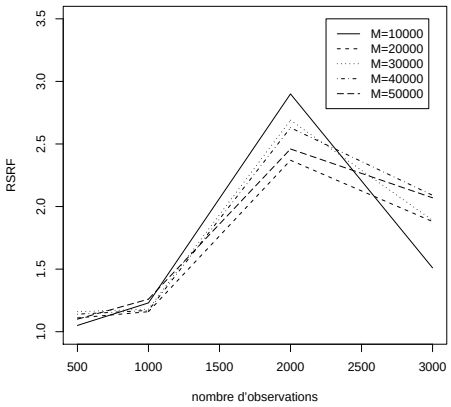
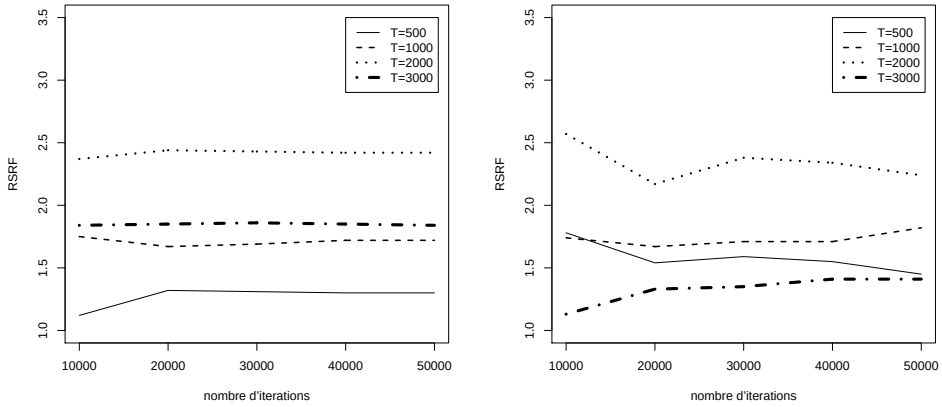


FIGURE V.28 – Evolution du R par rapport à T pour sigma.



(a) (b)

FIGURE V.29 – Evolution du R par rapport à M pour mu et phi.

trois paramètres sont stationnaires dans la majorité des cas.

Diagnostic de Rafetery and Lewis

Le tableau V.11 donne l'évolution du facteur de dépendance (I) du diagnostic de Raftery et Lewis par rapport au nombre d'itérations et T.

D'après les valeurs de I donnés dans le tableau V.11, on remarque que les

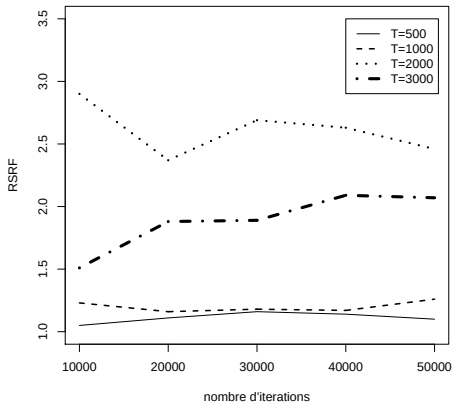


FIGURE V.30 – Evolution du R par rapport à M pour sigma.

	M	10^3	2×10^3	3×10^3	4×10^3	5×10^3
$T = 500$	μ	0.81	0.55	-0.81	-0.92	-0.22
	ϕ	-3.77	-0.55	-1.77	-2.85	-2.12
	σ^2	-3.85	0.72	1.91	2.65	2.01
$T = 1000$	μ	0.37	0.32	0.57	-0.11	-0.17
	ϕ	-0.13	-0.13	2.15	1.84	1.26
	σ^2	-0.09	0.10	-2.23	-1.66	-1.44
$T = 2000$	μ	-1.28	-0.47	-1.15	-1.12	-0.88
	ϕ	-2.82	-1.25	-1.16	-0.46	-0.30
	σ^2	2.81	1.20	1.24	0.41	0.25
$T = 3000$	μ	0.50	0.54	0.72	1.62	1.88
	ϕ	-1.56	0.54	-0.02	0.78	1.38
	σ^2	1.02	-0.69	-0.05	-0.90	-1.56

TABLE V.10 – Valeurs de la statistique de Geweke pour le modèle SV

valeurs de I sont toutes inférieures à 5 pour le paramètre μ tandis qu'elle sont toutes supérieures à 5 pour les paramètres ϕ et σ^2. Ces valeurs sont stationnaires par rapport à M et augmentent par rapport à T donc on peut conclure que pour les valeurs de la chaine représentant le paramètre μ ne sont pas très corrélées donc le nombre d'itérations est suffisant contrairement à celles de ϕ et σ^2 où elles restent encore corrélées notamment lorsque T augmente ce qui signifie qu'il faut executer des chaines de tailles plus importantes pour essayer d'atteindre la distribution cible.

	M	10^3	2×10^3	3×10^3	4×10^3	5×10^3
$T = 500$	μ	2.04	2.18	2.13	2.23	2.27
	ϕ	6.42	10.7	11.5	14.3	12.9
	σ^2	12.3	23.7	35.0	34.0	35.5
$T = 1000$	μ	1.22	1.18	1.17	1.18	1.16
	ϕ	8.97	15.2	12.6	13.7	14.9
	σ^2	25.6	39.6	35.1	44.0	41.7
$T = 2000$	μ	1.11	1.08	1.05	1.04	1.05
	ϕ	13.2	11.9	12.5	16.8	19.5
	σ^2	31.4	53.2	48.2	45.5	41.6
$T = 3000$	μ	1.04	1.04	1.03	1.04	1.04
	ϕ	20.1	19.7	20.6	23.9	22.3
	σ^2	30.5	41.2	50.8	51.7	44.9

TABLE V.11 – Evolution du facteur de dépendance du diagnostic de Raftery et Lewis pour le modèle SV

Diagnostic de Heidelberg and Welch

Le tableau V.12 porte les résultats des deux tests du diagnostic de Heidelberg et Welch.

	M	10^3	2×10^3	3×10^3	4×10^3	5×10^3
$T = 500$	μ	P-P	P-P	P-P	P-P	P-P
	ϕ	P-P	P-P	P-P	P-P	P-P
	σ^2	P-P	P-P	P-P	P-P	P-P
$T = 1000$	μ	P-P	P-P	P-P	P-P	P-P
	ϕ	P-P	P-P	P-P	P-P	P-P
	σ^2	P-P	P-P	P-P	P-P	P-P
$T = 2000$	μ	P-P	P-P	P-P	P-P	P-P
	ϕ	P-P	P-P	P-P	P-P	P-P
	σ^2	P-P	P-P	P-P	P-P	P-P
$T = 3000$	μ	P-P	P-P	P-P	P-P	P-P
	ϕ	P-P	P-P	P-P	P-P	P-P
	σ^2	P-P	P-P	P-P	P-P	P-P

TABLE V.12 – Résultats des deux tests de Heidelberg et Welch pour le modèle SV

D'après le tableau V.12 on remarque que les deux tests d'Heidelberg sont positifs donc les chaines générées sont stationnaires.

Conclusion
Après l'application des différents diagnostics de convergence sur les chaines représentants les différents paramètres du modèle SV dans le but de comparer l'évolution de la convergence en variant T et M, on constate que les chaines sont stationnaires pour les trois paramètres comme le montre le diagnostic de Geweke ainsi que celui de Heidelberg et Welch. On constate aussi que les valeurs des chaines restent dépendantes notamment pour les paramètres ϕ et σ^2 cette dépendance et beaucoup moins importante pour μ, on remarque aussi que cette dépendance devient plus importante lorsque T augmente (diagnostic de Gelman ainsi que le facteur d'inefficacité) cette conclusion est confirmée par le diagnostic de Raftery et Lewis qui montre qu'il faut plus d'itérations pour réduire la dépendance pour les paramètres ϕ et σ^2 tandis qu'il est suffisant pour le paramètre μ. Donc plus T augmente plus il faut d'itérations afin de réduire les dépendances et pouvoir atteindre la distribution cible.

V.4 Conclusion

D'après les résultats numériques obtenus, on remarque que les chaines générées montrent une bonne convergence pour μ mais pour ϕ et surtout σ^2 cette convergence est moins bonne. Les résultats sont assez identiques pour les deux modèles bien que légèrement meilleurs pour le modèle SV.
Les tests de convergence visuels montrent que les chaines générées tournent autour des vraies valeurs des paramètres bien que les autocorrélations sont élevées pour les paramètres ϕ et σ^2 et moins élevées pour μ. Alors que les diagnostics de Heidelberg et Geweke montrent que les chaines générées sont stationnaires pour tous les paramètres. Les valeurs du facteur d'inefficacité montrent que les autocorrélations restent importantes notamment pour ϕ et plus importantes pour σ^2, ces conclusions sont confirmées par les diagnostics de Gelman et Raftery qui montrent qu'il faut plus d'itérations pour atteindre la convergence.
On peut déduire également à partir de ces résultats que les réalisations des différntes chaines générées restent très corrélées en faisant augmenter T pour un même nombre d'itérations M comme le montre l'évolution du IF et du R de Gelman, legèrement pour μ et d'une manière plus importante pour ϕ et σ^2. Donc la convergence se déteriore lorsque T augmente.
Ceci est dû à la taille du vecteur d'états latents à estimer (de taille T) qui augmente lorsque T augmente. Donc le gain obtenu par la quantité d'information disponible au départ est perdu dans l'estimation des états inobservés car les deux vecteurs sont de même taille (T).
Donc les résultats obtenus montrent que lorsque T augmente il faut un plus

grand nombre d'itérations pour atteindre la distribution cible.

Conclusion

Dans ce mémoire, nous avons utilisé les méthodes MCMC pour l'inférence bayésienne dans les modèles à espace d'état non gaussiens en particuliers les deux modèles SCD et SV à temps discret ainsi que pour les modèles de diffusion qui sont des modèles à temps continu en prenant l'exemple du modèle SV à temps continu. L'inférence statistique bayésienne utilisée consiste en l'inférence sur la distribution conjointe des paramètres et des variables latentes conditionnellement aux données observées, et les méthodes MCMC fournissent un outil pour explorer ces distributions complexes de dimensions élevées. Ces méthodes sont particulièrement bien adaptées pour calculer des intégrales de grandes dimensions associées aux calculs des caractéristiques des densités a posteriori dans les applications impliquant des observations manquantes ou latentes.
Nous avons d'abord décrit les différents algorithmes et mécanismes associés à ces méthodes. Nous avons ensuite discuté la façon dont les diagnostics habituels de convergence peuvent être utilisées pour fournir une certaine assurance que l'inférence basée sur un échantillon donné est fiable. Mais généralement ces différentes méthodes donnent des conclusions contradictoires ou pas claires. Dans notre cas, ces diagnostics ont été utilisés pour controler la convergence des différentes densités a posteriori des paramètres vers leurs distributions cibles. Nous avons étudié cette convergence en faisant augmenter le nombre d'observations (quantité d'information) disponibles et le nombre d'itérations utilisé dans l'échantillonneur MCMC.
Les résultats numériques obtenus, ont montré que les chaines générées montrent une bonne convergence pour μ mais pour ϕ et surtout σ^2 cette convergence est moins bonne. Les résultats sont assez identiques pour les deux modèles bien que légèrement meilleurs pour le modèle SV.
Les diagnostics utilisés ont montré que les chaines obtenues sont stationnaires (diagnostics de Geweke et Heidelberg) mais que les réalisations restent très corréllées (Facteur IF et diagnostic de Gelman).
En ce qui concerne le comportement de la convergence avec l'augmentation de T, les résultats obtenus ont conclu que pour un même nombre d'itérations M

les résultats, pour les trois paramètres des modèles (μ, ϕ, σ^2) se déteriorent avec T legèrement pour μ et d'une manière plus importante pour ϕ et σ^2 comme le montre l'évolution du IF et du R de Gelman.
Cela peut s'expliquer par le fait que le gain obtenu par l'augmentation de l'information disponible au départ (nombre d'observations plus grand) est perdu dans l'estimation du vecteur des variables latentes (de longueur T) et donc plus on a d'observations plus on a d'états latents à estimer.
Donc les résultats obtenus montrent que lorsque T augmente il faut un plus grand nombre d'itérations pour atteindre la distribution cible surtout pour ϕ et σ^2. Une des solutions pour accélérer la convergence peut être la reparamétrisation des modèles. Certaines reparamétrisations ont été étudiées par Strickland et al (2008) où il est montré qu'elles conduisent à des gains d'efficacité du point de vue de la convergence par rapport au nombre d'itérations M. Mais le comportement de ces méthodes par rapport à T n'a pas été abordé.

Annexe A

Lois de probabilité usuelles en statistique bayésienne

Loi Gamma

On dit que la variable aléatoire X prenant ses valeurs dans R_+ est de loi Gamma avec les paramètres $a > 0$ et $b > 0$, si elle admet la densité

$$f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx)$$

- On note $X \sim G(a, b)$
- $\Gamma(a) = \int_0^{+\infty} e^{-x} x^{a-1} dx$
- $\Gamma(a+1) = a\Gamma(a)$
- $E(X) = ab^{-1}$
- $V(X) = ab^{-2}$
- Si $X \sim G(1, b)$ alors $X \sim Exp(b)$

Loi Beta sur l'intervalle $[0, 1]$

On dit que la variable aléatoire X prenant ses valeurs dans l'intervalle $[0, 1]$ est de loi Beta avec les paramètres $a > 0$ et $b > 0$ si elle admet la densité

$$\begin{aligned} f(x) &= \frac{1}{\beta(a,b)} x^{a-1}(1-x)^{b-1} 1_{[0,1]} \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} 1_{[0,1]} \end{aligned}$$

- $E(X) = a/(a+b)$
- $V(X) = ab/[(a+b)^2(a+b+1)]$
- Si $a, b > 1$, $f(x)$ a un mode unique en $(a-1)/(a+b-2)$
- Si $a = b = 1$, $X \sim U_{[0,1]}$

Loi Inverse Gamma

Une variable aléatoire X de valeurs dans $]0, \infty[$ est de loi inverse Gamma de paramètres $a > 0$ et $b > 0$, si admet la densité

$$f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(\frac{-b}{x})$$

- On note $X \sim IG(a, b)$
- $E(X) = \dfrac{b}{a-1}$ si $a > 1$.
- $V(X) = \dfrac{b^2}{(a-1)^2(a-2)}$ si $a > 2$.
- Si $X \sim IG(a, b)$ alors $1/X \sim G(a, b)$.
- L'utilisation typique de la densité Inverse Gamma dans l'analyse bayésienne est d'être une densité a priori conjuguée pour la variance d'un modèle normal i.e $\sigma^2 \sim IG(\nu_0/2, \nu_0\sigma_0^2/2)$. Une densité a priori impropre $p(\sigma^2) \propto 1/\sigma^2$ résulte quand $\nu_0 = 0$

Loi Normale Gamma

En analyse bayésienne d'un modèle normal on s'interesse généralement aux paramètres m qui désigne la moyenne et $\tau = 1/\sigma^2$ dit précision.
Supposons que

$$m/\tau, \mu, \lambda \sim N(\mu, 1/(\lambda\tau))$$

a une distribution normale avec une moyenne μ et une variance $1/(\lambda\tau)$ où

$$\tau/a, b \sim G(a, b)$$

de densité Gamma. Alors (m, τ) est de distribution normale Gamma, notée $(m, \tau) \sim NG(\mu, \lambda, a, b)$ de densité :

$$f(m, \tau/\mu, \lambda, a, b) = \frac{b^a\sqrt{\lambda}}{\Gamma(a)\sqrt{2\pi}} \tau^{a-\frac{1}{2}} \exp(-b\tau) \exp\left(-\frac{\lambda\tau(m-\mu)^2}{2}\right)$$

Donc la loi Normale Gamma est la conjuguée de la loi $N(m, 1/\tau)$

Loi Normale Inverse Gamma

Supposons que

$$m/\sigma^2, \mu, \lambda \sim N(\mu, \sigma^2/\lambda)$$

est de distribution normale avec une moyenne μ et une variance σ^2/λ, où

$$\sigma^2/a, b \sim IG(a, b)$$

est de distribution inverse Gamma.
Alors, (m, σ^2) est de distribution normale inverse Gamma, notée $(m, \sigma^2) \sim NIG(\mu, \lambda, a, b)$ de densité :

$$f(m, \sigma^2\ \mu, \lambda, a, b) = \frac{\sqrt{\lambda}}{\sigma\sqrt{2\pi}} \frac{b^a}{\Gamma(a)} \left(\frac{1}{\sigma^2}\right)^{a+1} \exp\left(-\frac{2b + \lambda(m-\mu)^2}{2\sigma^2}\right)$$

Donc la loi Normale Inverse Gamma est la conjuguée de la loi $N(m, \sigma^2)$

Annexe B

Algorithme de Metropolis Hastings acceptation rejet

L'algorithme de Metropolis Hastings acceptation rejet est une algorithme hybride entre la méthode d'acceptation rejet et l'algorithme de Metropolis Hastings.
Supposons que notre but est de simuler θ de densité a posteriori $p(\theta)$. On doit choisir une densité instrumentale notée $q(\theta)$ qui est indépendante de l'état actuel de la chaine. Pour construire un échantillon à partir de la distribution cible p en utilisant l'algorithme de Metropolis Hastings acceptation rejet on procède comme suit :

Algorithme

1. Initialiser θ^0, soit le premier élément de la chaine
2. Simuler une valeur candidate $\theta^* \sim q(.)$ indépendante de l'état actuel de la chaine de Markov.
3. Accepter θ^* avec probabilité $\alpha' = min\left[1, \frac{p(\theta^*)}{Cq(\theta^*)}\right]$, si θ^* est rejetée retourner à l'étape 2.
4. Calculer :

$$\alpha = \begin{cases} 1 & \text{si} \quad p(\theta^{(t-1)}) < cq(\theta^{(t-1)}) \\ \dfrac{p(\theta^{(t-1)})}{cq(\theta^{(t-1)})} & \text{si} \quad p(\theta^{(t-1)}) > cq(\theta^{(t-1)}) \\ & \text{et} \quad p(\theta^*) < cq(\theta^*) \\ \min\left[\dfrac{p(\theta^*)q(\theta^{(t-1)})}{p(\theta^{(t-1)})q(\theta^*)}, 1\right] & \text{si} \quad p(\theta^{(t-1)}) > cq(\theta^{(t-1)}) \\ & \text{et} \quad p(\theta^*) > cq(\theta^*) \end{cases}$$

5. Accepter θ^* avec probabilité α tel que

$$\theta^{(t)} = \begin{cases} \theta^* & \text{si avec probabilité } \alpha \\ \theta^{(t-1)} & \text{sinon} \end{cases}$$

Bibliographie

[1] Bauwens. L, et Verdas. D (1999). The stochastic conditional duration model : a latent variable model for analysis of financial durations. *Journal Of Economics,* **19**, 381-412.

[2] Brooks. S. P (1998). Markov chain Monte Carlo method and its application. *The Statistician,* **47**, 69-100.

[3] Carlin. B. P, Polson. N. G, Stoffer. D. S (1992). A Monte Carlo approach to nonnormal and nonlinear state space modeling. *Journal Of The American Statistical Association,* **87**, 493-500.

[4] Chib. S, Nardari, Shephard. N (1998). Markov chain Monte Carlo methods for generalised stochastic volatility models. *Journal Of Economics,* **108** (2), 281-316.

[5] Cowles. M. K et Carlin. B. P (2006). Markov chain Monte Carlo convergence diagnostics : a comparative review. *Journal Of The American Statistical Association,* **91**, 883-904.

[6] Durbin. J et Koopman. S. J (1998). Time series of non gaussian observations based on state space models from both classical and bayesian perspectives. *The Royal Statistical Society,* 62(1), 3-56.

[7] Durbin. J et Koopman. S. J (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrica,* **89**, 603-616.

[8] Eraker. B (2001). MCMC analysis of diffusion models with application to finance. *Journal Of Business Economic Statistics,* **19**, 177-191.

[9] Gelman. A et Rubin. D. B (1992). Inference from iterative simulation using multiple sequences. *Statistical Science,* **7**, 457-511.

[10] Hoff. P. D (2009). *A first course in Bayesian statistical methods.* Springer.

[11] Jackman. S (2009). *Bayesian analysis for the social sciences.* Wiley.

[12] Jacquier. E, Polson. N. G et Rossi. P. E (1994). Bayesian analysis of stochastic volatility models. *Journal Of Economic Statistics,* **20**, 69-87.

[13] Kim. S, Shephard. N et Chib. S (1998). Stochastic volatility : Likelihood inference and comparaison with ARCH models. *Review Of Economic Studies,* **65**, 361-393.

[14] Petris. G, Petrone. S et Campagnoli. P (2009). *Dynamic linear models with R.* Springer.

[15] Rachev. S. T, Hsu. J. S. J, Bagasheva. B. S et Fabozzi. F. J (2008). *Bayesian methods in finance.* Wiley.

[16] Robert. C (1990). *L'analyse statistique bayésienne.* Economica.

[17] Robert. C. P et Casella. G (2010). *Introducing Monte Carlo methods with R.* Springer.

[18] Robert. C. P et Casella. G (2011). *Méthodes de Monte Carlo avec R.* Springer.

[19] Strickland. C. M, Forbes. C. S et Martin. G. M (2005). Bayesian analysis of the stochastic conditional duration model. *Computational Statistics Data Analysis,* **50**(9), 2247-2267.

[20] Strickland. C. M, Martin. G. M et Forbes. C. S (2008). Parametrisation and efficient MCMC estimation of non-gaussian state space models. *Computational Statistics Data Analysis,* **52** (6), 2911-2930.

[21] Strickland. C. M, Turner. I. W et Denham. R (2008). Efficient bayesian estimation of multivariate state space models. *Computational Statistics Data Analysis,* **53**, 4116-4125.