On the Convergence of Monte Carlo Approximations to the Posterior Density

By
Charles J. Geyer[1] and Luke Tierney[2]
Technical Report No. 579
School of Statistics
University of Minnesota
June 1, 1992

[1]Research supported in part by grant DMS-9007833 from the National Science Foundation
[2]Research supported in part by grant DMS-9005858 from the National Science Foundation

Abstract

The Monte Carlo approximation of the posterior density using a mixture of complete data posteriors proposed by Tanner and Wong (1987) and Gelfand and Smith (1990) converges almost surely and in L^1 to the exact posterior. The coverages of level sets of the approximate posterior (highest posterior density regions) converge simultaneously in the Lévy metric to the exact coverages, as do the Monte Carlo approximations of coverages proposed by Wei and Tanner (1990). Some results are also given for problems in which the complete data likelihood must be calculated by Monte Carlo.

A method of calculating posterior densities in missing data problems from Monte Carlo simulations was proposed by Tanner and Wong (1987) and Gelfand and Smith (1990). It is assumed that we can calculate exactly the posterior density $p(\theta|x,y)$ of the parameter θ given both the observed data x and the missing data y, but the posterior density $p(\theta|x)$ of the parameter given only the observed data is analytically intractable (both are densities with respect to some σ-finite measure μ on the parameter space). Using some version of the Metropolis-Hastings algorithm (Metropolis, et al., 1953; Hastings, 1970) it is possible to generate samples (y_i, θ_i), $i = 1, 2, \ldots$ forming an irreducible Markov chain whose equilibrium distribution is the joint distribution of the parameter and the missing data given the observed data. The samples $\theta_1, \theta_2, \ldots$ of parameter values have the distribution of interest with density $p(\theta|x)$ and could be used for calculations about it, but Gelfand and Smith point out that it is better to use

$$h_n(\theta) = p_n(\theta|x) = \frac{1}{n}\sum_{i=1}^{n} p(\theta|x, y_i) \tag{1}$$

as an approximation of the exact posterior density

$$h(\theta) = p(\theta|x) = \int p(\theta|x,y)\, dP(y|x). \tag{2}$$

They show that (1) has smaller (pointwise) mean squared error than a kernel density estimator based on the θ_i. A more precise version of this result is given by Liu, Wong, and Kong (1991). This note discusses various senses in which (1) converges to (2).

Theorem 1 *If the complete data posterior $p(\theta|x,y)$ is a jointly measurable function of θ and y, then for almost all sample paths of the Markov chain, the Monte Carlo approximation (1) converges to the exact posterior density (2) almost everywhere $[\mu]$ and in L^1.*

PROOF. Let ω denote a point in the probability space of the Markov chain. Then (1) can be rewritten

$$h_n(\theta)(\omega) = \frac{1}{n}\sum_{i=1}^{n} p\Big(\theta \;\Big|\; x, y_i(\omega)\Big).$$

to explicitly show the dependence on ω. Note that h_n is a jointly measurable function of θ and ω. Let Q be the probability measure governing the Markov chain and

$$\begin{aligned} A &= \{\, (\theta,\omega) : h_n(\theta)(\omega) \not\to h(\theta) \,\} \\ A_\theta &= \{\, \omega : h_n(\theta)(\omega) \not\to h(\theta) \,\} \\ A_\omega &= \{\, \theta : h_n(\theta)(\omega) \not\to h(\theta) \,\} \end{aligned}$$

Then by ergodicity of the Markov chain $Q(A_\theta) = 0$ for each θ, so $\iint_A dQ\, d\mu = 0$ by Tonelli's theorem; hence, by another application of Tonelli, $\mu(A_\omega) = 0$, for almost all ω $[Q]$, which says that for almost all sample paths of the Markov chain $h_n(\theta) \to h(\theta)$ for almost all θ $[\mu]$, which is the first assertion of the theorem. That $\int |h_n - h|\, d\mu \to 0$ follows immediately by Scheffé's theorem. □

Level sets of the posterior, also called highest posterior density regions, are sets of the form

$$S_\gamma = \{\theta : h(\theta) \geq \gamma\}.$$

These are approximated by the level sets

$$S_{n,\gamma} = \{\theta : h_n(\theta) \geq \gamma\}$$

of the approximate posterior. Let H denote the posterior probability distribution: $H(A) = \int_A h\, d\mu$, and define the functions $F(\gamma) = 1 - H(S_\gamma)$ and $F_n(\gamma) = 1 - H(S_{n,\gamma})$. These would be distribution functions except that they are not necessarily right continuous at jumps. Still, we can use the Lévy distance to measure convergence of F_n to F, though technically it is a pseudometric here because of the lack of right continuity. F_n converges to F in the Lévy pseudometric if for every $\epsilon > 0$ there is an n_ϵ such that

$$F(\gamma - \epsilon) - \epsilon \leq F(\gamma) \leq F_n(\gamma + \epsilon) + \epsilon, \qquad \gamma \geq 0,\ n \geq n_\epsilon. \tag{3}$$

Wei and Tanner (1990) proposed calculating the posterior probability of the level sets using the simulations $\theta_1, \theta_2, \ldots$ of the parameter values. Let

$$H_n(A) = \frac{1}{n} \sum_{i=1}^{n} 1_{[\theta_i \in A]} \tag{4}$$

be the empirical approximation to the posterior distribution H, and let $F_{n,n}(\gamma) = 1 - H_n(S_{n,\gamma})$.

Theorem 2 *Under the joint measurability condition of Theorem 1, for almost all sample paths of the Markov chain, both F_n and $F_{n,n}$ converge to F in the Lévy pseudometric.*

PROOF. By Theorem 1, $h_n/h \to 1$ at almost every point where h is nonzero, hence for almost all θ $[H]$. Thus by Egoroff's theorem the convergence is almost uniform: for every $\epsilon > 0$ there is a set B_ϵ such that $H(B_\epsilon) < \epsilon$ and $h_n/h \to 1$ uniformly on the complement of B_ϵ. Hence there is an n_ϵ such that for $n \geq n_\epsilon$

$$(1 - \epsilon)h(\theta) \leq h_n(\theta) \leq (1 + \epsilon)h(\theta)$$

for $\theta \notin B_\epsilon$. This implies

$$S_{\gamma/(1-\epsilon)} \setminus B_\epsilon \subset S_{n,\gamma} \subset S_{\gamma/(1+\epsilon)} \cup B_\epsilon \tag{5}$$

for all γ, which in turn implies

$$H(S_{\gamma/(1-\epsilon)}) - \epsilon \leq H(S_{n,\gamma}) \leq H(S_{\gamma/(1+\epsilon)}) + \epsilon \tag{6}$$

which implies (3).

From (5) it follows that (6) holds with H replaced by H_n. Let $G_n(\gamma) = 1 - H_n(S_\gamma)$, then G_n converges uniformly to F except at jumps of F by the Glivenko-Cantelli theorem (which does not require independence, just ergodicity). Hence for

some $m_\epsilon \geq n_\epsilon$ we have $|G_n(\gamma) - F(\gamma)| \leq \epsilon$ for all γ that are not jumps of F and for all $n \geq m_\epsilon$. This implies

$$F\left(\frac{\gamma}{1+\epsilon}\right) - 2\epsilon \leq F_{n,n}(\gamma) \leq F\left(\frac{\gamma}{1-\epsilon}\right) + 2\epsilon$$

for all γ and ϵ such that $\gamma/(1 \pm \epsilon)$ is not a jump of F. This implies the convergence of $F_{n,n}$ to F. □

Wei and Tanner (1990) recommend calculating the αth quantile of $h_n(\theta_i)$, $i = 1, \ldots, m$ (call it $\gamma_{n,\alpha}$) and taking the level set $S_{n,\gamma_{n,\alpha}}$ to be the Monte Carlo approximation of the highest posterior density region with coverage $1 - \alpha$. This is reasonable if there is a γ such that $F(\gamma) = \alpha$, and F does not have a jump at γ. Otherwise, the coverage of this approximate level set may converge to anything between $F(\gamma-)$ and $F(\gamma+)$, so the coverage should be interpreted with caution. The exact sense in which the Monte Carlo approximations to the coverage of level sets converge is given by Theorem 2. Eventually $H(S_\gamma)$ is well approximated by $H_n(S_{n,\beta})$ for some β near γ but not necessarily well approximated by $H_n(S_{n,\gamma})$.

Smith (1992) suggested using Monte Carlo likelihood (Geyer and Thompson, 1992; Geyer, 1992) to construct posteriors in problems in which even the complete data likelihood is analytically intractable. This method can be extended to apply to the missing data case (Gelfand and Carlin, 1991; Geyer, 1992), but since the complete data likelihood is analytically intractable, the methods discussed above still do not apply.

These methods produce a Monte Carlo approximation $l_n(\theta)$ to the likelihood that converges pointwise almost surely to the exact likelihood $l(\theta)$. The Fubini trick of Theorem 1 gives convergence of l_n to l almost everywhere for almost all sample paths of the Monte Carlo, but now this does not buy much. The difficulty is that if $\pi(\theta)$ is a prior, $l_n(\theta)\pi(\theta)$ need not be integrable. Since Scheffé's theorem no longer applies, there is, in general, no guarantee that the integrals of $l_n(\theta)\pi(\theta)$ converge to integrals of $l(\theta)\pi(\theta)$.

Under mild continuity conditions (Geyer, 1992) it is true that l_n converges to l uniformly on compact sets. This assures that for almost all sample paths of the Monte Carlo and any compact set K

$$\int_K l_n(\theta)\pi(\theta)\, d\mu(\theta) \to \int_K l(\theta)\pi(\theta)\, d\mu(\theta)$$

by dominated convergence. Hence if we restrict our attention to some large compact set K everything is all right. In practice this may not make much difference, but it is unsatisfactory from a theoretical point of view.

This leaves us with the following scheme. Run a Metropolis-Hastings algorithm (on the sample space) to calculate a Monte Carlo approximation l_n to the likelihood. Choose a large compact set K. Run a second Metropolis-Hastings algorithm (on the parameter space) producing a Markov chain $\theta_1, \theta_2, \ldots$ whose equilibrium distribution is is concentrated on K and has density proportional to $l_n(\theta)\pi(\theta)$. Then

$$h_{m,n}(\theta) = \frac{l_n(\theta)\pi(\theta)}{\frac{1}{m}\sum_{i=1}^m l_n(\theta_i)\pi(\theta_i)}$$

approximates the posterior density conditioned on K, and the fraction of $h_{m,n}(\theta_i)$, $i = 1, \ldots, m$ that exceed some level γ approximates the content of the highest posterior density region above γ. In special cases, it seems that one should be able to control the tails of the Monte Carlo likelihood well enough to dispense with the compact set K, but it does not seem possible to do this in general.

References

Gelfand, A. E. and Carlin, B. P. (1991) Maximum likelihood estimation for constrained or missing data models. Research Report 91–002, Division of Biostatistics, University of Minnesota.

Gelfand, A. E. and Smith A. F. M. (1990) Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.*, **85**, 398–409.

Geyer, C. J. (1992) On the convergence of Monte Carlo maximum likelihood calculations. Technical Report No. 571, School of Statistics, University of Minnesota.

Geyer, C. J. and Thompson, E. A. (1992) Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. Roy. Statist. Soc. Ser. B*, to be published.

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Liu, J., Wong, W. H. and Kong A. (1991) Correlation structure and convergence rate of the gibbs sampler (I): application to the comparisons of estimators and augmentation schemes. Technical Report No. 299, Department of Statistics, University of Chicago.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953) Equation of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1092.

Smith, A. F. M. (1992) Discussion of the paper by Geyer and Thompson. *J. Roy. Statist. Soc. Ser. B*, to be published.

Tanner, M. A. and Wong, W. H. (1987) The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.*, **82**, 528–550.

Wei, G. C. G. and Tanner, M. A. (1990) Calculating the content and the boundary of the highest posterior density region via data augmentation. *Biometrika*, **77**, 649–652.