Une approche Monte Carlo adaptative pour l'approximation de lois a posteriori avec application à l'inférence de paramètres cosmologiques

Olivier Cappé[1], Christian P. Robert[2]

[1]LTCI, Télécom Paris, CNRS
46 rue Barrault, 75634 Paris cedex 13, France

[2]Ceremade, Université Paris-Dauphine, CREST-INSEE
Place du Maréchal de Lattre de Tassigny, 75775 Paris cedex 16, France
cappe à enst.fr, xian à ceremade.dauphine.fr

Résumé – Cette communication décrit une approche de Monte Carlo adaptative reposant sur l'échantillonnage préférentiel itéré et permettant d'approcher une loi a posteriori arbitraire (connue à une constante de proportionnalité près) par un mélange de lois gaussiennes. On discute des mérites de cette approche, comparée à l'état de l'art basé sur l'utilisation d'algorithmes de Monte Carlo par Chaîne de Markov (MCMC), dans le cadre de l'estimation de paramètres de modèles cosmologiques à partir d'observables comme le spectre du fond de rayonnement cosmique.

Abstract – This contribution presents an adaptive iterative importance sampling scheme in which the instrumental distribution is represented by a mixture distribution whose parameters are adjusted so as to minimise the Kullback divergence to the target distribution. Applications of this approach for inference of cosmological parameters from astrophysical measurements are discussed.

1 Introduction

De nombreux modèles utilisés en traitement de signal et au delà reposent sur l'approche bayésienne dans laquelle les observations Y sont supposées suivre un modèle probabiliste $\ell(y|x)$ dépendant d'un paramètre X lui-même muni d'une *loi a priori* $\mu(x)$. Dans ces conditions, l'inférence sur le paramètre X est liée à la détermination des propriétés (moyenne, variance, quantiles, mode, etc.) de sa *loi a posteriori* définie par la formule de Bayes

$$\pi(x|Y) = \frac{\ell(Y|x)\mu(x)}{\int \ell(Y|x')\mu(x')dx'} \tag{1}$$

Malgré la simplicité apparente de l'équation ci-dessus il est bien connu que l'inférence bayésienne ne conduit à des expressions explicites que dans un nombre réduit de modèles relativement simples. Dans le cas général, le calcul de *l'évidence* $\int \ell(Y|x)\mu(x)dx$ est impossible, de même que celui des moments (ou autres caractéristiques numériques) de la loi a posteriori $\pi(x|Y)$. La seule connaissance généralement exploitable est le fait que la loi a posteriori est donnée, à une constante de proportionnalité inconnue près, par $\ell(Y|x)\mu(x)$. Dans la suite, et pour alléger les notations, nous noterons simplement la loi d'intérêt $\pi(x)$, en omettant de figurer la dépendance en les données observées Y.

Si x est élément d'un espace $\mathcal{X}$ de grande dimension, les approches les plus simples reposant sur une exploration systématique de $\mathcal{X}$ (grilles, etc.) sont vouées à l'échec. Raison pour laquelle les méthodes d'inférence bayésienne reposant sur l'utilisation de simulations pseudo-aléatoires ont connu un intense développement depuis une quinzaine d'années. Dans ce contexte, l'approche la plus générique et la plus utilisée à ce jour est celle dite de Monte Carlo par Chaîne de Markov (MCMC) [1]. Nous nous intéressons ici à l'inférence de paramètres de modèles cosmologiques à partir de mesures d'observables comme le spectre du fond de rayonnement cosmique (mais également : effets de lentillage cosmique, caractéristiques issues de comptage de galaxies, ...) Dans ce cadre, l'approche MCMC est très largement utilisée du fait du caractère particulièrement complexe de la fonction de vraisemblance ℓ [2, 3]. Les difficultés pratiques rencontrées dans ce cadre sont, d'une part, la nécessité de calibrer des paramètres algorithmiques (opération qui requiert, parfois, une assez grande expertise) et, d'autre part, des questions de coût de calcul essentiellement liées à la complexité de l'évaluation de la vraisemblance ℓ.

Nous nous intéressons ici à l'approche dite *Adaptive Population Monte Carlo* qui permet de régler automatiquement, de façon adaptative, les paramètres d'une classe d'algorithmes de simulation [4, 5, 6] s'inspirant du principe des méthodes de Monte Carlo séquentiel (ou « filtrage particulaire »). Cette contribution se propose de montrer comment l'approche de [5] peut être étendue pour déterminer, via une méthode d'échantillonnage préférentiel itératif, la meilleure approximation (dans un sens précisé ci-dessous) de la loi a posteriori $\pi(x)$ dans une famille q_θ de lois de mélange de la forme

$$q_\theta(x) = \sum_{k=1}^{d} \alpha_k f_{\lambda_k}(x) \tag{2}$$

où $(\alpha_k)_{1\leq k\leq n}$ sont les poids du mélange ($0 < \alpha_k < 1$ et $\sum_{k=1}^{d} \alpha_k = 1$) et $\{f_\lambda; \lambda \in \Lambda\}$ est une famille paramétrique de densités. Par rapport à [5], l'optimisation a lieu non seulement sur les poids α_k mais également sur les paramètres λ_k des composantes. Même si on conserve la notation λ_k par souci de généralité, nous nous intéresserons ici en pratique au cas du mélange de gaussiennes (multivariées) $f_{\lambda_k}(x) = \mathcal{N}_{\mu_k,\Sigma_k}(x)$ pour lequel λ_k recouvre le vecteur moyen μ_k et la matrice de covariance Σ_k.

2 Les principes

Comme dans [5], on utilise ici le critère de divergence de Kullback (ou entropie relative)

$$K(\pi||q_\theta) = \int \log \frac{\pi(x)}{q_\theta(x)} \pi(x)dx \tag{3}$$

pour mesurer la proximité entre la loi a posteriori π et le mélange q_θ. Une fois la phase d'adaptation effectuée, on utilisera q_θ comme distribution instrumentale pour approximer les espérances de fonctions h sous la loi a posteriori par *échantillonnage préférentiel auto-normalisé*, c'est à dire :

$$\bar{h} = \int h(x)\pi(x)dx \approx \hat{h}_n = \sum_{i=1}^{n} \omega_i h(X_i) \tag{4}$$

où $(X_i)_{1\leq i\leq n}$ sont simulés sous la loi q_θ et $(\omega_i)_{1\leq i\leq n}$ désignent les *poids d'importance normalisés* définis par

$$\omega_i = \frac{\pi(X_i)/q_\theta(X_i)}{\sum_{j=1}^{n} \pi(X_j)/q_\theta(X_j)}$$

Sous la condition $\int (1 + h^2(x))\pi^2(x)/q_\theta(x)dx < \infty$, $\hat{h}_n$ est un estimateur asymptotiquement normal de $\bar{h}$ (à la vitesse $1/\sqrt{n}$) avec une variance asymptotique donnée par [7]

$$\int (h(x) - \bar{h})^2 \pi^2(x)/q_\theta(x)dx \tag{5}$$

Il est également possible d'estimer empiriquement cette variance limite — ce qui constitue un avantage substantiel de l'échantillonnage préférentiel par rapport aux approches MCMC — sous la forme $n \sum_{i=1}^{n} \omega_i^2 (h(X_i) - \hat{h}_n)^2$.

Notons à ce stade deux remarques importantes. Tout d'abord, il est possible d'étendre le critère (3) à une classe plus large d'algorithmes que le simple échantillonnage préférentiel [5], même si le caractère global de ce critère semble mieux adapté au cas de l'échantillonnage préférentiel. Il est également tentant d'utiliser comme critère d'optimalité, non pas la divergence de Kullback définie en (3) mais directement la variance asymptotique de l'estimateur final définie en (5). Cette approche, suivie dans [6] présente néanmoins deux limitations : d'une part, il faut privilégier a priori une fonction particulière d'intérêt h puisque la variance limite dépend de la fonction considérée ; d'autre part, la formule de mise à jour des poids $(\alpha_k)_{1\leq k\leq d}$ du mélange obtenue dans [6] repose sur un constat qu'il n'est pas possible d'étendre au cas plus général où l'on souhaite également adapter les paramètres des composantes f_{λ_k} du mélange. Par ailleurs, un intérêt du critère (3) qui n'a pas été relevé dans la littérature est le fait que

$$\exp\{-K(\pi||q_\theta)\} = \exp\left(\int -\log \frac{\pi_{\mathrm{nn}}(x)}{q_\theta(x)} \pi(x)dx\right) \left(\int \pi_{\mathrm{nn}}(x)dx\right) \tag{6}$$

où π_{nn} désigne la version non normalisée de π, c'est-à-dire la seule quantité dont on dispose en pratique. En estimant la première intégrale de (6) par échantillonnage préférentiel auto-normalisé,

$$-\sum_{i=1}^{n} \omega_i \log \frac{\pi_{\mathrm{nn}}(X_i)}{q_\theta(X_i)}$$

et la seconde par échantillonnage préférentiel classique $(1/n \sum_{i=1}^{n} \pi_{\mathrm{nn}}(X_i)/q_\theta(X_i))$, on en déduit que $\exp(H_n)/n$, où $H_n = -\sum_{i=1}^{n} \omega_i \log \omega_i$ désigne l'*entropie* associée aux poids normalisés, est un estimateur consistant de (6). En d'autre termes, la minimisation de la divergence de Kullback $K(\pi||q_\theta)$ est directement reliée à la maximisation de la *perplexité* ($\exp(H_n)$) normalisée des poids d'importance normalisés $(\omega_i)_{1\leq i\leq n}$, qui est l'un des critères empiriques classiquement utilisés pour jauger la performance de l'échantillonnage préférentiel (une valeur proche de 1 indiquant une situation proche de l'échantillonnage sous π, tandis qu'une valeur de proche de 0 correspond à des poids d'importance très inégaux et donc à un mauvais choix de la loi instrumentale q_θ) [7].

3 L'algorithme d'adaptation

La minimisation de la divergence définie en (3) est équivalente à la *maximisation* du critère

$$C(\theta) = \int \log q_\theta(x)\pi(x)dx = \int \log \left(\sum_{k=1}^{d} \alpha_k f_{\lambda_k}(x)\right) \pi(x)dx$$

Cette maximisation est bien sûr difficile, même en admettant pour l'instant que l'intégration par rapport à la loi $\pi(x)$ soit faisable. Une remarque peu connue, bien qu'implicite dans la construction dite *information bottleneck* [8], est qu'il existe une approche — que nous appellerons, faute de mieux, *algorithme EM intégré* en référence à l'algorithme *Expectation-Maximisation* — permettant de maximiser $C(\theta)$ itérativement. En introduisant la variable indicatrice Z du mélange telle que $\mathrm{P}(Z = k) = \alpha_k$, pour $k = 1, \ldots, d$, et $p(x|Z = k) = f_{\lambda_k}(x)$, on peut définir les probabilités conditionnelles

$$\rho_\theta(k|x) = \mathrm{P}_\theta(Z = k|x) = \frac{\alpha_k f_{\lambda_k}(x)}{\sum_{\ell=1}^{d} \alpha_\ell f_{\lambda_\ell}(x)}$$

puis la quantité intermédiaire

$$Q_\theta(\theta') = \int \left\{\sum_{k=1}^{d} \rho_\theta(k|x) \log\left(\alpha'_k f_{\lambda'_k}(x)\right)\right\} \pi(x)dx$$

En utilisant la concavité du logarithme, on montre directement, comme dans le cas de l'algorithme EM usuel appliqué aux mélanges, que

$$Q_\theta(\theta') - Q_\theta(\theta) \leq C(\theta') - C(\theta)$$

c'est à dire que les maximisations successives de la quantité intermédiaire $Q_\theta(\theta')$ par rapport à θ' conduisent à une séquence de valeurs croissantes du critère $C(\theta)$. On vérifie aisément que les formules de mises à jour des paramètres sont similaires à celles de l'algorithme EM usuel appliqué au cas d'un mélange, à l'intégration par π près. Dans le cas du mélange de gaussiennes, on obtient

$$\alpha'_k = \int \rho_\theta(k|x)\pi(x)dx, \quad \mu'_k = \int x\rho_\theta(k|x)\pi(x)dx,$$
$$\Sigma'_k = \int (x-\mu'_k)(x-\mu'_k)^T\rho_\theta(k|x)\pi(x)dx \qquad (7)$$

Evidemment, l'intégration nécessaire au calcul des trois quantités ci-dessus n'est pas réalisable exactement et on se propose d'utiliser l'échantillonnage préférentiel avec la densité instrumentale q_θ définie par les paramètres courant θ pour les estimer. Chaque itération de l'algorithme adaptatif se décompose comme suit :

1. Etant données les estimations courantes des paramètres $(\alpha_k, \mu_k, \Sigma_k)_{1\le k\le d}$, on génère des points sous la loi de mélange, c'est à dire des couples $\{Z_i, X_i\}_{1\le i\le n}$ tels que $\mathrm{P}(Z_i = k) = \alpha_k$ et X_i suive une loi $\mathcal{N}_{\mu_k,\Sigma_k}$ lorsque $Z_i = k$.
2. On calcule les poids d'importance
$$\frac{\pi(X_i)}{\sum_{k=1}^d \alpha_k \mathcal{N}_{\mu_k,\Sigma_k}(X_i)}$$
en les normalisant par leur somme pour en déduire $(\omega_i)_{1\le i\le n}$.
3. Les quantités définies en (7) sont approchées par échantillonnage préférentiel sous la forme
$$\alpha'_k = \sum_{i=1}^n \omega_i \mathbb{1}\{Z_i = k\}, \quad \mu'_k = \sum_{i=1}^n \omega_i X_i \mathbb{1}\{Z_i = k\},$$
$$\Sigma'_k = \sum_{i=1}^n \omega_i (X_i-\mu'_k)(X_i-\mu'_k)^T \mathbb{1}\{Z_i = k\} \qquad (8)$$

Les formules ci-dessus ont une forme très simple et raisonnablement intuitive : pour la mise à jour du poids α_k, par exemple, α'_k correspond à la fraction totale des poids d'importance normalisés pour les points effectivement simulés sous la k-ième composante du mélange ; α'_k sera donc d'autant plus grand lorsque les points simulés sous la k-ième composante du mélange conduisent, en moyenne, à des poids d'importance élevés. Pour justifier plus formellement ces relations, notons que pour une fonction h quelconque

$$\mathrm{E}\left[\frac{\pi(X_i)}{\sum_{\ell=1}^d \alpha_\ell f_{\lambda_\ell}(X_i)} h(X_i)\mathbb{1}\{Z_i = k\}\right]$$
$$= \mathrm{P}(Z_i = k)\mathrm{E}\left[\frac{\pi(X_i)}{\sum_{\ell=1}^d \alpha_\ell f_{\lambda_\ell}(X_i)} h(X_i)\,\middle|\, Z_i = k\right]$$
$$= \alpha_k \int \frac{\pi(x)}{\sum_{\ell=1}^d \alpha_\ell f_{\lambda_\ell}(x)} h(x) f_{\lambda_k}(x)dx$$
$$= \int h(x)\rho_\theta(k|x)\pi(x)dx$$

Par ailleurs, nous avons déjà observé que la somme des poids d'importance avant normalisation est un estimateur de la constante de normalisation de π (cf. discussion après l'équation 6). Les équations (8) constituent donc des approximations obtenues par échantillonnage préférentiel auto-normalisé des relations exactes de mise à jour définies par (7).

4 Expérimentations

Considérons tout d'abord un exemple jouet en dimension deux où la loi cible est représentée sur la figure 1.

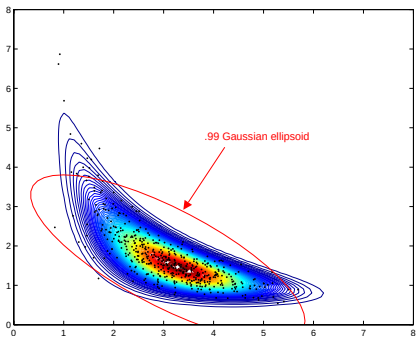

FIG. 1 – Loi cible π

Méthode	Perp. norm.	Var. X_1	Var. X_2
MC (sous π)	1	0.7	0.4
RW-MH (cov. opt.)	-	9.9	10.3
EP uniforme	0.08	7.3	3.2
EP adapt. $d = 1$	0.81	3.5	4.7
EP adapt. $d = 3$	0.98	0.9	0.7

TAB. 1 – Performance de différentes méthodes en terme de perplexité normalisée et de variance pour les fonctions de projections sur chacun des deux axes. MC : Monte Carlo direct sous la loi π ; RW-MH : Algorithme de Metropolis Hastings à marche aléatoire gaussien avec covariance de la loi de proposition optimisée ; EP : échantillonnage préférentiel, les deux dernières lignes correspondent à des lois instrumentales déterminées avec l'algorithme proposé.

Dans ce cas très simple, on constate que l'algorithme proposé permet d'obtenir une loi instrumentale sous la forme d'un mélange de $d = 3$ gaussiennes telle que la perplexité normalisée soit très proche de 1. Pour cette loi instrumentale, les poids d'importance ont donc peu de variabilité et, sans surprise, les variance pour les deux fonctions coordonnées sont assez proche de ce que l'on obtiendrait en utilisant directement la méthode de Monte Carlo (sans poids) en simulant sous la loi π. La variance de simulation obtenue est en particulier bien plus faible que celle de l'algorithme MCMC de type Metropolis-Hastings où la matrice de covariance de la loi de proposition a été ajustée à partir de la covariance a posteriori estimée (représentée en trait plein sur la figure 1) en suivant les arguments de [9]. Pour des lois cibles de grande dimension, les résultats de *scaling* discutés dans [9] suggèrent que la variance de l'estimation obtenue par l'algorithme MCMC (toujours pour

les fonctions coordonnées) augmente proportionnellement à la dimension du problème. Même pour un problème bidimensionnel, on constate que l'augmentation de variance due aux corrélations dans la chaîne peut être significative (ici d'un facteur 5 à 20 ; on trouverait un facteur 8 pour une cible gaussienne bidimensionnelle).

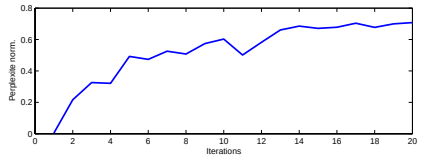

FIG. 2 – Perplexité normalisée en fonctions du nombre d'itérations ($n = 10000$).

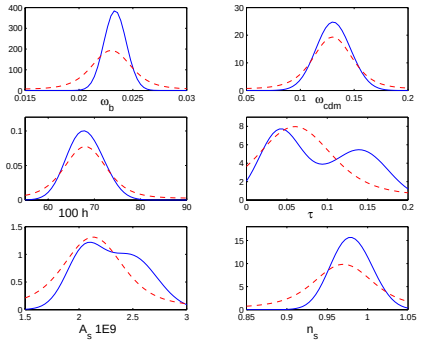

FIG. 3 – Marginales de la loi instrumentale avant (tirets) et après (trait plein) adaptation.

Dans le cadre du projet ANR ECOSSTAT (*Exploration du modèle cosmologique par fusion statistique de grands relevés hétérogènes*, en partenariat avec l'Institut d'Astrophysique de Paris et le Laboratoire d'Astrophysique de Marseille) nous nous intéressons aux applications de cette approche pour l'inférence de paramètres cosmologiques. Nous présentons ici simplement un exemple de résultat obtenu à partir de mesures du spectre du fond de rayonnement cosmique (données WMAP1 [10]) et un modèle cosmologique classique à six paramètres. Ici la tâche est plus ardue du fait de la dimensionalité plus élevée de l'espace des paramètres. Il est en particulier difficile de trouver une loi instrumentale initiale conduisant à un estimateur d'échantillonnage préférentiel viable. La figure 2 montre tout de même qu'avec $n = 10000$ simulations à chaque itération, il est possible d'obtenir un résultat assez appréciable en partant d'une perplexité normalisée de 0.004 (c'est à dire d'une loi instrumentale initialement très inefficace). On constate par ailleurs sur la figure 2 que la perplexité normalisée augmente de façon très régulière montrant une bonne approximation des relations exactes de mise à jour des paramètres décrites dans la section 3. La figure 3 n'est malheureusement pas très informative sur le travail d'adaptation effectué par l'algorithme (adaptation pourtant très significative puisqu'on passe d'une perplexité normalisée de 0.004 à 0.71). Une des raisons en est que la loi instrumentale de départ correspond au produit des marginales représentées en tirets sur la figure 3 et que l'adaptation consiste, pour une large part, à apprendre les corrélations entre les variables. On observe néanmoins des effets supplémentaires que l'on aurait manifestement du mal à capturer avec une simple loi de proposition gaussienne multivariée, c'est-à-dire avec $d = 1$ (attention par ailleurs à une mauvaise interprétation de figure 3 : la loi instrumentale déterminée peut ne pas ressembler dans les détails à la la loi cible, sauf si la perplexité normalisée est extrêmement proche de 1, ce qui n'est pas le cas ici).

Dans le cadre de l'inférence de modèles cosmologiques, un intérêt potentiel supplémentaire de l'approche décrite ci-dessus est lié au fait que l'évaluation de la vraisemblance ℓ présente un coût de calcul très élevé (les 20×10000 calculs de vraisemblance nécessaires pour produire les résultats des figures 2–3 correspondent à plusieurs jours de temps de calcul). La possibilité de distribuer les calculs de vraisemblance simultanément sur plusieurs machines avec l'échantillonnage préférentiel constitue donc, dans ce cadre, un avantage très net sur les techniques MCMC. Compte tenu de cet argument ainsi que des gains obtenus (dans des cas simples) en termes de variance d'estimation, cette approche d'échantillonnage préférentiel adaptatif nous semble présenter un fort potentiel pour le type d'applications discutées ici.

Références

[1] C. P. Robert et G. Casella. *Monte Carlo Statistical Methods*. Springer, New York, 2nd edition, 2004.

[2] A. Lewis et S. Bridle. Cosmological parameters from CMB and other data: A Monte-Carlo approach. *Phys. Rev. D*, 66, 2002. astro-ph/0205436.

[3] M. Doran et C. M. Mueller. Analyze this! A cosmological constraint package for CMBEASY. *J. Cosmol. Astropart. Phys.*, 2004. astro-ph/0311311.

[4] O. Cappé, A. Guillin, J. M. Marin et C. P. Robert. Population Monte Carlo. *J. Comput. Graph. Statist.*, 13(4):907–929, 2004.

[5] R. Douc, A. Guillin, J-M. Marin et C. P. Robert. Convergence of adaptive mixtures of importance sampling schemes. *Ann. Statist.*, 35(1), 2007.

[6] R. Douc, A. Guillin, J-M. Marin et C. P. Robert. Minimum variance importance sampling via population Monte Carlo. A paraître dans *ESAIM Probab. Statist.*, 2007.

[7] O. Cappé, E. Moulines et T. Rydén. *Inference in Hidden Markov Models*. Springer, 2005.

[8] N. Slonim et Y. Weiss. Maximum likelihood and the information bottleneck. In *Advances in Neural Information Processing Systems (NIPS)*, 15:335–342, 2003.

[9] G. O. Roberts et J. S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statist. Sci.*, 16(4):351–367, 2001.

[10] G. Hinshaw *et al.* First year wilkinson microwave anisotropy probe (WMAP) observations: Angular power spectrum. *Astrophys. J. Suppl.*, 148, 2003. astro-ph/0302217.