Mélange de distributions comme fonction d'importance dans l'échantillonnage préférentiel combiné avec l'algorithme de Monte Carlo par Chaîne de Markov

Dorota Gajda*, Chantal Guihenneuc-Jouyaux**
Judith Rousseau***

*Biostatistiques, CESP Centre de recherche en Epidémiologie et Santé des Populations,
U1018, Inserm, F-94807, Villejuif, France
Université Paris Sud, UMRS1018, Villejuif, F-94807, France
dorota.gajda@inserm.fr,
** EA 4064 (épidémiologie environnementale : impact sanitaire des pollutions),
Faculté des Sciences Pharmaceutiques et Biologiques, Université Paris Descartes,
4, av de l'Observatoire, 75006 Paris, France
***Université Paris Dauphine, Paris, France

Résumé. Les algorithmes de Monte Carlo par Chaîne de Markov (MCMC) sont très souvent utilisés pour estimer les lois a posteriori ainsi que leurs moments dans le cadre d'un modèle Bayésien. En effet, selon les modèles, les lois a posteriori ou leurs moments peuvent ne pas avoir d'expression analytique et le recours a des méthodes d'approximation est donc indispensable. Lors de l'étude empirique d'estimateurs, différents jeux donnés sont simulés sous le même modèle (et avec les mêmes valeurs de paramètres) et pour chaque jeu de données, les estimations a posteriori des paramètres sont obtenus via MCMC. Cette procédure est réitérée pour d'autres valeurs des paramètres. Globalement, les temps de calcul peuvent être très importants. L'échantillonnage préférentiel (Importance Sampling en anglais, IS) combiné avec les algorithmes MCMC est une solution permettant de réduire ce temps de calcul. En effet, l'IS nécessite le choix d'une fonction d'importance que nous proposons construite comme mélange de lois a posteriori présélectionnées sur quelques jeux données simulés, lois a posteriori déjà estimées via MCMC. Les autres calculs ne nécessitent plus le recours aux algorithmes MCMC. Les approches évoquées ici sont illustrées sur deux exemples de modèles de Poisson.

1 Introduction

L'échantillonnage préférentiel (Importance Sampling en anglais, IS) est présenté ici comme une méthode d'optimisation algorithmique dans le cas de l'étude empirique d'un estimateur. En effet, même s'il est possible d'avoir des propriétés asymptotiques des estimateurs, les études empiriques sont nécessaires afin d'évaluer leurs comportements dans un cadre non asymptotique. La démarche consiste alors à définir ces situations caractéristiques (taille d'échantillon, valeurs

des paramètres) et pour chacune d'entre elles, à simuler plusieurs jeux de données sur lesquels, les paramètres sont estimés. Cette démarche permet de caractériser les performances des estimateurs selon différentes situations en contrôlant les fluctuations aléatoires. Notre travail se place dans le cadre d'une modélisation paramétrique bayésienne où de manière générique, les données sont notées X et les paramètres θ. Dans le contexte Bayésien, des lois a priori $\pi(\theta)$ sont spécifiées sur les paramètres et sont supposées les mêmes dans toutes les situations étudiées. La démarche bayésienne, comme abordée dans la vaste littérature (Cf. par exemple Robert (2007)), consiste à combiner l'information a priori des paramètres représentée par les lois a priori avec l'information provenant des données à travers la vraisemblance $\pi(X|\theta)$ pour obtenir la loi a posteriori des paramètres conditionnelle aux données, $\pi(\theta|X)$. Cette loi, d'après le Théorème de Bayes, s'écrit comme

$$\pi(\theta|X) = \frac{\pi(X|\theta)\pi(\theta)}{\pi(X)} \tag{1}$$

où $\pi(X)$ est la loi marginale de X. Quand la loi a posteriori ou quand les moments de cette loi ne sont pas explicites, une approximation est obtenue par les algorithmes stochastiques dits de Monte Carlo par Chaînes de Markov (MCMC) comme présentés par Hastings (1970) ou par Geman et Geman (1984). Ces algorithmes itératifs permettent d'obtenir des réalisations Markoviennes sous la loi a posteriori recherchée et, via la théorie ergodique, d'obtenir ainsi des estimations de ses moments. Dans le cas d'études empiriques, l'algorithme itératif MCMC doit être utilisé pour chaque jeu de données simulé. L'utilisation répétée de ces algorithmes peut être très coûteuse en temps de calcul. Pour réduire ce temps de calcul, nous proposons une alternative consistant en l'utilisation combinée d'échantillonnage préférentiel et de MCMC. L'idée d'utilisation simultanée de l'IS a été déjà proposée entre autre par Geyer et Thompson (1992) pour l'évaluation de vraisemblances, par Gelfand et al. (1992) pour des critères de validation croisée mais pas dans le contexte d'études empiriques d'estimateurs. La partie 2 présente le principe de l'échantillonnage préférentiel combiné avec MCMC dans ce contexte ainsi que les propriétés des estimateurs obtenus. Le choix de la fonction d'importance est également discuté. Lors d'une première étude (Gajda et al. (2010)), nous avons proposé certains choix pour cette fonction amenant à soit des réductions importantes de temps de calcul mais avec parfois de mauvaises estimations, soit de très bonnes estimations mais avec une réduction moins nette du temps de calcul. Afin d'améliorer cette procédure, nous proposons ici une fonction d'importance comme mélange de distributions. Enfin, dans la partie 3, ces approches sont illustrées dans le cadre de modèles poissonniens.

2 Méthodes

2.1 Généralités

Soit le modèle paramétrique M_θ où θ est le vecteur des paramètres. Pour une situation particulière en $\theta = \theta_0$, K jeux de données, $(X^{(1)}, \ldots, X^{(K)})$ sont simulés sous le modèle M_{θ_0}. Pour un jeu de données i ainsi obtenu, l'algorithme MCMC permet d'obtenir la réalisation d'une chaîne de Markov $\theta_1, \cdots, \theta_N$ admettant comme loi stationnaire la loi a posteriori $\pi(\theta|X^{(i)})$. Par le théorème ergodique, la moyenne empirique $\frac{1}{N}\sum_{j=1}^{N} g(\theta_j)$ converge presque sûrement vers $E_{\theta|X^{(i)}}(g(\theta))$, espérance de $g(\theta)$ par rapport à la loi a posteriori, et

ceci pour toute fonction intégrable g. Le choix de la fonction g permet ainsi d'obtenir des estimations des différents moments de la distribution a posteriori. Malgré son efficacité incontestable dans les problèmes où les vraies lois ou les vraies valeurs des espérances a posteriori ne sont pas accessibles, l'utilisation des algorithmes MCMC dans le cadre bayésien d'une étude de simulations peut devenir extrêmement longue, car doit être répétée autant de fois que de nombre de jeux de données simulés. Pour $K = 100$ jeux de données, la démarche classique pour estimer $(E_{\pi(\theta|X^i)}(g(\theta)), i = 1, \cdots, 100)$ est illustrée sur la figure 1.

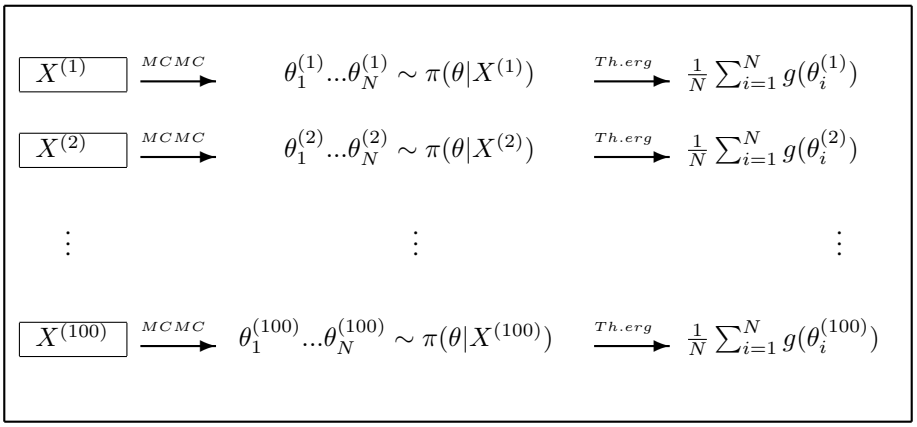

FIG. 1 – *Démarche classique via MCMC*

Le principe de base de l'échantillonnage préférentiel est d'estimer une espérance selon une densité f grâce à des réalisations obtenues sous une autre densité h. Par exemple, si on désire estimer $E_f(g(\theta)) = \int g(\theta)f(\theta)d\theta$, il est simple de montrer que $E_f(g(\theta)) = \int \frac{g(\theta)f(\theta)}{h(\theta)}h(\theta)d\theta$ pourvu que le support de f soit inclus dans le support de h. Cette simple remarque permet de fournir deux estimateurs : - l'estimateur classique, $\frac{1}{N}\sum_{j=1}^N g(\theta_j)$ où $(\theta_j)_{j=1,...,N} \sim f$ - et l'estimateur IS, $\frac{1}{N}\sum_{j=1}^N \frac{g(\theta_j)f(\theta_j)}{h(\theta_j)}$ où $(\theta_j)_{j=1,...,N} \sim h$. La fonction h est appelée fonction d'importance et en général, est choisie comme étant simple, rapide à simuler et respectant la condition sur les supports.

Dans le cas particulier de l'étude empirique d'estimateurs, le choix de la fonction d'importance est guidé par d'autres considérations. Le principe de cette méthode peut être présenté pour deux jeux de données $X^{(m)}$ et $X^{(k)}$. Nous supposons disposer déjà des résultats de l'algorithme MCMC relatifs au premier jeu de données $X^{(m)}$ et donc en particulier, des réalisations Markoviennes de θ sous la loi stationnaire $\pi(\theta|X^{(m)})$ obtenues par MCMC et des approximations ergodiques correspondantes de $E_{\pi(\theta|X^{(m)})}g(\theta)$. On désire estimer $E_{\pi(\theta|X^{(k)})}(g(\theta))$ mais cette fois-ci sans utiliser d'algorithmes MCMC. D'après le résultat suivant :

$$E_{\pi(\theta|X^{(k)})}(g(\theta)) = E_{\pi(\theta|X^{(m)})}(g(\theta)\frac{\pi(\theta|X^{(k)})}{\pi(\theta|X^{(m)})}) \tag{2}$$

Une réalisation Markovienne $\{\theta_1, \ldots, \theta_N\}$ sous la loi stationnaire $\pi(\theta|X^{(m)})$ permet alors d'estimer $E_{\pi(\theta|X^{(k)})}(g(\theta))$ par la moyenne empirique $\frac{1}{N}\sum_{i=1}^{N}\frac{g(\theta_i)\pi(\theta_i|X^{(k)})}{\pi(\theta_i|X^{(m)})}$ ou bien par sa version "normalisée" (self-normalised Importance Sampling estimator) :

$$\frac{\sum_{i=1}^{N} g(\theta_i)\frac{\pi(X^{(k)}|\theta_i)}{\pi(X^{(m)}|\theta_i)}}{\sum_{i=1}^{N}\frac{\pi(X^{(k)}|\theta_i)}{\pi(X^{(m)}|\theta_i)}} \rightarrow E_{\pi(\theta|X^{(k)})}[g(\theta)] \quad p.s.. \tag{3}$$

Cette version normalisée fait intervenir uniquement les vraisemblances et non plus les lois a posteriori évitant ainsi le calcul des lois marginales. En effet il est possible de démontrer que :

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\pi(X^{(k)}|\theta_i)}{\pi(X^{(m)}|\theta_i)} \rightarrow \frac{\pi(X^{(k)})}{\pi(X^{(m)})} \quad p.s.. \tag{4}$$

La loi a posteriori $\pi(\theta|X^{(m)})$ est ici la fonction d'importance. D'après la loi forte des grands nombres, les estimateurs ainsi obtenus par échantillonnage préferentiel dans leur version normalisée (formule 3) sont des estimateurs consistants de $E_{\pi(\theta|X^{(k)})}(g(\theta))$. Dans le cas indépendant (et non avec dépendance Markovienne), Geweke (1989) montre sous certaines conditions que les estimateurs IS suivent asymptotiquement une loi normale, résultat étendu au cas markovien par Doss (1994) dans la discussion de l'article de Tierney (1994). Les détails des propriétés asymptotiques des estimateurs se trouvent dans l'article Gajda et al. (2010).

2.2 Choix de la fonction d'importance

La méthode d'échantillonnage préférentiel nécessite le choix d'une fonction d'importance, choix souvent délicat à faire. En effet, dans notre contexte, la fonction d'importance est une loi a posteriori et pour obtenir une bonne estimation, le support de la fonction d'importance doit couvrir le support de la fonction d'intérêt. Plus précisemment, sauf dans le cas particulier de modèles où un paramètre définirait le support, le problème est plutôt de choisir une fonction d'importance qui ne tende pas vers 0 plus rapidement que la loi a posteriori d'origine. Ceci n'est pas évident à savoir quand les densités a posteriori n'ont pas d'expressions explicites. Lors d'une précédente étude (Gajda et al. (2010)), nous avons proposé deux stratégies de choix de la fonction d'importance appelée ici loi a posteriori de " référence " pour le calcul de l'IS. La première stratégie (appelée par la suite stratégie 1) consiste à choisir la fonction d'importance par simple tirage au hasard (équiprobable) d'une seule loi a posteriori. Cette loi a posteriori (appelée "référence fixe") est ensuite utilisée comme unique fonction d'importance pour toutes les estimations. Cette première stratégie est illustrée sur la figure 2 pour une série de $K = 100$ réplications de jeux de données avec $\pi(\theta|X^{(1)})$ comme référence fixe.

L'idée d'utiliser la même distribution a posteriori comme fonction d'importance pour l'ensemble de toutes les estimations peut paraître trop restrictive. Choisir une fonction d'importance parmi un sous-ensemble de distributions pour chaque estimation est une alternative séduisante. La première étape de la deuxième stratégie (stratégie 2) consiste en la présélection d'un petit nombre de jeux de données (au hasard ou par une procédure automatique). La seconde étape consiste à choisir pour chaque nouvelle estimation (et donc chaque nouveau jeu de données) une loi a posteriori de référence (appelée "référence choisie")

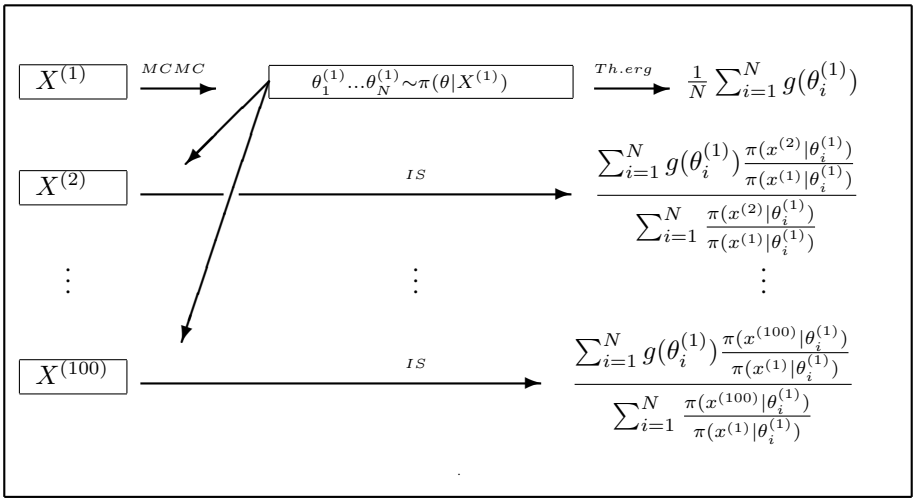

FIG. 2 – *Démarche via IS avec référence fixe*

parmi les distributions présélectionnées. Pour un total de 100 jeux de données et 10 jeux de données présélectionnés, cette démarche est présentée sur la figure 3. Cette deuxième stratégie nécessite donc de savoir choisir une loi a posteriori de référence (parmi les distributions présélectionnées) "adaptée" au jeu de données sur lequel les estimations sont faites. Pour cela, plusieurs critères de choix ont été proposés : Le premier est fondé sur la minimisation de la norme L_1 de la différence entre deux lois a posteriori, le deuxième sur la minimisation de la divergence de Kullback-Leibler et le troisième sur la minimisation de la variance de l'estimation MCMC.

L'avantage de la première stratégie est la rapidité car une seule fonction d'importance est utilisée pour toutes les estimations mais parfois au prix d'une estimation moins performante alors que la deuxième stratégie amène à de meilleures estimations mais moins rapidement.

Dans ce travail, nous proposons une troisième stratégie (stratégie 3) où la fonction d'importance est une densité de mélange des lois présélectionnées. Ainsi, cette démarche offre l'avantage à nouveau d'être fondée sur une seule fonction d'importance pour l'ensemble des estimations avec, de plus, un support plus large que la fonction d'importance de la première stratégie. Si J est le nombre de lois a posteriori présélectionnés (dans notre exemple, J=10), alors la fonction d'importance π_{mix} est de la forme

$$\pi_{mix}(\theta) = \sum_{j=1}^{J} \frac{n_j}{|n|} \pi(\theta|X^{(j)}) = \sum_{j=1}^{J} \frac{n_j}{|n|} c_j h_j(\theta)$$

où $\pi(\theta|X^{(j)}) = c_j h_j(\theta)$ (version normalisée ou non de la $j^{\text{ième}}$ densité a posteriori), n_j est le nombre de réalisations Markoviennes obtenues via MCMC sous la loi $\pi(\theta|X^{(j)})$ et $|n| = n_1 + \cdots + n_J$.

Mélange de distributions et échantillonnage préférentiel combiné avec MCMC

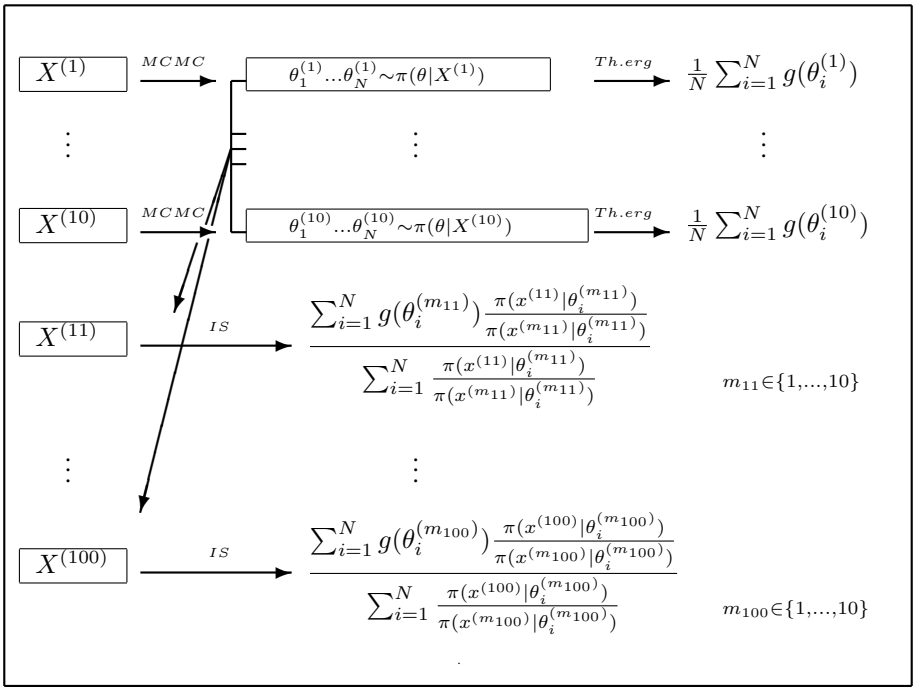

FIG. 3 – *Démarche via IS avec référence choisie*

Si $(\theta_1^{(j)}, \cdots, \theta_{n_j}^{(j)})$ est l'ensemble des n_j réalisations sous la loi $\pi(\theta|X^{(j)})$ alors, pour $k > J$, :

$$\frac{1}{|n|}\sum_{j=1}^{J}\sum_{i=1}^{n_j} g(\theta_i^{(j)})\frac{\pi(\theta_i^{(j)}|X^{(k)})}{\pi_{mix}(\theta_i^{(j)})} \to E_{\pi(\theta|X^{(k)})}(g(\theta))$$

Les coefficients du mélange dépendent de constantes de normalisation $(c_1, \cdots c_J)$ qui ne sont pas explicites. Nous proposons de les estimer par deux méthodes décrites dans la partie 2.3.

Les constantes de normalisations estimées sont notées $(\hat{c}_1, \cdots \hat{c}_J)$. L'estimation de $E_{\pi(\theta|X^{(k)})}(g(\theta))$ par échantillonnage préférentiel est :

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j} w(\theta_i^{(j)}) g(\theta_i^{(j)})$$

où

$$w(\theta_i^{(j)}) = \frac{\pi(X^{(k)}|\theta_i^{(j)})/\pi_{mix}(\theta_i^{(j)})}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\pi(X^{(k)}|\theta_i^{(j)})/\pi_{mix}(\theta_i^{(j)})}$$

et

$$\pi_{mix}(\theta) = \sum_{j=1}^{J}\frac{n_j}{|n|}\hat{c}_j h_j(\theta).$$

Pour évaluer et comparer les qualités des estimateurs obtenus par les trois stratégies, nous avons calculé les racines carrées des erreurs quadratiques moyennes, D et $\tilde{D}$. D peut être interprétée comme une distance entre les estimations et les vraies valeurs a posteriori et $\tilde{D}$ comme une distance entre les estimations via IS et les estimations obtenues par l'approche classique MCMC. En effet, si les expressions des espérances sous la loi a posteriori ne sont pas explicites alors la distance D n'est pas calculable et sera donc remplacée par $\tilde{D}$. Plus précisemment, pour les $K-J$ jeux de données (en dehors des jeux de données présélectionnés), les expressions de D et de $\tilde{D}$ sont les suivantes :

$$D = \left[\frac{1}{K-J} \sum_{k=J+1}^{K} (IS_k - E_{\pi(\theta|X^{(k)})}(g(\theta)))^2 \right]^{1/2} \tag{5}$$

et

$$\tilde{D} = \left[\frac{1}{K-J} \sum_{k=J+1}^{K} (IS_k - MCMC_k)^2 \right]^{1/2} \tag{6}$$

où IS_k est l'estimation de $E_{\pi(\theta|X^{(k)})}(g(\theta))$ obtenue par échantillonnage préférentiel et $MCMC_k$ celle obtenue par l'approche "classique" via MCMC. Le critère de comparaison choisi est l'erreur quadratique moyenne comme lors de l'étude précédente (Gajda et al. (2010)). D'autres critères pourraient être utilisés comme par exemple le risque bayésien intégré (Robert (2007)) dont l'admissibilité et l'optimalité pourraient être ensuite jugées par rapport à la borne de Cramér-Rao bayésienne (Gill et Levit (1995)).

Un premier exemple présente les résultats dans le cas d'un modèle où toutes les expressions a posteriori sont explicites. Les estimations seront donc comparées aux vraies valeurs en utilisant D. Dans un deuxième exemple, les calculs analytiques n'étant plus possibles, $\tilde{D}$ sera évaluée.

2.3 Constantes de normalisation

Nous proposons deux méthodes pour estimer les coefficients du mélange $(c_1, \cdots c_J)$ intervenant dans π_{mix}.

La première méthode dite de "Reverse Logistic Regression" a été proposée par Geyer (1993). En utilisant la reparamétrisation suivante :

$$\eta_j = \log c_j + \log \frac{n_j}{|n|}$$

alors la probabilité que θ appartienne à la $j^{\text{ème}}$ composante du mélange est :

$$p_j(\theta, \eta) = \frac{h_j(\theta) e^{\eta_j}}{\sum_{i=1}^{J} h_i(\theta) e^{\eta_i}}$$

où $\eta = (\eta_1, \cdots, \eta_J)$. Geyer (1993) propose d'estimer les coefficients η_j et donc les coefficients c_j par maximisation de la log-vraisemblance $l(\eta) = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \log p_j(\theta_i^{(j)}, \eta)$. Les détails des procédures itératives de maximisation se trouvent dans Geyer (1993).

La seconde méthode utilise à nouveau la théorie ergodique et l'échantillonnage préférentiel. En effet, d'après l'égalité $\pi(\theta|X^{(j)}) = c_j h_j(\theta) = c_j \pi(X^{(j)}|\theta)\pi(\theta)$ alors

$$c_j = \frac{\pi(\theta|X^{(j)})}{\pi(X^{(j)}|\theta)\pi(\theta)}.$$

Si f est une densité de probabilité dont le support est inclus dans le support de $\pi(\theta|X^{(j)})$ alors

$$c_j = \int \frac{f(\theta)}{\pi(X^{(j)}|\theta)\pi(\theta)} \pi(\theta|X^{(j)}) d\theta.$$

Ainsi, d'après la théorie ergodique, si $\{\theta_1^{(j)}, \ldots, \theta_{n_j}^{(j)}\}$ sont des réalisations Markoviennes sous la loi stationnaire $\pi(\theta|X^{(j)})$ alors

$$\hat{c}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} \frac{f(\theta_i^{(j)})}{\pi(X^{(j)}|\theta_i^{(j)})\pi(\theta_i^{(j)})} \to c_j \quad p.s.$$

Pour ce calcul, $f(\theta_i^{(j)})$ a été choisie comme la densité d'une loi normale de moyenne $\hat{\mu}_j = \frac{\sum_{i=1}^{n_j} \theta_i^{(j)}}{n_j}$ (estimation de l'espérance a posteriori de θ sous $\pi(\theta|X^{(j)})$) et de variance $\hat{\sigma}_j^2 = \frac{\sum_{i=1}^{n_j} \theta_i^{(j)^2}}{n_j} - (\hat{\mu}_j)^2$ (estimation de la variance a posteriori de θ sous $\pi(\theta|X^{(j)})$).

3 Exemples

Dans cette partie, deux exemples sont présentés dans le cadre poissonnien. Le premier exemple est un modèle simple où les lois a posteriori sont explicites et le second exemple est un modèle linéaire généralisé offrant l'avantage d'être plus réaliste mais où cette fois-ci les lois a posteriori ne sont plus explicites. Les approches présentées dans ce papier reposent sur la convergence des algorithmes MCMC. Il est donc essentiel de la vérifier. Comme cela est suggéré dans les articles de Brooks et Roberts (1998) et Mengersen et al. (1999), plusieurs diagnostics de convergence doivent être utilisés (et non un seul). Dans notre étude, la convergence a été étudiée en vérifiant que les erreurs de Monte Carlo sont inférieures à 5% des écarts types a posteriori des paramètres, et en utilisant les diagnostics de Gelman et Rubin ainsi que des outils graphiques (traces des paramètres avec bonne mélangeance, autocorrelations des paramètres). L'ensemble de ces outils est accessible sous R dans la librairie BRuGS (R Development Core Team (2008)). Concernant les résultats présentés ici, 50000 itérations avec un "temps de chauffe" (Burn-in) de 5000 itérations ont été utilisées dans les algorithmes MCMC.

3.1 Modèle de Poisson

Le premier exemple d'application des méthodes présentées ci-dessus est un modèle simple de Poisson avec le paramètre de moyenne λ. L'avantage de cet exemple est qu'en choisissant comme loi a priori la loi conjugée Gamma, la loi a posteriori de λ est explicite. Ainsi, la comparaison avec les vraies valeurs a posteriori est également possible. En effet, pour un jeu de données $X^{(k)} = (X_1^{(k)}, \ldots, X_n^{(k)})$ où les $X_i^{(k)}$ sont indépendantes et de même loi de Poisson

$\mathcal{P}(\lambda)$ et pour la loi a priori $\lambda \sim \mathcal{G}(\alpha, \beta)$, la densité a posteriori de λ sachant $X^{(k)}$ est aussi une loi Gamma $\mathcal{G}(\sum_{i=1}^{n} X_i^{(k)} + \alpha, \beta + n)$. La vraie moyenne et la vraie variance a posteriori sont égales à $(\sum_{i=1}^{n} X_i^{(k)} + \alpha)/(\beta + n)$ et $(\sum_{i=1}^{n} X_i^{(k)} + \alpha)/(\beta + n)^2$ respectivement.

Nous avons simulé un total de 100 jeux de données de taille $n = 20$ selon la loi de Poisson avec la paramètre $\lambda = 20$. Ensuite, pour chaque simulation, la moyenne et la variance de la loi a posteriori du paramètre ont été estimées de deux manières différentes : classiquement via MCMC et via Importance Sampling combiné avec MCMC comme présenté précédemment selon les trois stratégies.

La première stratégie consiste à utiliser une seule loi a posteriori de référence choisie au hasard pour l'ensemble des autres jeux de données. Afin d'envisager toutes les possibilités, nous avons étudié les résultats pour toutes les lois a posteriori possibles de référence : $\pi(\lambda|X^{(1)}), \ldots, \pi(\lambda|X^{(100)})$. Les résultats présentés dans le tableau 1 sur les lignes intitulées "la pire réf. fixe" et "la meilleur réf. fixe" correspondent respectivement aux résultats associés au choix des lois a posteriori de référence qui donnaient les plus grandes valeurs et les plus petites valeurs du critère D par rapport aux vraies valeurs a posteriori. Pour la deuxième stratégie, les dix lois a posteriori présectionnées correspondent simplement aux dix premières.

Ces dix premières lois a posteriori préselectionnées ont été aussi utilisées pour la construction de la loi mélange comme fonction d'importance selon la troisième stratégie.

Le tableau 1 présente les valeurs de D (voir formule 5) par rapport aux vraies valeurs des moyennes a posteriori selon les différentes statégies.

Statégie		D
	mcmc	9.5
1	la pire réf. fixe	346.8
1	la meilleure réf. fixe	33.9
2	critère 1	6.5
2	critère 2	11.2
2	critère 3	6.7
3	mélange	6.3

TAB. 1 – *Valeurs de D (multipliées par 10^3), racine carrée des erreurs quadratiques moyennes de la moyenne a posteriori dans le modèle de Poisson avec $\lambda = 20$.*

L'ordre de grandeur des erreurs quadratiques reste très petit sauf dans le cas de la première stratégie, en particulier dans le cas de la mauvaise référence fixe. Les deux critères et la technique du mélange ont donné des erreurs quadratiques nettement inférieures à celles de la référence fixe (première stratégie). De plus, les deuxième et troisième stratégies présentent l'avantage d'éviter une fonction d'importance choisie par hasard parmi les 100. L'approche classique par MCMC étant elle-même une approximation, il est intéressant de voir ici que les résultats obtenus avec IS sont légèrement meilleurs qu'avec l'approche classique. Concernant les variances a posteriori (résultats non montrés), les conclusions sont à peu près identiques.

3.2 Modèle linéaire généralisé

Le second exemple concerne un exemple de modèle linéaire généralisé. Afin d'étudier les avantages des approches combinées avec l'échantillonnage préférentiel en terme de précisions d'estimation et de temps de calcul, cet exemple concerne une grande taille d'échantillon (n=1000) avec 10 covariables. Le modèle est le suivant :

$X_i|\lambda \sim \mathcal{P}(\lambda_i)$

$log(\lambda_i) = a + \sum_{j=1}^{10} b_j Z_{ij}$

Des lois a priori vagues $\mathcal{N}(0, 10^5)$ ont été choisies pour les coefficients a et $\{b_j, j = 1, \ldots, 10\}$. Pour les simulations des 100 jeux de données, les valeurs choisies sont : $a = 0$, $b_j = 0.05$ pour $j = 1, \ldots, 10$. Les valeurs de $\tilde{D}$ (voir formule 6) ont été calculées dans ce cas car les expressions analytiques des lois a posteriori et de leurs moments ne sont plus explicites. Concernant les résultats des estimations des coefficients $\{b_j, j = 1, \ldots, 10\}$, la moyenne des 10 valeurs de $\tilde{D}$ a été évaluée pour chaque stratégie. Pour la stratégie 2, les trois critères ont donné des résultats très proches (0.8×10^{-2}, 1.6×10^{-2} and 1.1×10^{-2} pour les critères 1, 2 and 3 respectivement). Concernant la stratégie 1, la moyenne de $\tilde{D}$ est de 3.8×10^{-2} en choisissant comme référence fixe la première distribution a posteriori. La moyenne de $\tilde{D}$ pour la stratégie 3 est 0.6×10^{-2}. Ces résultats confirment les résultats du premier exemple simple à savoir de meilleures précisions pour les stratégies 2 et 3 que pour la stratégie 1 avec ici une légère meilleure performance de la stratégie 3.

Dans ce deuxième exemple, les performances en terme de temps de calcul ont été évaluées. La table 2 montre les rapports des temps entre les différentes stratégies et le temps de calcul de l'approche "classique" MCMC. Tous ces calculs ont été faits sur le même ordinateur et avec le même nombre d'itérations. La stratégie 1 qui ne nécessite qu'une fonction d'importance pour l'ensemble des estimations réduit nettement le temps de calcul par rapport à l'approche MCMC car demande environ que 6% du temps de l'approche MCMC classique. Par contre, les gains en temps pour la stratégie 2 sont beaucoup moins nets voir supérieurs pour le critère 3. Rappelons que pour ce critère, une fonction d'importance différente est choisie pour chaque nouveau jeu de données mais également pour chaque fonction g intervenant dans l'espérance. Le gain de la stratégie 3 reste moins bon que celui de la stratégie 1 mais nettement meilleur que celui de la stratégie 2. En effet, la stratégie 3 demande moins de la moitié du temps (42%) que l'approche classique MCMC pour des qualités d'estimation proches de celles obtenues par la stratégie 2.

TAB. 2 – *Rapport des temps de calculs entre les méthodes combinées avec l'échantillonnage préférentiel et la méthode classique MCMC concernant la régression de Poisson avec* $n = 1,000$ *et* 10 *covariables.*

Stratégie 1	Stratégie 2 critère 1 (L_1)	Stratégie 2 critère 2 (KL)	Stratégie 2 critère 3	Stratégie 3 Mélange
0.057	0.957	0.771	1.156	0.423

4 Conclusions

Les estimations par échantillonnage préférentiel sont, en général, proches des estimations obtenues classiquement via MCMC en étant également proches des vraies valeurs a posteriori. Concernant la comparaison des différentes stratégies proposées avec IS, les erreurs quadratiques calculées avec les lois a posteriori de référence choisies par les critères (stratégie 2) ou avec la loi mélange (stratégie 3), ont toujours été plus petites que les erreurs quadratiques obtenues avec la stratégie 1 "référence fixe", en particulier quand la fonction d'importance correspondait au pire choix comme référence fixe.

Les exemples présentés ci-dessus montrent que dans le cadre des estimations répétées, le choix de la fonction d'importance n'est pas évident à faire même pour un modèle simple. Parmi toutes les références fixes possibles, il existe des cas qui peuvent amener à de très mauvais résultats. Il est impossible d'identifier la loi de référence fixe qui aboutirait à des résultats convenables. Il est donc utile de mettre en place d'autres stratégies. L'idée d'une loi de mélange comme forme de loi de référence unique a donné des erreurs quadratiques qui se sont révélées être toujours plus petites que celles avec la stratégie "référence fixe" et de très bons résultats en terme de gain de temps de calculs. Ainsi, cette approche offre le double avantage d'avoir une seule fonction d'importance pour l'ensemble des estimations (donc d'être rapide) et de donner de bons résultats en terme de qualités d'estimations.

Concernant les algorithmes MCMC, le nombre d'itérations utilisé dans les estimations est bien souvent inférieur au nombre d'itérations réellement effectué et ceci pour plusieurs raisons. En premier lieu, une partie des itérations initiales (dite "période de chauffe") est écartée afin de garantir que la chaîne de Markov soit sous le régime stationnaire. En second lieu, souvent seul un sous-échantillonnage des itérations est conservé, par exemple une itération sur dix ou vingt. Même s'il est toujours avantageux en terme de précision d'estimateur de garder toutes les itérations (MacEachern et Berliner (1994), Geyer (1992)), ce sous-échantillonnage peut être intéressant pour des raisons d'amélioration de la mélangeance de la chaîne ou bien simplement de stockage. Ainsi, une proportion parfois non négligeable des itérations est rejetée. Enfin, lors de chaque itération, des simulations sont nécessaires. Par exemple, l'algorithme de Gibbs requiert des simulations sous la loi conditionnelle complète de chaque paramètre. Si ces lois conditionnelles ne sont pas connues, une étape de Hasting Metropolis est souvent introduite nécessitant une simulation sous une loi dite instrumentale. Ainsi, il est souvent plus coûteux en temps d'obtenir les réalisations de la loi a posteriori via un algorithme MCMC que de les manipuler ensuite comme dans le calcul de l'échantillonnage préférentiel, et ceci même quand la proportion des itérations rejetées est faible. Typiquement, la troisième stratégie montre un gain en temps de calcul alors qu'elle requiert un nombre d'évaluations des vraisemblances plus important que par une procédure MCMC classique. En effet, ces évaluations doivent être faites sur l'ensemble des réalisations markoviennes obtenues par les J procédures MCMC. Cette troisième stratégie est donc particulièrement adaptée quand la vraisemblance est explicite et non approchée numériquement (par à nouveau une procédure itérative). En effet, si dans le déroulement de l'algorithme MCMC, l'évaluation de la vraisemblance (ou d'une de ses composantes) est non négligeable par rapport à l'utilisation de générateurs de nombre pseudo aléatoire, elle le sera aussi dans le calcul d'échantillonnage préférentiel. La comparaison exacte de la complexité numérique entre MCMC et les stratégies combinées utilisant l'échantillonnage préférentiel est difficile à faire de manière générale car dépend du modèle étudié et du type d'algorithme MCMC utilisé, ceci reste néanmoins une piste de recherche à approfondir.

Afin de poursuivre ce travail, il serait intéressant d'étudier les performances de ces différentes approches dans le cadre des modèles linéaires généralisés mixtes qui demandent souvent des temps de calculs importants.

Références

Brooks, S. P. et G. O. Roberts (1998). Convergence assessment techniques for Markov chain Monte Carlo. *Statistics and Computing 8*(4), 319–335.

Doss, H. (1994). Discussion of the paper "Markov chains for exploring posterior distributions" by luke tierney. *Ann. Statist. 22*(4), 1728–1734.

Gajda, D., C. Guihenneuc-Jouyaux, J. Rousseau, K. Mengersen, et D. Nur (2010). Use in practice of importance sampling for repeated MCMC for Poisson models. *Electron. J. Statist. 4*, 361–383.

Gelfand, A., D. Dey, et H. Chang (1992). Model determination using predictive distributions with implementation via sampling-based methods. In B. J. D. A. Bernardo, J.M. et A. Smith (Eds.), *Bayesian Statistics*, Volume 4, pp. 147–167. Oxford University Press.

Geman, S. et D. Geman (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence 6*, 721–740.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica 57*(6), 1317–1339.

Geyer, C. et E. Thompson (1992). Constrained Monte Carlo Maximum Likelihood for Dependent Data (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological) 54*(3), 657–699.

Geyer, C. J. (1992). Practical Markov Chain Monte Carlo. *Stat. Sci. 7*(4), 473–511.

Geyer, C. J. (1993). Estimating Normalizing Constants and Reweighting Mixtures in Markov Chain Monte Carlo (revision). Technical Report No. 568 R(4), School of Statistics, University of Minnesota, http ://www.stat.umn.edu/PAPERS/tr568r.html.

Gill, R. et B. Levit (1995). Application of the van Trees inequality : a Bayesian Cramér-Rao bound. *Bernoulli 1*, 59–79.

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika 57*(1), 97–109.

MacEachern, S. N. et M. L. Berliner (1994). Subsampling the Gibbs Sampler. *Am. Stat. 48*(3), 188–190.

Mengersen, K. L., C. P. Robert, et C. Guihenneuc-Jouyaux (1999). MCMC convergence diagnostics : a reviewww. In *Bayesian statistics, 6 (Alcoceber, 1998)*, New York, pp. 415–440. Oxford Univ. Press.

R Development Core Team (2008). *R : A Language and Environment for Statistical Computing*. Vienna, Austria : R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Robert, C. (2007). *The Bayesian Choice. From Decision-Theoretic Foundations to Computational Implementation* (2 ed.). Springer Texts in Statistics. Springer.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist. 22*(4), 1701–1762. With discussion and a rejoinder by the author.

Summary

The Importance Sampling method is used in combination with MCMC in Bayesian simulation study. In the particular context of numerous simulated data sets, MCMC algorithms have to be called several times which may become computationally expensive. We propose to use MCMC on a preselected set of the simulated data in order to obtain Markovian realisations of each corresponding posterior distribution. The estimates for the other simulated data are computed via IS based on this preselected data set. Since Importance Sampling requires the choice of an importance function, we propose a strategy for this choice based on a mixture of the preselected posterior distributions. The featured methods are illustrated in simulations studies under two different Poisson models.