An Adaptive Monte Carlo Metropolis-Hastings Algorithm for Bayesian inference of spatial autologistic models

Faming Liang

January 3, 2011

Abstract

The problem of simulating from distributions with intractable normalizing constants has received much attention in recent literature. In this talk, we introduce a new algorithm, the so-called adaptive Monte Carlo Metropolis-Hastings (AMCMH) algorithm, for tackling this problem. At each iteration, AMCMH replaces the unknown normalizing constant ratio by a Monte Carlo estimate which is calculated using all samples generated so far in the run. Under mild conditions, we show that AMCMH is ergodic, and the weak law of large numbers still holds for it. AMCMH represents a new type of adaptive MCMC algorithms for which the stationary distribution is changed from iteration to iteration.

The problem

Suppose we have a dataset X generated from a statistical model with the likelihood function

$$f(x|\theta) = \frac{1}{\kappa(\theta)} \exp\{-U(x,\theta)\}, \quad x \in \mathcal{X},\ \theta \in \Theta, \tag{1}$$

where θ is the parameter vector of the model, and $\kappa(\theta)$ is the normalizing constant which depends on θ and is not available in closed form.

The posterior density of θ is given by

$$\pi(\theta|x) \propto \frac{1}{\kappa(\theta)} \exp\{-U(x,\theta)\}\pi(\theta). \tag{2}$$

How to sample from $\pi(\theta|x)$ puts a great challenge on current statistical methods due to the intractable constant $\kappa(\theta)$.

Maximum Pseudo-likelihood Approach

Besag (1974) proposed to approximate the likelihood function by a tractable pseudo-likelihood function which ignores neighboring dependence of the data.

The method is easy to use, but it typically performs less satisfactory for the models with strong neighboring dependence.

Besag, J.E. (1974). Spatial interaction and the statistical analysis of lattice systems. JRSS-B, 36, 192-236.

Monte Carlo MLE

Geyer and Thompson (1992) proposed to approximate $\kappa(\theta)$ using an importance sampling approach.
Let θ^* denote an initial guess of θ, and let $y_1, \ldots, y_m$ denote random samples simulated from $f(y|\theta^*)$. Then

$$\log f_m(x|\theta) = -U(x,\theta) - \log(\kappa(\theta^*)) - \log\left(\frac{1}{m}\sum_{i=1}^{m} \exp\{U(y_i,\theta^*) - U(y_i,\theta)\}\right), \quad (3)$$

will approach to $\log f(x|\theta)$ as $m \to \infty$.

The estimator $\hat{\theta} = \arg\max_\theta \log f_m(x|\theta)$ is called the MCMLE of θ.

Geyer, C. and Thompson, E. (1992), "Constrained Monte Carlo Maximum Likelihood for Dependent Data," JRSS-B, ***54****, 657-699.*

Exchange Algorithm

1. Propose a candidate point ϑ from a proposal distribution denoted by $q(\vartheta|\theta, x)$.
2. Generate an auxiliary variable $y \sim f(y|\vartheta)$ using a perfect sampler (Propp and Wilson, 1996).
3. Accept ϑ with probability $\min\{1, r(\theta, \vartheta|x)\}$, where

$$r(\theta, \vartheta|x) = \frac{\pi(\vartheta)f(x|\vartheta)f(y|\theta)q(\theta|\vartheta, x)}{\pi(\theta)f(x|\theta)f(y|\vartheta)q(\vartheta|\theta, x)}.$$

Møller et al. (2006), "An Efficient Markov chain Monte Carlo Method for Distributions with Intractable Normalizing Constants," Biometrika, **93***, 451-458.*
Murray et al. (2006), "MCMC for Doubly-Intractable Distributions," Proc. 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI).

MCMH Algorithm

Let θ_t denote the current draw of θ, and let $y_1^{(t)}, \ldots, y_m^{(t)}$ denote the auxiliary samples simulated from the distribution $f(y|\theta_t)$, which can be drawn by either a MCMC algorithm or an automated rejection sampling algorithm.

1. Draw ϑ from some proposal distribution $Q(\theta_t, \vartheta)$.
2. Estimate the normalizing constant ratio $R(\theta_t, \vartheta) = \kappa(\vartheta)/\kappa(\theta_t)$ by

$$\widehat{R}_m(\theta_t, \mathbf{y}_t, \vartheta) = \frac{1}{m} \sum_{i=1}^{m} \frac{g(y_i^{(t)}, \vartheta)}{g(y_i^{(t)}, \theta_t)},$$

where $g(y, \theta) = \exp\{-U(y, \theta)\} = C\dot{f}(y|\theta)$, and $\mathbf{y}_t = (y_1^{(t)}, \ldots, y_m^{(t)})$ denotes the collection of auxiliary samples.

Liang, F. and Jin, I.H. (2010) A Monte Carlo Metropolis-Hastings Algorithm for Sampling from Distributions with Intractable Normalizing Constants. Submitted Manuscript.

MCMH Algorithm (continue)

3. Calculate the Monte Carlo MH ratio

$$\tilde{r}_m(\theta_t, \mathbf{y}_t, \vartheta) = \frac{1}{\widehat{R}_m(\theta_t, \mathbf{y}_t, \vartheta)} \frac{g(x, \vartheta)\pi(\vartheta)}{g(x, \theta_t)\pi(\theta_t)} \frac{Q(\vartheta, \theta_t)}{Q(\theta_t, \vartheta)},$$

   where $\pi(\theta)$ denotes the prior distribution imposed on θ.
4. Set $\theta_{t+1} = \vartheta$ with probability $\tilde{\alpha}(\theta_t, \mathbf{y}_t, \vartheta) = \min\{1, \tilde{r}_m(\theta_t, \mathbf{y}_t, \vartheta)\}$, and set $\theta_{t+1} = \theta_t$ with the remaining probability.
5. If the proposal is rejected in step 4, set $\mathbf{y}_{t+1} = \mathbf{y}_t$. Otherwise, draw samples $\mathbf{y}_{t+1} = (y_1^{(t+1)}, \ldots, y_m^{(t+1)})$ from $f(y|\theta_{t+1})$ using either a MCMC algorithm or an automated rejection sampling algorithm.

AMCMH: Algorithm Setting

- θ_t: the current draw of θ;
- $y_t = (y_1^{(t)}, \ldots, y_m^{(t)})$: a collection of m auxiliary samples simulated from the distribution $f(y|\theta_t)$, which can be drawn by either a MCMC algorithm or an automated rejection sampling algorithm;
- S_t: the set of all distinct samples of θ drawn by iteration t.

Liang, F. and Song, Q.(2010) An adaptive Monte Carlo MH algorithm for Bayesian inference of spatial autologistic models. Submitted Manuscript.

AMCMH: Algorithm

1. Draw ϑ from some proposal distribution $Q(\theta_t, \vartheta)$.
2. Estimate the normalizing constant ratio $R(\theta_t, \vartheta) = \kappa(\theta_t)/\kappa(\vartheta)$ by

$$\widehat{R}(\theta_t, \vartheta) = \frac{1}{m_0 + m_0 \sum_{\theta_i \in S_t \setminus \{\theta_t\}} I(\|\vartheta - \theta_i\| \leq \eta)} \left\{ \sum_{\theta_i \in S_t \setminus \{\theta_t\}} \left[I(\|\vartheta - \theta_i\| \leq \eta) \sum_{j=1}^{m_0} \frac{g(z_j^{(i)}, \theta_t)}{g(z_j^{(i)}, \vartheta)} \right] + \sum_{j=1}^{m_0} \frac{g(z_j^{(t)}, \theta_t)}{g(z_j^{(t)}, \vartheta)} \right\}, \quad (4)$$

where η is a pre-specified threshold value which defines a neighborhood region of ϑ, $g(z, \theta) = \exp\{-U(z, \theta)\}$, and $(z_1^{(i)}, \ldots, z_{m_0}^{(i)})$ denotes a subset of importance samples drawn from the set $y_i = (y_1^{(i)}, \ldots, y_m^{(i)})$ with each being drawn with a probability proportional to $g(z, \vartheta)/g(z, \theta_i)$.

AMCMH: Algorithm (continue)

3. Calculate the Monte Carlo MH ratio

$$\tilde{r}(\theta_t, \vartheta) = \widehat{R}(\theta_t, \vartheta) \frac{g(x, \vartheta)\pi(\vartheta)}{g(x, \theta_t)\pi(\theta_t)} \frac{Q(\vartheta, \theta_t)}{Q(\theta_t, \vartheta)}.$$

   Set $\theta_{t+1} = \vartheta$ with probability $\tilde{\alpha}(\theta_t, \vartheta) = 1 \wedge \tilde{r}(\theta_t, \vartheta)$, and set $\theta_{t+1} = \theta_t$ with the remaining probability, where $a \wedge b = \min(a, b)$.

4. If the proposal is accepted in step 4, set $S_{t+1} = S_t \cup \{\vartheta\}$ and draw samples $y_{t+1} = (y_1^{(t+1)}, \ldots, y_m^{(t+1)})$ from $f(y|\theta_{t+1})$ using either a MCMC algorithm or an automated rejection sampling algorithm. Otherwise, set $\theta_{t+1} = \theta_t$, $y_{t+1} = y_t$, and $S_{t+1} = S_t$.

AMCMH: On generation of auxiliary samples

AMCMH requires the auxiliary samples to be drawn at equilibrium, if a MCMC algorithm is used for generating the auxiliary samples.

To ensure this requirement to be satisfied, the initial auxiliary sample can be generated at each iteration through an importance resampling procedure; that is, set $y_0^{(t+1)} = y_i^{(t)}$ with a probability proportional to its importance weight

$$w_i = g(y_i^{(t)}, \theta_{t+1})/g(y_i^{(t)}, \theta_t). \tag{5}$$

As long as $y_0^{(t+1)}$ follows correctly from $f(y|\theta_{t+1})$, this procedure ensures that all samples, y_{t+1}, y_{t+2}, y_{t+3}, $\ldots$, drawn in the followed iterations will follow correctly from the respective distributions, provided that θ does not change dramatically at each iteration.

Ergodicity

Theorem 1. Consider an adaptive MCMC algorithm with transition kernels P_{γ_k}, $k = 0, 1, 2, \ldots$, on the state space $(\mathcal{X}, \mathcal{F})$. The adaptive algorithm is ergodic if the following conditions are satisfied:

(a) (Stationarity) There exists a stationary distribution $\pi_{\gamma_k}(\cdot)$ for each transition kernel P_{γ_k}.

(b) (Asymptotic Simultaneous Uniform Ergodicity) For any $\epsilon > 0$, there exists $K > 0$ and $N > 0$ such that

$$\|P^n_{\gamma_k}(x, \cdot) - \pi(\cdot)\| \leqslant \epsilon, \qquad \text{for all } x \in \mathcal{X} \text{ and } k > K, n > N.$$

(c) (Diminishing Adaptation) $\lim_{k \to 0} D_k = 0$ in probability, where

$$D_k = \sup_{x \in \mathcal{X}} \|P_{\Gamma_{k+1}}(x, \cdot) - P_{\Gamma_k}(x, \cdot)\|$$

is a $\mathcal{G}_{k+1}$-measurable random variable (depending on the random values Γ_k and Γ_{k+1}).

Weal Law of Large Numbers

Theorem 2. Consider an adaptive MCMC algorithm with transition kernels P_{γ_k}, $k = 0, 1, 2, \ldots$, on the state space $(\mathcal{X}, \mathcal{F})$. Suppose that conditions (a), (b) and (c) of Theorem 1 hold and that all kernels P_{γ_k} converge uniformly to their respective stationary distributions. Let $g(\cdot)$ be a bounded measurable function. Then, for any starting values $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$, conditional on $X_0 = x$ and $\Gamma_0 = \gamma$, we have

$$\frac{\sum_{i=1}^n g(X_i)}{n} \to \pi(g)$$

in probability as $n \to \infty$, where $\pi(g) = \int_{\mathcal{X}} g(x)\pi(dx)$.

Theorems 1 and 2 can be proved using the coupling approach in a similar way to

Roberts and Rosenthal (2007) J. Appl. Prob.

Assumptions

Consider the MH transition kernel

$$P(\theta, \vartheta) = \alpha(\theta, \vartheta) Q(\theta, \vartheta) + \mathcal{I}(\theta \in d\vartheta)[1 - \int_{\Theta} \alpha(\theta, \vartheta) Q(\theta, \vartheta) d\vartheta],$$

which is induced by the proposal $Q(\cdot, \cdot)$ under the assumption that $R(\theta, \vartheta)$ is analytically available.

(A_1) P is irreducible and aperiodic, and admits the posterior $\pi(\theta|x)$ as its stationary distribution.

(A_2) There exists a large constant $M > 1$ such that

$$\sup_{(\theta,\vartheta)\in\Theta\times\Theta} \frac{f(x|\theta)\pi(\theta)}{Q(\vartheta, \theta)} \leq M < \infty.$$

(A_3) Both the prior $\pi(\theta)$ and the unnormalized likelihood function $g(x, \theta) = \exp(-U(x, \theta))$ are bounded away from 0 and ∞ for every $\theta \in \Theta$.

Stationarity of AMCMH Kernels

Theorem 3. Assume conditions (A_1), (A_2) and (A_3) hold, then there exists a stationary distribution for each kernel P_{γ_t}.

Asymptotic simultaneous uniform ergodicity

Theorem 4. Consider the adaptive Markov chain induced by the AMCMH algorithm. If the conditions (A_1), (A_2) and (A_3) are satisfied and the drift function of P satisfies $\sup_{\theta \in \Theta} V(\theta) < \infty$, then the kernels $\{P_{\gamma_t}\}$ are asymptotic simultaneous uniform ergodic.

Diminishing adaptation condition

Theorem 5. Consider the adaptive Markov chain induced by the AMCMH algorithm. If the conditions (A_1), (A_2) and (A_3) are satisfied, then the transition kernels $\{P_{\gamma_t}\}$ satisfy the diminishing adaptation condition.

Convergence of AMCMH Algorithm

Theorem 6. Consider the AMCMH algorithm. If Conditions (A_1), (A_2) and (A_3) are satisfied, then the following results hold:

(i) The algorithm is ergodic with respect to the posterior distribution $\pi(\theta|x)$.

(ii) For a bounded measurable function $g(\cdot)$, as $n \to \infty$,

$$\frac{\sum_{i=1}^{n} g(\theta_i)}{n} \longrightarrow \int g(\theta)\pi(\theta|x)d\theta \quad \text{in probability.}$$

Spatial Autologistic Model

Let $\mathbf{x} = \{x_i : i \in D\}$ denote the observed binary data, where D is the set of indices of the spins. Let $n(i)$ denote the set of neighbors of spin i. The likelihood function of the model is given by

$$f(\mathbf{x}|\alpha, \beta) = \frac{1}{Z(\alpha, \beta)} \exp\left\{ \alpha \sum_{i \in D} x_i + \frac{\beta}{2} \sum_{i \in D} x_i \Big(\sum_{j \in n(i)} x_j \Big) \right\}, \quad (\alpha, \beta) \in \Theta, \qquad (6)$$

where the parameter α determines the overall proportion of $x_i = +1$, the parameter β determines the intensity of interaction between x_i and its neighbors. An exact evaluation of $Z(\alpha, \beta)$ is prohibited even for a moderate system.
For Bayesian analysis, a uniform prior

$$(\alpha, \beta) \in \Theta = [-1, 1] \times [0, 1]$$

is assumed for the model.

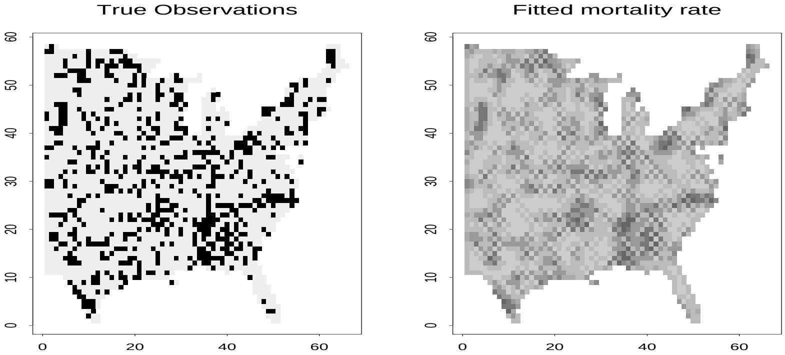

Figure: US cancer mortality data. Left: The mortality map of liver and gallbladder cancers (including bile ducts) for white males during the decade 1950-1959. Black squares denote counties of high cancer mortality rate, and white squares denote counties of low cancer mortality rate. Right: Fitted cancer mortality rates by the spatial autologistic model with the parameters being replaced by its AMCMH estimates.

Table: Computational results for the U.S. cancer mortality data. CPU: The CPU time cost by a single run on a 3.0GHz personal computer. The numbers in the parentheses denote the standard error of the estimates.

Algorithm	Setting	$\widehat{\alpha}$	$\widehat{\beta}$	CPU(s)
AMCMH	$m = 20$	-0.3017 (7.4×10^{-4})	0.1232 (4.0×10^{-4})	5.0
	$m = 50$	-0.3019 (7.4×10^{-4})	0.1228 (3.8×10^{-4})	10.2
	$m = 100$	-0.3017 (6.6×10^{-4})	0.1228 (3.6×10^{-4})	22.5
Exchange	—	-0.3013 (7.7×10^{-4})	0.1231 (4.6×10^{-4})	13.1

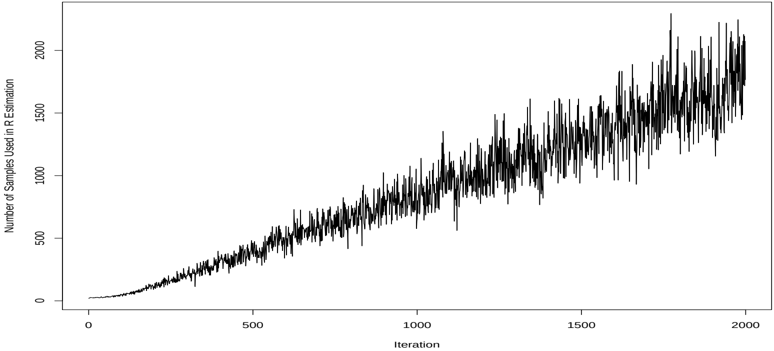

Figure: Average number of samples used in estimation of the normalizing constant ratio versus iterations. The average is taken over 50 runs with $m = 50$ and $m_0 = 20$.

Table: Results for the simulated U.S. cancer mortality data.

	AMCMH			Exchange algorithm		
(α, β)	$\widehat{\alpha}$	$\widehat{\beta}$	CPU(s)	$\widehat{\alpha}$	$\widehat{\beta}$	CPU(s)
(0,0.1)	−.0037 (.0024)	.1003 (.0018)	10.9	−.0042 (.0025)	.1013 (.0019)	13.2
(0,0.2)	−.0025 (.0021)	.2008 (.0018)	9.4	−.0025 (.0020)	.2008 (.0019)	52.3
(0,0.3)	−.0007 (.0014)	.2977 (.0018)	9.5	−.0006 (.0014)	.2973 (.0017)	89.8
(0,0.4)	.0006 (.0011)	.3965 (.0016)	16.4	.0002 (.0006)	.3980 (.0011)	1180.2
(0.1,0.1)	.1035 (.0025)	.0986 (.0022)	10.8	.1035 (.0026)	.0987 (.0022)	12.6
(0.3,0.3)	.3014 (.0099)	.3005 (.0044)	10.6	.3034 (.0099)	.2999 (.0044)	38.2
(0.5,0.5)	.5085 (.0224)	.5016 (.0080)	17.3	— —	— —	—

Discussion

- AMCMH is an adaptive version of the MH algorithm. At each iteration, it replaces the unknown normalizing constant ratio $R(\theta_t, \vartheta)$ by a Monte Carlo estimate calculated using all auxiliary samples generated so far in the simulation. Although it violates the detailed balance condition, it is still ergodic with respect to the desired target distribution and the weak law of large numbers still holds for bounded measurable functions. AMCMH represents a new type of adaptive MCMC algorithms for which the stationary distribution changes from iteration to iteration.

Discussion

- Unlike the auxiliary variable MCMC algorithms, AMCMH avoids the requirement for perfect sampling, and thus can be applied to many statistical models for which perfect sampling is not available or very expensive.
- For estimation of the normalizing constant ratio, only the simple importance sampling method is presented here. Other normalizing constant ratio estimators, such as ratio importance sampling (Chen and Shao, 1997) and bridge sampling (Meng and Wong, 1996), should also work well for AMCMH, although they are only asymptotically unbiased. In particular, the ratio importance sampling estimator relies on the samples generated from a third distribution, other than $f(x|\theta_t)$ and $f(x|\vartheta)$, and can fit well into the framework of AMCMH. The past samples that are close to both θ_t and ϑ can be selected for construction of the "third" distributions.

Acknowledgement

- NSF
- KAUST: The award made by King Abdullah University of Science and Technology (KAUST).