LOCAL WEAK CONSISTENCY OF MARKOV CHAIN MONTE CARLO METHODS WITH APPLICATION TO MIXTURE MODEL

Kamatani, Kengo
Graduate School of Engineering Science, Osaka University

https://doi.org/10.5109/1563534

出版情報：Bulletin of informatics and cybernetics. 45, pp.103-123, 2013-12. 統計科学研究会
バージョン：
権利関係：

LOCAL WEAK CONSISTENCY OF MARKOV CHAIN MONTE CARLO METHODS WITH APPLICATION TO MIXTURE MODEL

by

Kengo Kamatani

Reprinted from the Bulletin of Informatics and Cybernetics
Research Association of Statistical Sciences, Vol.45

FUKUOKA, JAPAN
2013

Bulletin of Informatics and Cybernetics, Vol. 45, 2013

LOCAL WEAK CONSISTENCY OF MARKOV CHAIN MONTE CARLO METHODS WITH APPLICATION TO MIXTURE MODEL

By

Kengo Kamatani*

Abstract

Markov chain Monte Calro methods (MCMC) are commonly used in Bayesian statistics. In the last twenty years, many results have been established for the calculation of the exact convergence rate of MCMC methods. We introduce another rate of convergence for MCMC methods by approximation techniques. This rate can be obtained by the convergence of the Markov chain to a diffusion process. We apply it to a simple mixture model and obtain its convergence rate. Numerical simulations are performed to illustrate the effect of the rate.

Key Words and Phrases: Markov chain Monte Carlo, Asymptotic Normality, Diffusion process.

1. Introduction

Markov chain Monte Carlo (MCMC) method has become an essential tool in any study that has a complicated posterior calculation problem. Various new MCMC methods have been developed in the last two decades. Theoretical support of this strategy has also been developed such as Roberts and Tweedie (1996), Mengersen and Tweedie (1996) and many others. In particular, it was shown that the usual MCMC method produces an ergodic Markov chain (see Tierney (1994) and Roberts and Rosenthal (2004)).

In practice, it is of great interest to study the convergence speed of these Markov chains. Various quantitative bounds have been developed from the spectral approach by such as Diaconis and Stroock (1991) and Diaconis et al. (2008), and from the so-called (double) drift condition approach by Meyn and Tweedie (1994) and Rosenthal (1995), Rosenthal (2002). For an ergodic Markov chain $\{X_m; m \geq 0\}$ on the state space $(E, \mathcal{E})$ with the transition kernel P, they calculated the upper bound of

$$\|\mathcal{L}(X_m) - \Pi\| = 2 \sup_{A \in \mathcal{E}} |\mathbb{P}(X_m \in A) - \Pi(A)| \tag{1}$$

where Π is the invariant distribution and $\mathcal{L}(X_m)$ is the law of X_m. In the former approach, if we can calculate the eigenvalues and the eigenfunctions for $P : L^2(\Pi) \to L^2(\Pi)$, then it is possible to calculate the almost exact bounds. On the other hand, although the latter approach does not provide tight bound, it is relatively easy to apply.

To compare different MCMC methods, the above approaches may have difficulties, since we need to calculate tight (upper and lower) bounds for two or more MCMC

* Graduate School of Engineering Science, Osaka University Machikaneyama-cho 1-3, Toyonaka-si, Osaka, 560-0043, Japan. Email: kamatani@sigmath.es.osaka-u.ac.jp.

methods. However, without calculating such bounds, sometimes it is possible to compare different MCMC methods by the asymptotic variance σ_f^2 in the following limit in $M \to \infty$:

$$\sqrt{M}(I_M - I) \Rightarrow N(0, \sigma_f^2),\ I_M = \frac{1}{M}\sum_{m=0}^{M-1} f(X_m),\ I = \int_E f(x)\Pi(dx).$$

For this comparison, it is sufficient to show positivity of an operator in $L^2(\Pi)$ sense. This approach was studied in Peskun (1973), and later developed by Tierney (1998) and Mira (1998). Although the application area of this is limited, this approach is particularly useful for the comparison of the so-called data-augmentation (DA) procedure with its parameter-expanded extension (see Hobert and Marchev (2008)).

These analysis on MCMC procedures obtain the exact bound of the convergence rate or the exact comparison of MCMC procedures. We took a different approach in Kamatani (2013). Usually, MCMC procedures are complicated that prevent us from exact analysis. On the other hand, by using approximation theory, such as the traditional large sample theory, sometimes it is easy to perform theoretical comparison among MCMC procedures. For this approximation, we introduce an index n, which tends to ∞. As $n \to \infty$, if the following holds for any $M_n \to \infty$, the MCMC procedure is said to have the consistency in Kamatani (2013):

$$I^n_{M_n} - I^n = o_{\mathbb{P}}(1),\ I^n_M = \frac{1}{M}\sum_{m=0}^{M-1} f(X^n_m),\ I^n = \int_E f(x)\Pi_n(dx) \tag{2}$$

where $\{X^n_m; m \geq 0\}$ are sequences of Π_n-invariant Markov chains generated by MCMC procedures. By Theorem 1 of Kamatani (2013), under some regularity conditions, the DA procedure satisfies this property. In practice, if an MCMC procedure has the consistency, it works fairly well. On the other hand, many popular MCMC methods do not satisfy this good convergence property but satisfy a bad property

$$I^n_M - I^n_1 = o_{\mathbb{P}_n}(1) \tag{3}$$

for any fixed $M \in \mathbb{N}$. This property means that the Monte Carlo estimation using M iteration is no more helpful than that using only one iteration. Therefore we can classify MCMC procedure into two categories (2) or (3). Although these two categories do not cover all of the cases, this classification is useful in practice. However it does not tell the rate of convergence.

In this paper, we introduce a further step of this approach. As mentioned earlier, the rate of convergence is useful to predict sufficient number of iteration until convergence, or to compare different MCMC procedures in details. We call r_n the order of the weak consistency if (2) is satisfied for any $M_n \to \infty$ such that $M_n/r_n \to \infty$. If the MCMC procedure has the consistency, we can take $r_n = 1$. On the other hand, the order can be high if the performance of MCMC procedure is poor, that is, the condition (3) is satisfied. The order r_n can be interpreted as the order of the sufficient number of iteration.

As an example we will consider the DA procedure for a simple mixture model $pN(\epsilon, 1) + (1-p)N(0, 1)$ for unknown p but for known ϵ. Since the performance of the DA procedure heavily depends on the parameter ϵ, we let $\epsilon \to 0$ to illustrate the effect. This DA procedure works quite poorly if the true model is close to $N(0, 1)$. The index

n is the sample size. It has the order $r_n = \epsilon^{-1} n^{1/2}$ and this shows the effects of both ϵ and the sample size n. This result comes from the fact that the trajectory of the DA procedure tends to a path of the stochastic process defined by

$$dX_t = (\alpha_1 + X_t z - X_t^2 I)dt + \sqrt{2X_t} dW_t \tag{4}$$

where I corresponds to the Fisher information matrix and z corresponds to the scaled maximum likelihood estimator (see Theorem 3.1). It is probably well recognized that the trajectory of poor behaved MCMC procedure looks like a path of a diffusion process. This result is the first validation for this observation.

This paper is organized as follows. In Section 2 we define (local) weak consistency. In Section 3 we apply this to the simple mixture model. Numerical results is provided in Section 4 which shows the effect of the order of the weak consistency.

2. Local weak consistency of MCMC

We write $[x]$ for the integer part of $x \in \mathbb{R}$.

2.1. Definition of local weak consistency

In this section, we review the (local) consistency and degeneracy and also, we define the order of the weak consistency. Let $\Theta = \mathbb{R}^d$ be a parameter space. Suppose that observation x_n is an element of a set X_n, and we are interested in the approximation of the posterior distribution $P(d\theta|x_n)$. We assume Bernestein von-Mises's theorem, that is, for some $c_n \to \infty$ and some $\mathbb{R}^d$-valued random variable $u_n(x_n)$ such that

$$c_n(\theta - u_n) = O_{\mathbb{P}}(1) \tag{5}$$

where $\theta|x_n \sim P(d\theta|x_n)$. We will consider asymptotic properties of the scaled parameter $c_n(\theta - u_n)$.

MCMC procedure generates a sequence $\theta_\infty = (\theta_0, \theta_1, \ldots)$ such that the law of $\theta_\infty|x_n$ is a Markov chain with the invariant distribution $P(d\theta|x_n)$. We assume stationarity of the process $\theta_\infty|x_n$, that is, the initial guess θ_0 is generated from the posterior distribution. This is impractical setting, but this assumption can be weakened. See Section B for the detail. Let

$$I_M^n(f) = \frac{1}{M}\sum_{m=0}^{M-1} f(c_n(\theta_i - u_n)),\ I^n(f) = \int f(c_n(\theta - u_n))P(d\theta|x_n). \tag{6}$$

We expect that $I_M^n(f)$ is a good approximation of $I^n(f)$. For this in mind, we define local consistency.

Definition 2.1 Local consistency. MCMC procedure is said to have the local consistency if $I_{M_n}^n(f) - I^n(f) = o_{\mathbb{P}}(1)$ for any continuous, bounded function f and for any $M_n \to \infty$.

The MCMC procedure does not always work well. We also define a property of this inefficient behavior. Essentially, the good behavior, local consistency, and the bad behavior, local degeneracy defined below are exclusive (see Kamatani (2011)).

DEFINITION 2.2 LOCAL DEGENERACY. MCMC procedure is said to have the local degeneracy if $I_M^n(f) - I_1^n(f) = o_{\mathbb{P}}(1)$ for any continuous, bounded function f and for any $M \in \mathbb{N}$.

As a measure of poor behavior, local degeneracy is sometimes too wide and so we define a kind of order of convergence among degenerate MCMC procedures.

DEFINITION 2.3 LOCAL WEAK CONSISTENCY. MCMC procedure is said to have the local weak consistency, if $I_{M_n}^n(f) - I^n(f) = o_{\mathbb{P}}(1)$ for any continuous, bounded function f and for any M_n such that $M_n/r_n \to \infty$. We call r_n the order of the local weak consistency.

We can interpret M_n as the sufficient number of iterations for good approximation. Therefore if r_n is large, the MCMC procedure requires many iterations to have a good result. Under the local consistency, we can take $r_n = 1$. We can compare different algorithms by this order r_n.

2.2. Useful lemma

Let $r_n \to \infty$. For each $n = 1, 2, \ldots$, consider a $\mathbb{R}^d$-valued semi-Markov process $\{\theta_t^n; t \geq 0\}$, which jumps at $r_n^{-1}, 2r_n^{-1}, \ldots$ on a probability space $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$.

LEMMA 2.4. *Assume that the embedded Markov chain $\{\theta_{m/r_n}^n; m = 0, 1, \ldots\}$ is a stationary Markov chain with the invariant distribution P_n. If $\{\theta_t^n; t \geq 0\}$ converge in law to a stationary ergodic process, then*

$$\frac{1}{T_n}\int_{[0,T_n)} f(\theta_t^n)dt - \int_{\mathbb{R}^d} f(\theta)P_n(d\theta) = o_{\mathbb{P}_n}(1) \tag{7}$$

for any $T_n \to \infty$ for any bounded and continuous function f.

PROOF. We omit the superscript n from θ_t^n. Proof is almost the same as that of Lemma 2 of Kamatani (2013). Without loss of generality, we can assume $\sup_{x\in\mathbb{R}^d} |f(x)| \leq 1$. Write I_{T_n} and I^n for the first and the second terms in the left-hand side of (7), respectively, and write $I_{i,S}$ for $\frac{1}{S}\int_{t\in[0,S)} f(\theta_{iS+t})dt$. Then

$$I_T = \frac{S}{T}\sum_{i=0}^{[T/S]-1} I_{i,S} + \frac{1}{T}\int_{[S[T/S],T)} f(\theta_t)dt.$$

Note that $I_{i,S}$ $(i = 0, 1, \ldots)$ is not identically distributed in general. If we take $S_n = ([Sr_n]+1)/r_n$, then I_{i,S_n} $(i = 0, 1, \ldots)$ have the same law under $\mathbb{P}_n$. Hence as in Lemma 2 of Kamatani (2013), we have

$$\begin{aligned}\mathbb{E}_n[|I_T - I^n|] &\leq \frac{S_n}{T}\Big[\frac{T}{S_n}\Big]\mathbb{E}_n[|I_{0,S_n} - I^n|] + 2\frac{T - S_n[T/S_n]}{T} \\ &\leq \frac{S_n}{T}\Big[\frac{T}{S_n}\Big]\Big\{\frac{S}{S_n}\mathbb{E}_n[|I_{0,S} - I^n|] + 2\frac{S_n - S}{S_n}\Big\} + 2\frac{T - S_n[T/S_n]}{T}\end{aligned}$$

where in the second inequality, we used $I_{0,S_n} = (S/S_n)I_{0,S} + \int_S^{S_n} f(\theta_t)dt/S_n$. As $T = T_n \to \infty$, the second and the third terms in the left-hand side vanishes, and

$S_n/T[T/S_n]S/S_n \to 1$ for any fixed S. Hence if $\mathbb{E}_n[|I_{0,S} - I^n|]$ can be arbitrary small, the claim follows.

Let P be the limit of P_n and write $I = \int_{\mathbb{R}^d} f(\theta)P(d\theta)$. Then $\mathbb{E}_n[|I_{0,S} - I^n|] \to \mathbb{E}[|I_{0,S} - I|]$ where $\mathbb{E}$ is the expectation with respect to the limit probability measure $\mathbb{P}$. Since θ_t is ergodic under $\mathbb{P}$, this value tends to 0 as $S \to \infty$. Hence the claim follows.

3. Application to mixture model

Let $F_t(dx) = f_t(x)dx$ be probability measures on $(E, \mathcal{E})$ with parameter $t \geq 0$, and write $F = F_0$ and $f = f_0$. We assume that f_t is always strictly positive. Consider the following simple mixture model:

$$P^\epsilon(dx|\theta) = p^\epsilon_\theta(x)dx = (1-\theta)F(dx) + \theta F_\epsilon(dx). \tag{8}$$

MCMC procedures for general k-component mixture model have been developed to perform better posterior inference. See monographs such as Robert and Casella (2004) and Frühwirth-Schnatter (2006). It is well known that for general k-component mixture model, the posterior distribution is multi-modal, and if these peeks are close, then the posterior inference becomes difficult due to the so-called label-switching problem (see Stephens (2000), Marin et al. (2005) and Jasra et al. (2005)). We address here a separate issue. In fact, under such a situation, another problem, local degeneracy occurs. We illustrate this effect by using the order of the weak consistency.

For this reason, we assume over-parametrized situation, that is, the observation $x_n = \{x^1, \ldots, x^n\}$ are independent draw from a one-component model F. We will show that if two components F_0 and F_ϵ are close, the performance becomes even worse. To illustrate the effect, we let $\epsilon = \epsilon_n \to 0$. Write $r_n = \epsilon^{-1}n^{1/2}$ and $c_n = \epsilon n^{1/2}$. There is an obvious relation $r_n c_n = n$. As in p902 of Gassiat (2002) we assume the following regularity condition. Write $L^2(F)$ for the set of F-square integrable functions with norm $\|\cdot\|$ defined by $\|g\|^2 = \int g(x)^2 F(dx)$.

Assumption 3.1 *There exists $d \in L^2(F)$ such that $r_\epsilon(x) := f_\epsilon(x)/f(x) - 1 - \epsilon d(x) = o(\epsilon)$ in $L^2(F)$. Moreover, $\|d\|^2 = I \neq 0$. The prior distribution is assumed to be* Beta(α_1, α_0) *for $\alpha_0, \alpha_1 > 0$.*

Under the assumption, $\int d(x)F(dx) = 0$ because $|\int d(x)F(dx)| = \epsilon^{-1}|\int r_\epsilon(x)F(dx)| \leq \epsilon^{-1}\|r_\epsilon\| \to 0$. One step of the DA procedure is

$$\begin{cases} y^i|x^i, \theta \sim \text{Bernoulli}(p_i),\ p_i = \frac{\theta f_\epsilon(x^i)}{(1-\theta)f(x^i) + \theta f_\epsilon(x^i)}\ (i = 1, \ldots, n), \\ \theta|x_n, y_n \sim \text{Beta}(\alpha_1 + n_1, \alpha_0 + n - n_1) \end{cases} \tag{9}$$

where $y_n = \{y^1, \ldots, y^n\}$ and n_1 is the number of heads in $y_1, \ldots, y_n$. For this model, it is natural to take state space scaling as

$$\theta \mapsto c_n\theta \tag{10}$$

where we take $u_n \equiv 0$ in (5). We will discuss the local consistency and the local degeneracy under this localization.

The DA output $\theta_0, \theta_1, \ldots$ behaves poorly, and the sequence converges to the following diffusion process after the state space scaling in (10) with suitable time scaling.

Let

$$dX_t = (\alpha_1 + X_t z - X_t^2 I)dt + \sqrt{2X_t}dW_t;\ X_0|z \sim P^*(dx|z) \propto \exp(xz - x^2 I/2)x^{\alpha_1 - 1}dx \tag{11}$$

where $\{W_t; t \geq 0\}$ is the standard Brownian motion independent of $z \sim N(0, I)$. Write the law of $\{X_t; t \geq 0\}$ given z by $\mathcal{L}(\{X_t; t \geq 0\}|z)$. For each z, there is a weak solution $\mathcal{L}(\{X_t; t \geq 0\}|z)$ which is ergodic with invariant measure $P^*(dx|z)$ (see Theorem 2.3 in Bibby et al. (2005). See also Section 5.5 of Karatzas and Shreve (1991)). By convergence to this process, we obtain the following.

THEOREM 3.1. *Suppose that $x^1, \ldots, x^n$ are independent sample from F. Under Assumption 3.1, the DA procedure has the local weak consistency with the order $r_n = \epsilon^{-1} n^{1/2}$ if $c_n = \epsilon n^{1/2} \to \infty$.*

PROOF. Let $\{\theta_m^n; m = 0, 1, \ldots\}$ be the stationary Markov chain generated by (9). Let $X_t^n = c_n \theta_{[r_n t]}^n$. Then by Theorem A6, $\mathcal{L}(\{X_t^n; t \geq 0\}|x_n)$ tends to $\mathcal{L}(\{X_t; t \geq 0\}|z)$ in distribution, where $z \sim N(0, I)$. As mentioned above, $\mathcal{L}(\{X_t; t \geq 0\}|z)$ is stationary and ergodic. Together with the separability of the Skorohod topology, and Skorohod's representation theorem (see Theorem 6.7 of Billingsley (1999)), there is a probability space such that $\mathcal{L}(\{X_t^n; t \geq 0\}|x_n) \to \mathcal{L}(\{X_t; t \geq 0\}|z)$. Hence by Lemma 2.4, for any bounded continuous function f on $[0, \infty)$,

$$\frac{1}{T_n}\int_{[0,T_n)} f(X_t^n)dt - \int_{[0,1]} f(c_n\theta)P(d\theta|x_n) = o_{\mathbb{P}_n}(1)$$

for any $T_n \to \infty$, where $P(d\theta|x_n)$ is the posterior distribution, which is the invariant distribution of X_t^n. Take $M_n/r_n = T_n$ and rewrite

$$\frac{1}{T_n}\int_{[0,T_n)} f(X_t^n)dt = \frac{1}{M_n}\sum_{m=0}^{M_n-1} f(c_n\theta_m).$$

Then the convergence of probability in the above means weak consistency of the DA procedure on the order r_n.

Although this result for the large sample scaling limit is new, the scaling limit to a diffusion process have been studied in other directions by Gelfand and Mitter (1991) for small variance asymptotics, and by Roberts et al. (1997) for high-dimensional small variance asymptotics. In particular, the latter approach is still very active. See a recent review by Sherlock et al. (2010) and the references their in. It is worth mentioning that we can apply the local weak consistency for these results.

4. Numerical results

4.1. Metropolis-Hastings procedure

To illustrate poor performance of the DA procedure, we consider a simple independent type Metropolis-Hastings (IMH) procedure as an alternative and compare it with the DA procedure. Note that we prepare this IMH procedure just for comparison and may work well only for this simple mixture model. However related methods may work well for general mixture model and this direction will be studied in elsewhere.

We briefly review the IMH procedure. If we want to approximate probability distribution P on $(E, \mathcal{E})$, we prepare the so-called proposal distribution Q, such that there exists a Radon-Nikodým derivative $dP/dQ = h(x)$. Then IMH procedure iterates the following; Suppose that we have the current value $\theta \in E$. Then

$$\textbf{simulate } \theta^* \sim Q, \textbf{ set } \theta \leftarrow \begin{cases} \theta^* \text{ with probability } \alpha(\theta, \theta^*) \\ \theta \text{ with probability } 1 - \alpha(\theta, \theta^*) \end{cases},$$

where $\alpha(\theta, \theta^*) = \min\{1, h(\theta^*)/h(\theta)\}$. This iteration resulted in a Markov chain $\theta_0, \theta_1, \ldots$ with the invariant distribution P. Hence if it is ergodic, we obtain an approximation of P without simulation from P.

Now we apply this IMH procedure to the simple mixture model. The key is the choice of the proposal distribution. Set $Q^\epsilon(dx|\theta)$ which is close to $P^\epsilon(dx|\theta)$ in such as the Kullback-Leibler distance. Calculate the posterior distribution $Q(d\theta|x_n)$ for observation $x_n = \{x^1, \ldots, x^n\}$ with the model $Q^\epsilon(dx|\theta)$. We use $Q(d\theta|x_n)$ as the proposal distribution.

Next section, we will consider $F_t = N(t, 1)$. For this, take $Q^\epsilon(dx|\theta) = N(\epsilon\theta, 1)$ with the uniform prior. Then $Q(d\theta|x_n) = N(\sum_{i=1}^n x^i/n\epsilon, 1/n\epsilon^2)$ truncated to $[0, 1]$. It is not difficult to check the local consistency of this IMH procedure. See Section C for the details.

4.2. Simulation

We compare the DA and the IMH procedures through numerical simulations. Consider the normal mixture model $F_\epsilon(dx) = N(\epsilon, 1)$. To illustrate the difference of the DA and the IMH procedures, first we plot the trajectories of θ_m $(m = 0, 1, 2, \ldots)$ under fairly large sample size $n = 10^4$ with the true model $N(0, 1)$ and $\epsilon = 0.5$. Unlike the IMH procedure, the trajectory from the DA procedure behaves like a stochastic diffusion process (Figure 1) and this is true by Theorem A6. By Theorem 3.1, the order of the weak consistency is $r_n = \epsilon^{-1} n^{1/2} = 200$ for the DA procedure but it is $r_n = 1$ for the IMH procedure.

Next we check the effect of n, ϵ and the underlying true model. To illustrate the differences of the performance, we plot empirical autocorrelations. If an MCMC procedure has poor mixing property, empirical autocorrelation does not converges to 0 quickly. First we check the effect of n for $\epsilon = 0.5$ by the different sample sizes $50, 250$ and 1250. Orders of the weak consistency of the DA procedure are $r_n = 14.14\ldots, 31.62\ldots$ and $70.71\ldots$. Recall that r_n corresponds to the number of iteration for good convergence, and so we take the window size as $\max\{25, 2r_n\}$. As the sample size becomes larger, the mixing property of the DA procedure becomes worse, as the empirical autocorrelations suggest (Figure 2).

Similarly, next we check the effect of ϵ for $n = 250$ by the different values $\epsilon = 0.1, 0.5$ and 1. Orders of the weak consistency of the DA procedure are $r_n = 158.11\ldots, 31.62\ldots$ and $15.81\ldots$ (Figure 3).

Finally, we check the effect of the underlying model. We only checked the behaviors of MCMC procedures under the true model $N(0, 1)$. Now we check those for $pN(0, 1) + (1-p)N(1, 1)$ for $p = 0$, 0.25 and 0.5. However as Lemma A1 and Le Cam's third lemma suggest (see the comment after Lemma A1), the effect of the difference of the underlying true model may be small under the assumption $\epsilon n^{1/2} \to \infty$ (Figure 4).

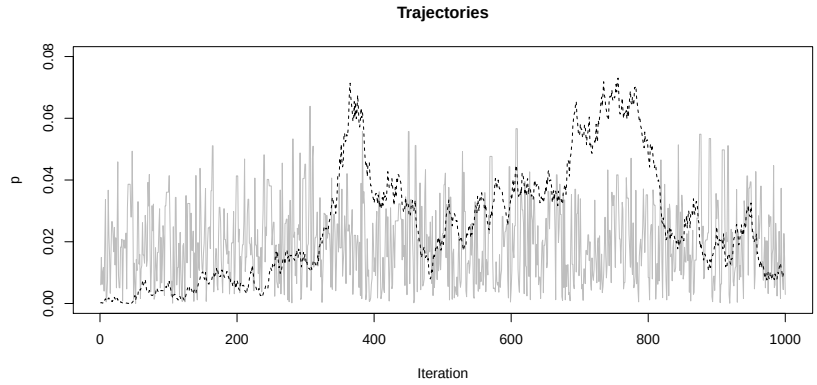

Figure 1: Trajectories of MCMC procedure for $n = 10^4$. The dashed line is the trajectory from the DA procedure and the solid gray line is that of the IMH procedure.

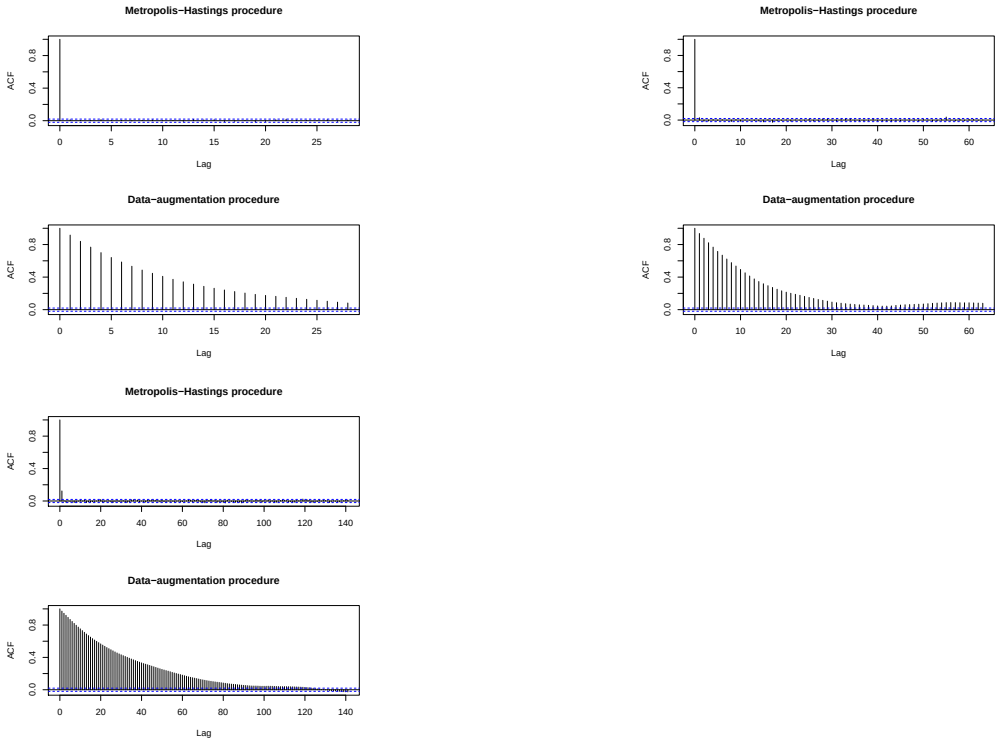

Figure 2: Empirical autocorrelations for $n = 50$ (left), 250 (middle), and 1250 (right), for the IMH procedure (top) and the DA procedure (bottom).

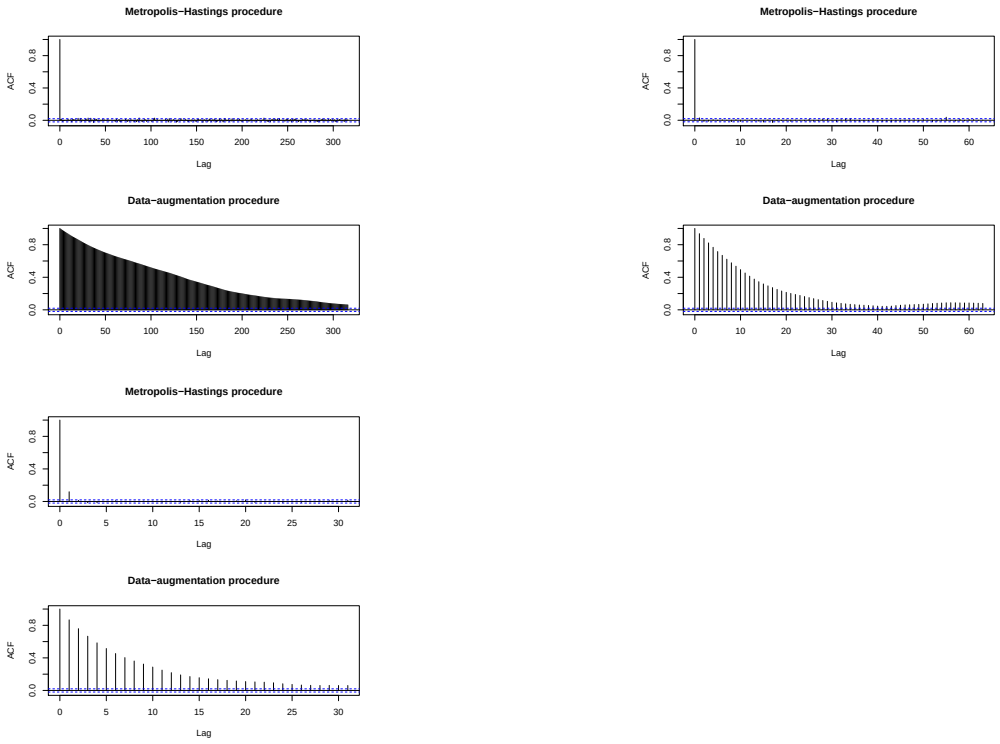

Figure 3: Empirical autocorrelations for $\epsilon = 0.1$ (left), 0.5 (middle), and 1 (right) with the sample size $n = 250$, for the IMH procedure (top) and the DA procedure (bottom).

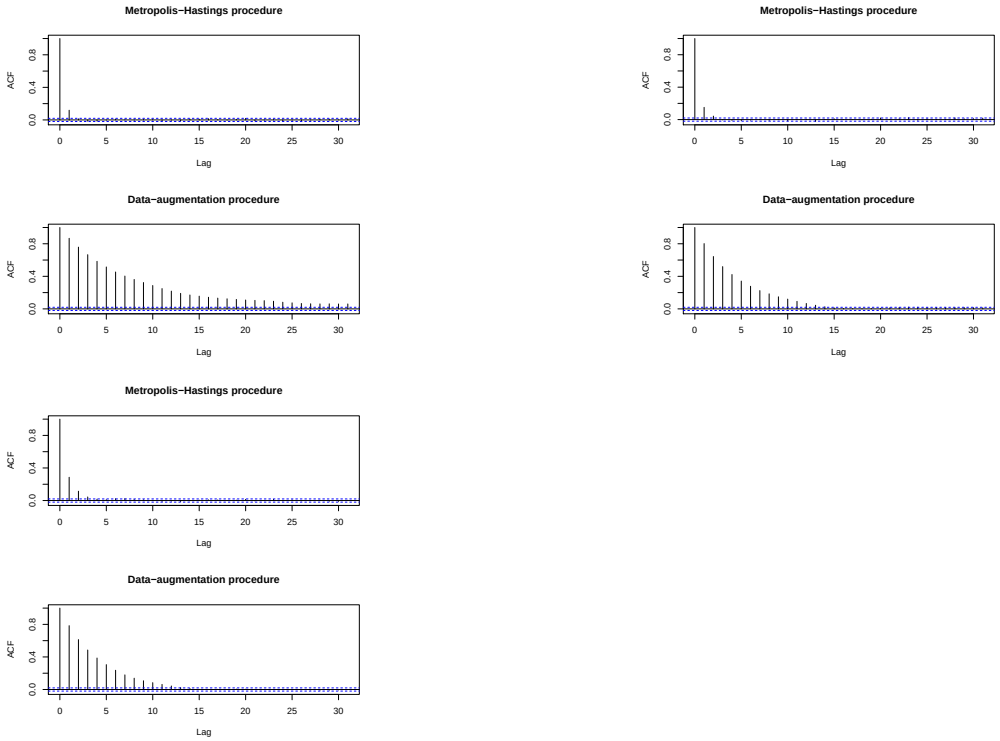

Figure 4: Empirical autocorrelations for $p = 0$ (left), 0.25 (middle), and 0.5 (right) with the sample size $n = 250$, and $\epsilon = 1$ for the IMH procedure (top) and the DA procedure (bottom).

4.3. Discussion

In this article we presented a definition of the order of the weak consistency, and applied it to a simple mixture model. The simulation results suggest that this order is a good measure of the convergence. It remains true that the verification of the weak consistency for complicated model is technically difficult. However estimation of r_n is not difficult if the corresponding Markov chain convergence to a diffusion process such as (4). Such a convergence is probably true for more general MCMC methods. For example, the DA procedure for a simple probit model converges to the Ornstein-Uhlenbeck process with the rate $r_n = n$. With this in mind, r_n can be estimated by an empirical estimate of

$$\frac{\mathbb{V}_n[\theta|x_n]}{\mathbb{V}_n[\theta_1^n - \theta_0^n|x_n]} \approx r_n$$

where the numerator is the variance of the posterior distribution, and the denominator is the variance of $\theta_1^n - \theta_0^n$, where $\{\theta_m^n; m = 0, 1, \ldots\}$ is an output of an MCMC procedure. A similar statistic was proposed in Roberts and Rosenthal (1998). This can be used as a measure of efficiency of MCMC methods.

Acknowledgement

I am deeply grateful to Professor Nakahiro Yoshida, for encouragement and valuable suggestions. We also thank the referee for their useful comments and suggestions. This work was supported in part by Grant-in-Aid for JSPS Fellows (19-3140) and Grant-in-Aid for Young Scientists (B) 22740055.

A Limit results

A1. Some properties of simple mixture model

We define

$$t_g = \frac{g - f}{f},\ l_n(g) = \sum_{i=1}^{n} \log g(x^i).$$

For $g = p_\theta^\epsilon$, we have $t_{p_\theta^\epsilon} = \theta t_{f_\epsilon}$ and $t_{f_\epsilon} = \epsilon d + r_\epsilon$. Note that $\int g(x)F(dx) = \int r_\epsilon(x)F(dx) = 0$.

LEMMA A1 LOCAL ASYMPTOTIC NORMALITY. *Under Assumption 3.1, for any $H > 0$,*

$$\sup_{h\in[0,H]} |l_n(p^\epsilon_{h/c_n}) - l_n(f) - hn^{-1/2}\sum_{i=1}^{n} d(x^i) + h^2 I/2| = o_{\mathbb{P}_n}(1). \tag{12}$$

PROOF. Set R to be $\log(1 + x) = x - x^2/2 + x^2 R(x)$, where $R(x) \to 0$ $(x \to 0)$. As in the proof of Theorem 3.1 of Gassiat (2002) we have

$$l_n(g) - l_n(f) = \sum_{i=1}^{n} t_g(x^i) - \frac{1}{2}\sum_{i=1}^{n} t_g(x^i)^2 + \sum_{i=1}^{n} t_g(x^i)^2 R(t_g(x^i)).$$

For $g = p^{\epsilon}_{h/c_n}$, we have $t_g = hn^{-1/2}(d + r_\epsilon/\epsilon)$. Hence the first term of the right-hand side is

$$hn^{-1/2}\sum_{i=1}^{n} d(x^i) + hn^{-1/2}\epsilon^{-1}\sum_{i=1}^{n} r_\epsilon(x^i) = hn^{-1/2}\sum_{i=1}^{n} d(x^i) + o_{\mathbb{P}_n}(1)$$

uniformly in $h \in [0, H]$. By similar arguments, $\sup_{h\in[0,H]} |\sum_{i=1}^{n} t_{p^{\epsilon}_{h/c_n}}(x^i)^2 - h^2 I| = o_{\mathbb{P}_n}(1)$ in probability. For the third term,

$$\sup_{h\in[0,H]} \max_{i=1,\dots,n} |t_{p^{\epsilon}_{h/c_n}}(x^i)| \le H \max_{i=1,\dots,n} |c_n^{-1} t_{f_\epsilon}(x^i)|.$$

However $\max_{i=1,\dots,n} |c_n^{-1} t_{f_\epsilon}(x^i)| = o_{\mathbb{P}_n}(1)$ by the inequality $\mathbb{P}(\max_{i=1,\dots,n} |X_i| > c) \le c^{-2}\sum_{i=1}^{n} \mathbb{E}[|X_i|^2, |X_i| > c)]$. Hence the third term is negligible. By these, the convergence (12) follows.

Let $P_n(dx_n) = \prod_{i=1}^{n} F_0(dx^i)$ and $Q_n(dx_n) = \prod_{i=1}^{n} P^{\epsilon}_{h/c_n}(dx^i)$ where $x_n = \{x^1, \dots, x^n\}$. Then P_n and Q_n are mutually contiguous by Le Cam's third lemma. Therefore if a statement is true in P_n-probability, then it is also true in Q_n-probability, and vice versa. Thus Theorem 3.1 holds even if the observation is an i.i.d. draw from P^{ϵ}_{h/c_n} for any fixed $h > 0$.

LEMMA A2 CONSISTENCY OF THE POSTERIOR DISTRIBUTION. *Under Assumption 3.1, for any $M_n \to \infty$,*

$$\int_{M_n/c_n}^{1} P(d\theta|x_n) = o_{\mathbb{P}_n}(1).$$

PROOF. Let $P(d\theta) = \mathrm{Beta}(\alpha_1, \alpha_0)$. By definition, for $M, H > 0$,

$$\begin{aligned}\int_{M/c_n}^{1} P(d\theta|x_n) &= \frac{\int_{M/c_n}^{1} \exp(l_n(p^{\epsilon}_\theta) - l_n(f))P(d\theta)}{\int_0^1 \exp(l_n(p^{\epsilon}_\theta) - l_n(f))P(d\theta)} \le \frac{\int_{M/c_n}^{1} \exp(l_n(p^{\epsilon}_\theta) - l_n(f))P(d\theta)}{\int_0^{H/c_n} \exp(l_n(p^{\epsilon}_\theta) - l_n(f))P(d\theta)} \\ &\le \frac{\int_{M/c_n}^{1} \exp(l_n(p^{\epsilon}_\theta) - l_n(f))P(d\theta)}{\inf_{\theta\in[0,H/c_n]} \exp(l_n(p^{\epsilon}_\theta) - l_n(f)) \int_0^{H/c_n} P(d\theta)} \qquad (13)\end{aligned}$$

By Lemma A1, the infimum in the denominator is bounded below from 0 in probability, and $\int_0^{H/c_n} P(d\theta)$ is on the order of $c_n^{-\alpha_1}$. Therefore, to prove the claim, it is sufficient to show that the numerator is $o_{\mathbb{P}_n}(c_n^{-\alpha_1})$ as $M = M_n \to \infty$.

As in the proof of Inequality 1.2 of Gassiat (2002), since $\log(1 + x) \le x - (x_-)^2/2$ where $x_- = \min\{0, x\}$, we have

$$l_n(p^{\epsilon}_\theta) - l_n(f) \le \theta\sum_{i=1}^{n} t_{f_\epsilon}(x^i) - \frac{\theta^2}{2}\sum_{i=1}^{n} (t_{f_\epsilon})^2_-(x^i).$$

Note that $|(t_{f_\epsilon})_- - \epsilon d_-| \le |r_\epsilon|$. Since $P(d\theta) \le C\theta^{\alpha_1 - 1}d\theta$ for some constant $C > 0$, up to this constant, an upper bound of the numerator of (13) is

$$\int_{M/c_n}^{\infty} \exp(\theta\sum_{i=1}^{n} t_{f_\epsilon}(x^i) - \frac{\theta^2}{2}\sum_{i=1}^{n} (t_{f_\epsilon})^2_-(x^i))\theta^{\alpha_1 - 1}d\theta = c_n^{-\alpha_1}\int_M^{\infty} \exp(hA_n - \frac{h^2}{2}B_n)h^{\alpha_1 - 1}dh$$

where $A_n := c_n^{-1} t_{f_\epsilon}(x^i) \Rightarrow N(0, I)$ and $B_n := \sum_{i=1}^n c_n^{-2} (t_{f_\epsilon})_-^2(x^i) \to \|d_-\|^2$ in probability. Hence as $M = M_n \to \infty$, this is on the order of $o_{\mathbb{P}_n}(c_n^{-\alpha_1})$ and hence the claim follows.

Let $P^*(d\theta|z)$ be a probability measure on $[0, \infty)$ such that

$$P^*(d\theta|z) \propto \exp(\theta z - \theta^2 I/2)\theta^{\alpha_1 - 1} d\theta \tag{14}$$

where $z \in \mathbb{R}$. The following is a Bernstein-von Mises theorem. We omit this proof since it come from exact the same way as the proof of the usual Bernstein-von Mises theorem for regular parametric family. See for example, p142 of van der Vaart (1998). In the following, $\|\cdot\|$ is the total variation distance defined in (1).

PROPOSITION A3 BERNSTEIN-VON MISES THEOREM. *Under Assumption 3.1,*

$$\|P^*(d\theta|x_n) - P^*(d\theta|z = n^{-1/2} \sum_{i=1}^n d(x^i))\| = o_{\mathbb{P}_n}(1)$$

where $P^*(d\theta|x_n)$ *is defined by* $\int_{[0,H]} P^*(d\theta|x_n) = \int_{[0,H/c_n]} P(d\theta|x_n)$.

A2. Convergence to a diffusion process

Let $\theta_0^n, \theta_1^n, \ldots$ be an output of the DA procedure for given observation x_n. In this section, we will show that $\{h_m^n; m = 0, \ldots, \}$ has a diffusion limit where $h_m^n = c_n \theta_m^n$. Note that by using t_g, $\mathbb{P}(y^i = 1|\theta, x^i)$ and $\mathbb{P}(y^i = 0|\theta, x^i)$ become

$$\frac{\theta f_\epsilon(x^i)}{p_\theta^\epsilon(x^i)} = \frac{\theta + t_{p_\theta^\epsilon}(x^i)}{1 + t_{p_\theta^\epsilon}(x^i)}, \quad \frac{(1-\theta) f(x^i)}{p_\theta^\epsilon(x^i)} = \frac{1-\theta}{1 + t_{p_\theta^\epsilon}(x^i)}, \tag{15}$$

with respectively. Throughout in this section, we set $\theta = h/c_n$ and consider uniform convergence property with respect to h. We use the both notation θ and h. Write $X_\theta = o_{\mathbb{P}_n}(\delta_n)$ or $X_\theta = O_{\mathbb{P}_n}(\delta_n)$ for some sequence $\delta_n > 0$ if $\sup_{\theta \in [0, H/c_n]} |X_\theta|$ is $o_{\mathbb{P}_n}(\delta_n)$ or $O_{\mathbb{P}_n}(\delta_n)$ for any $H > 0$ with respectively.

LEMMA A4. *Let* $n_1 = \sum_{i=1}^n y^i$ *such that* $y^i|\theta, x^i \sim P(dy|\theta = h/c_n, x^i)$. *Under Assumption 3.1,*

$$\mathbb{E}_n[n_1|\theta, x_n] - n\theta = h n^{-1/2} \sum_{i=1}^n d(x^i) - h^2 I + o_{\mathbb{P}_n}(1), \tag{16}$$

$$r_n^{-1} \mathbb{V}_n[n_1|\theta, x_n] = h + o_{\mathbb{P}_n}(1). \tag{17}$$

In particular, both $\mathbb{E}_n[n_1|\theta, x_n]$ *and* $\mathbb{V}_n[n_1|\theta, x_n]$ *are* $O_{\mathbb{P}_n}(r_n)$. *Moreover, for* $k = 3, 4$,

$$\mathbb{E}_n[(n_1 - \mathbb{E}_n[n_1|\theta, x_n])^k|\theta, x_n] = O_{\mathbb{P}_n}(r_n^{k-2}). \tag{18}$$

PROOF. Note

$$\frac{\theta + x}{1 + x} = \theta + (1 - \theta)\Big\{x - x^2 + \frac{x^3}{1 + x}\Big\}.$$

Thus, by (15),

$$\mathbb{E}_n[n_1|\theta,x_n]=\sum_{i=1}^n\frac{\theta+t_{p_\theta^\epsilon}(x^i)}{1+t_{p_\theta^\epsilon}(x^i)}=n\theta+(1-\theta)\{\sum_{i=1}^n t_{p_\theta^\epsilon}(x^i)-\sum_{i=1}^n t^2_{p_\theta^\epsilon}(x^i)+\sum_{i=1}^n\frac{t^3_{p_\theta^\epsilon}(x^i)}{1+t_{p_\theta^\epsilon}(x^i)}\}.$$

We already have a similar expansion in Lemma A1 (take $R(x)=x/(1+x)$ here), and so for $\theta=h/c_n$, the inside of the bracket is

$$hn^{-1/2}\sum_{i=1}^n d(x^i)-h^2I+o_{\mathbb{P}_n}(1).$$

Thus we obtain (16). To show (17), observe

$$\begin{aligned}\mathbb{V}_n[n_1|\theta,x_n] &= \sum_{i=1}^n\frac{\theta+t_{p_\theta^\epsilon}(x^i)}{1+t_{p_\theta^\epsilon}(x^i)}\frac{1-\theta}{1+t_{p_\theta^\epsilon}(x^i)}\\ &= (1-\theta)\sum_{i=1}^n\Big\{\theta+\frac{(1-\theta)t_{p_\theta^\epsilon}(x^i)}{1+t_{p_\theta^\epsilon}(x^i)}\Big\}\Big\{1-\frac{t_{p_\theta^\epsilon}(x^i)}{1+t_{p_\theta^\epsilon}(x^i)}\Big\}.\end{aligned}$$

The reading term of the above is $n(1-\theta)\theta$, and the remaining terms are dominated by

$$|\sum_{i=1}^n\frac{t_{p_\theta^\epsilon}(x^i)}{1+t_{p_\theta^\epsilon}(x^i)}|+\sum_{i=1}^n\frac{t^2_{p_\theta^\epsilon}(x^i)}{(1+t_{p_\theta^\epsilon}(x^i))^2}.$$

For $\theta=h/c_n$, these are $O_{\mathbb{P}_n}(1)$. This is clear by expansion $x/(1+x)=x-x^2/(1+x)$ with $\sum_{i=1}^n t_{p_\theta^\epsilon}(x^i)=O_{\mathbb{P}_n}(1)$ and $\max_i|t_{p_\theta^\epsilon}(x^i)|=o_{\mathbb{P}_n}(1)$. This proves (17).

Next we show (18). If $X^i\sim\text{Bernoulli}(p^i)$, $|\mathbb{E}(X^i-p^i)^k|=|p^i(1-p^i)^k+(-p^i)^k(1-p^i)|\le 2p^i=2\mathbb{E}[X^i]$. Let $S=\sum_{i=1}^n X_i$. Then

$$\begin{aligned}&|\mathbb{E}[(S-\mathbb{E}[S])^3]|=|\textstyle\sum_{i=1}^n\mathbb{E}[(X^i-\mathbb{E}[X^i])^3]|\le 2\mathbb{E}[S],\\ &|\mathbb{E}[(S-\mathbb{E}[S])^4]|=|\textstyle\sum_{i=1}^n\mathbb{E}[(X^i-\mathbb{E}[X^i])^4]|+\sum_{i\neq j}\mathbb{V}[X^i]\mathbb{V}[X^j]\\ &\qquad\qquad\le 2\mathbb{E}[S]+\mathbb{V}[S]^2\le 2\mathbb{E}[S]+4\mathbb{E}[S]^2.\end{aligned}$$

Using this fact,

$$\begin{aligned}&|\mathbb{E}_n[(n_1-\mathbb{E}_n[n_1|\theta,x_n])^3|\theta,x_n]|\le 2\mathbb{E}_n[n_1|\theta,x_n]=O_{\mathbb{P}_n}(r_n),\\ &\mathbb{E}_n[(n_1-\mathbb{E}_n[n_1|\theta,x_n])^4|\theta,x_n]\le 2\mathbb{E}_n[n_1|\theta,x_n]+4\mathbb{E}_n[n_1|\theta,x_n]^2=O_{\mathbb{P}_n}(r_n^2).\end{aligned}$$

PROPOSITION A5 CONVERGENCE OF THE COEFFICIENTS. *Let θ^* be the output of (9) when θ is the previous value. Then under Assumption 3.1,*

$$n\mathbb{E}_n[\theta^*-\theta|\theta,x_n] = \alpha_1+hn^{-1/2}\sum_{i=1}^n d(x^i)-h^2I+o_{\mathbb{P}_n}(1), \tag{19}$$

$$nc_n\mathbb{V}_n[\theta^*|\theta,x_n] = 2h+o_{\mathbb{P}_n}(1) \tag{20}$$

and

$$n^2c_n^2\mathbb{E}_n[(\theta^*-\mathbb{E}_n[\theta^*|\theta,x_n])^4|\theta,x_n]=O_{\mathbb{P}_n}(1). \tag{21}$$

PROOF. Since $\theta^*|x_n, y_n$ follows $\mathrm{Beta}(\alpha_1 + n_1, \alpha_0 + n - n_1)$, we have

$$\begin{aligned}\mathbb{E}_n[\theta^* - \theta|\theta, x_n] &= \mathbb{E}_n[\mathbb{E}_n[\theta^*|x_n, y_n]|\theta, x_n] - \theta = \frac{\alpha_1 + \mathbb{E}_n[n_1|\theta, x_n]}{\alpha_1 + \alpha_0 + n} - \theta \\ &= \Big(1 - \frac{\alpha_1 + \alpha_0}{\alpha_1 + \alpha_0 + n}\Big)\frac{\{\alpha_1 + \mathbb{E}_n[n_1|\theta, x_n] - n\theta\} - \theta(\alpha_1 + \alpha_0)}{n}\end{aligned}$$

and thus (19) follows by (16). Observe that

$$\mathbb{E}_n[(\theta^* - \theta)^2|x_n, y_n] = (\mathbb{E}_n[\theta^*|x_n, y_n] - \theta)^2 + \mathbb{V}_n[\theta^*|x_n, y_n]. \tag{22}$$

Expectation of the first term in the right-hand side is

$$\begin{aligned}\mathbb{E}_n[(\mathbb{E}_n[\theta^*|x_n, y_n] - \theta)^2|\theta, x_n] &= \mathbb{E}_n[(\mathbb{E}_n[\theta^*|x_n, y_n] - \mathbb{E}_n[\theta^*|\theta, x_n])^2|\theta, x_n] \\ &\quad + \mathbb{E}_n[\theta^* - \theta|\theta, x_n]^2 \\ &= \frac{\mathbb{V}_n[n_1|\theta, x_n]}{(\alpha_1 + \alpha_0 + n)^2} + \mathbb{E}_n[\theta^* - \theta|\theta, x_n]^2 \\ &= h/nc_n + O_{\mathbb{P}_n}(1/n^2)\end{aligned}$$

where the last equation comes from (17, 19). Expectation of the second term of the right-hand side of (22) is

$$\begin{aligned}\mathbb{E}_n[\mathbb{V}_n[\theta^*|x_n, y_n]|\theta, x_n] &= \mathbb{E}_n\Big[\frac{(\alpha_1 + n_1)(\alpha_0 + n - n_1)}{(\alpha_1 + \alpha_0 + n)^2(\alpha_1 + \alpha_0 + 1 + n)}|\theta, x_n\Big] \\ &= \mathbb{E}_n\Big[\frac{nn_1 - n_1^2 + O(n)}{n^3 + O(n^2)}|\theta, x_n\Big] \\ &= \frac{\mathbb{E}_n[n_1|\theta, x_n]}{n^2} + O_{\mathbb{P}_n}\Big(\frac{r_n^2}{n^3}\Big) = \frac{h}{nc_n} + O_{\mathbb{P}_n}\Big(\frac{r_n^2}{n^3}\Big), \end{aligned} \tag{23}$$

where in the third equality, we used $\mathbb{E}_n[n_1^2|\theta, x_n] = \mathbb{E}_n[n_1|\theta, x_n]^2 + \mathbb{V}_n[n_1^2|\theta, x_n] = O_{\mathbb{P}_n}(r_n^2)$, and $n^{-2} = o(r_n^2/n^3)$. This proves (20).

Last we show (21). By Jensen's inequality,

$$\mathbb{E}_n[(\theta^* - \mathbb{E}_n[\theta^*|\theta, x_n])^4|\theta, x_n] \le \mathbb{E}_n[\mathbb{E}_n[(\theta^* - \mathbb{E}_n[\theta^*|x_n, y_n])^4|x_n, y_n]|\theta, x_n].$$

Recall that if $X \sim \mathrm{Beta}(\alpha, \beta)$, the kurtosis is

$$\frac{\mathbb{E}[(X - \mathbb{E}[X])^4]}{\mathbb{V}[X]^2} = 3 + 6\frac{(\alpha - \beta)^2(\alpha + \beta + 1) - \alpha\beta(\alpha + \beta + 2)}{\alpha\beta(\alpha + \beta + 2)(\alpha + \beta + 3)} \le 3 + \frac{6}{\min\{\alpha, \beta\}}$$

where in the second inequality, we used simple fact such as $|\alpha - \beta| \le \max\{\alpha, \beta\}$. Since $\theta^*|x_n, y_n \sim \mathrm{Beta}(\alpha_1 + n_1, \alpha_0 + n - n_1)$ and $\min\{\alpha_1 + n_1, \alpha_0 + n - n_1\} \ge \min\{\alpha_1, \alpha_0\} > 0$, we have

$$\mathbb{E}_n[(\theta^* - \mathbb{E}_n[\theta^*|x_n, y_n])^4 \le C\mathbb{V}_n[\theta^*|x_n, y_n]^2$$

for some constant $C > 0$. Hence (21) follows by (23).

Recall that $\theta_0^n, \theta_1^n, \ldots$ are scaled to $h_m^n = c_n\theta_m^n$. Furthermore, we introduce an interpolated process

$$X_t^n = h_{[r_n t]}^n = c_n\theta_{[r_n t]}^n$$

where $[x]$ is the integer part of x. Write $\mathcal{L}(\{X_t^n; t \geq 0\}|x_n)$ for the law of $\{X_t^n; t \geq 0\}$ given x_n. We show that the convergence of $\{X_t^n; t \geq 0\}$ to $\{X_t; t \geq 0\}$. There are many studies for the convergence of Markov chain to Markov process. See Section 4.8 of Ethier and Kurtz (1986) and references therein. We apply Theorem IX.4.21 of Jacod and Shiryaev (2003) that shows the convergence of pure jump Markov process to a diffusion process. Note that $\{X_t^n; t \geq 0\}$ is not a Markov process, since it jumps at deterministic time, but still we can apply the theorem by Proposition VI.6.37 (b) of Jacod and Shiryaev (2003). Indeed, we apply Theorem IX.4.21 of Jacod and Shiryaev (2003) to

$$Y_t^n = h^n_{T_t^n} = c_n \theta^n_{T_t^n}$$

that is a step Markov process, and we will obtain convergence of Y^n to X in the proof of the following theorem. We can check the condition of Proposition VI.6.37 (b) for $\tau_t^n := N^n_{[nt]}$ where N_i^n is the i-th jump time of T^n. Then by this proposition, we obtain the convergence of X^n to X as in the statement of the following theorem.

THEOREM A6 CONVERGENCE OF THE DA PROCEDURE TO A DIFFUSION PROCESS. *By Assumption 3.1, $\mathcal{L}(\{X_t^n; t \geq 0\}|x_n)$ tends to $\mathcal{L}(\{X_t; t \geq 0\}|z)$ in distribution, where $z \sim N(0, I)$.*

PROOF. By Skorohod's representation theorem, we may assume that there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $z_n(\omega) := n^{-1/2} \sum_{i=1}^n d(x^i) \to z(\omega)$ for any $\omega \in \Omega$. Let $b(\xi; z) = \alpha_1 + \xi z - \xi^2 I$ and $c(\xi; z)^2 = 2\xi$, and set

$$\begin{cases} b^n(\xi; x_n) = \lambda_n c_n \mathbb{E}_n[\theta^* - \theta | \theta = c_n\xi, x_n], \ c^n(\xi; x_n) = \lambda_n c_n^2 \mathbb{V}_n[\theta^* | \theta = c_n\xi, x_n]. \\ d^n(\xi; x_n) = \lambda_n c_n^4 \mathbb{E}_n[(\theta^* - \mathbb{E}_n[\theta^* | \theta = c_n\xi, x_n])^4 | \theta = c_n\xi, x_n]. \end{cases}$$

In Proposition A5, we proved that

$$\mathbb{P}(\sup_{|\xi|<H} |b^n(\xi; x_n) - b(\xi; z_n)| + |c^n(\xi; x_n) - c(\xi; z_n)| + |d^n(\xi; x_n)| > \epsilon) \to 0$$

for any $H > 0$, $\epsilon > 0$. Thus by local uniform continuity of $b(\xi; z)$ and $c(\xi; z)$ in (ξ, z), we have $U_n = o_{\mathbb{P}}(1)$ where

$$U_n = \sum_{m=1}^{\infty} 2^{-m} \min\{1, \sup_{|\xi|<m} |b^n(\xi; x_n) - b(\xi; z)| + |c^n(\xi; x_n) - c(\xi; z)| + |d^n(\xi; x_n)|\}.$$

Thus by Theorem IX.4.21 of Jacod and Shiryaev (2003), $\mathcal{L}(\{X_t^n; t \geq 0\}|x_n)$ converges to $\mathcal{L}(\{X_t; t \geq 0\}|z)$ in probability. Indeed, by Skorohod's representation theorem again, we may assume $z_n \to z$ and $U_n \to 0$ on a probability space $(\Omega', \mathcal{F}', \mathbb{P}')$. Then $b^n(\xi; x_n(\omega)) \to b(\xi; z(\omega))$, $c^n(\xi; x_n(\omega)) \to c(\xi; z(\omega))$ and $d^n(\xi; x_n(\omega)) \to 0$, local uniformly in ξ for any $\omega \in \Omega'$.

B Stationarity condition

Let x_n, $\theta_\infty = (\theta_0, \theta_1, \ldots)$ and a probability measure $\mathbb{P}$ be as in Section 2.1. Stationarity of $\theta_\infty | x_n$ implies $\theta_0 | x_n \sim P(d\theta | x_n)$. We relax the condition and replace $P(\cdot|x_n)$ by some probability measure $Q(\cdot|x_n)$. The choice of $Q(\cdot|x_n)$ is of great interest in the literature, and in practice, we usually choose θ_0 close to the "center" (such as MLE or

MAP) of $P(d\theta|x_n)$. Lemma 4 of Kamatani (2013) shows that these choices are reasonable. As an example of $Q(\cdot|x_n)$, we will show that $Q(\cdot|x_n)$ defined in the end of Section 4.1. works well (more precisely, Theorem 3.1 holds) in our model.

Write $\mathbb{Q}$ for a probability measure as the same as $\mathbb{P}$ replacing $\theta_0|x_n \sim P(d\theta|x_n)$ by $\theta_0|x_n \sim Q(d\theta|x_n)$. We call that $\mathbb{Q}$ is contiguous to $\mathbb{P}$ if for any bounded functions $\phi_n(x_n, \theta_\infty)$, $\phi_n(x_n, \theta_\infty) = o_{\mathbb{P}}(1)$ implies $\phi_n(x_n, \theta_\infty) = o_{\mathbb{Q}}(1)$. In particular, if an MCMC procedure has local (weak) consistency in $\mathbb{P}$, then the same property holds in $\mathbb{Q}$. Thanks to Markov property, contiguity holds if for any bounded functions $\psi_n(x_n, \theta)$,

$$\psi_n(x_n, \theta_0) = o_{\mathbb{P}}(1) \Rightarrow \psi_n(x_n, \theta_0) = o_{\mathbb{Q}}(1).$$

To see this, take $\psi_n(x_n, \theta_0) = \mathbb{E}[\phi_n(x_n, \theta_\infty)|x_n, \theta_0]$. Write $\mu \le \nu$ if $\mu(A) \le \nu(A)$ for any measurable set A. Note that $\|\nu - \mu\| = 2 \sup_A |\nu(A) - \mu(A)| \le 2\epsilon$ implies $\mu \le \nu + \epsilon$.

Lemma B1 Lemma 4 of Kamatani (2013). *Suppose that for any $\epsilon > 0$, there exists $c > 0$ such that*

$$\liminf_{n\to\infty} \mathbb{P}(A_{n,\epsilon}) > 1 - \epsilon, \ A_{n,\epsilon} := \{Q(\cdot|x_n) \le cP(\cdot|x_n) + \epsilon\}. \tag{24}$$

Then $\mathbb{Q}$ is contiguous to $\mathbb{P}$.

Indeed, if $\mathbb{P}(\psi_n) = o(1)$ for $[0,1]$-valued function ψ_n, then $\mathbb{Q}(\psi_n) \le \mathbb{P}(A^c_{n,\epsilon}) + c\mathbb{P}(\psi_n) + \epsilon$ and $\limsup_{n\to\infty} \mathbb{Q}(\psi_n) \le 2\epsilon$. Now we apply Lemma B1 to our model. Let $Q^\epsilon(dx|\theta) = F_{\epsilon\theta}$ and write $Q(d\theta|x_n)$ for the posterior distribution for the flat prior distribution as in the end of Section 4.1.

Scale $Q(d\theta|x_n)$ by $\theta \mapsto c_n\theta$ and write it $Q^*(d\theta|x_n)$. As in Lemmas A1-A3, it is not difficult to conclude

$$\|Q^*(d\theta|x_n) - Q^*(d\theta|z_n)\| = o_{\mathbb{P}}(1), \ z_n = n^{-1/2} \sum_{i=1}^n d(x^i) \tag{25}$$

where $Q^*(d\theta|z)$ is a probability measure on $[0, \infty)$ such that

$$Q^*(d\theta|z) \propto \exp(\theta z - \theta^2 I/2) d\theta.$$

Thus for any $\epsilon > 0$, we have

$$\liminf_{n\to\infty} \mathbb{P}(A^1_{n,\epsilon}) = 1, \ A^1_{n,\epsilon} := \{Q^*(\cdot|x_n) \le Q^*(\cdot|z = z_n) + \epsilon/3\}.$$

Choose a compact set $K \subset (0, \infty)$ so that $\liminf \mathbb{P}(z_n \in K) \ge 1 - \epsilon$. If $z \in K$, we can choose $c > 0$ so that $Q^*(d\theta|z) \le cP^*(d\theta|z) + \epsilon/3$ where $P^*(d\theta|z)$ is as in (14). By Proposition A3,

$$\liminf_{n\to\infty} \mathbb{P}(A^2_{n,\epsilon}) = 1, \ A^2_{n,\epsilon} := \{P^*(d\theta|z = z_n) \le P^*(d\theta|x_n) + \epsilon/3c\}.$$

Together with these results, the conclusion holds, since $A_{n,\epsilon} \supset \{z_n \in K\} \cap \bigcup_{i=1,2} A^i_{n,\epsilon}$ and we have

$$\liminf_{n\to\infty} \mathbb{P}(A_{n,\epsilon}) \ge \liminf_{n\to\infty} \left(1 - \mathbb{P}(z_n \in K^c) - \sum_{i=1}^2 \mathbb{P}((A^i_{n,\epsilon})^c)\right) \ge 1 - \epsilon.$$

Thus Theorem 3.1 holds under $\mathbb{Q}$ as claimed.

C Consistency of independent type Metropolis-Hastings procedure

We prove local consistency of the independent type Metropolis-Hastings (IMH) defined in Section 4.1. For probability measures P, Q and a transition kernel K on $\mathbb{R}^d$, we introduce an operator $\otimes$

$$(P \otimes K)(A \times B) = \int_A P(dx)K(x, B),\ (P \otimes Q)(A \times B) = P(A)Q(B).$$

and extend them to probability measures on $\mathbb{R}^{2d}$. We introduce another operator $\wedge$ by

$$(P \wedge Q)(dx) = \min\{p(x), q(x)\}\sigma(dx).$$

if $dP/d\sigma(x) = p(x)$ and $dQ/d\sigma(x) = q(x)$ for a σ-finite measure σ. Let $X^n = \{X_i^n; i \in \mathbb{N}_0\}$ be a Markov chain with the transition kernel K_n with the initial distribution P_n, and let $X = \{X_i; i \in \mathbb{N}_0\}$ be that for the transition kernel K with the initial distribution P.

LEMMA C1 LEMMAS 2 AND 3 OF KAMATANI (2013). *Let K and K_n $(n = 1, 2, \ldots)$ be transition kernels that have the invariant probability distributions P and P_n with respectively. If $\|P_n \otimes K_n - P \otimes K\| \to 0$, then X^n tends to X in law. Moreover, if K is ergodic, then for any bounded continuous function f and any $M_n \to \infty$,*

$$\int f(x)P_n(dx) - M_n^{-1} \sum_{m=0}^{M_n-1} f(X_m^n) = o_{\mathbb{P}}(1).$$

Thus $\|P_n \otimes K_n - P \otimes K\| \to 0$ with ergodicity of K is a set of sufficient condition for (local) consistency. Assume that there exist Radon-Nikodým derivatives $dP/dQ(x) = h(x)$ and $dP_n/dQ_n(x) = h_n(x)$. As in Section 4.1, let

$$\begin{aligned}
K(x, dy) &= Q(dy)\min\{1, h(y)/h(x)\} + R(x)\delta_x, \\
R(x) &= 1 - \int_{y \in E} Q(dy)\min\{1, h(y)/h(x)\}, \\
K_n(x, dy) &= Q_n(dy)\min\{1, h_n(y)/h_n(x)\} + R_n(x)\delta_x, \\
R_n(x) &= 1 - \int_{y \in E} Q_n(dy)\min\{1, h_n(y)/h_n(x)\}.
\end{aligned}$$

where δ_x is the Dirac measure. By definition,

$$(P \otimes K)(dxdy) = (P \otimes Q) \wedge (Q \otimes P)(dxdy) + P(dx)R(x)\delta_x(dy).$$

We also have

$$P(dx)R(x) = P(dx) - (P \otimes Q) \wedge (Q \otimes P)(dx \times E). \tag{26}$$

Note that by Corollary 2 of Tierney (1994), K is ergodic if P and Q is equivalent, that is, absolutely continuous to each other.

LEMMA C2. *Suppose K_n and K has the above form. Assume that P_n and P are absolutely continuous to Q_n and Q with respectively. Then*

$$\|P_n \otimes K_n - P \otimes K\| \le 5\|P_n - P\| + 4\|Q_n - Q\|.$$

PROOF. By triangular inequality,

$$\|P_n \otimes K_n - P \otimes K\| \leq \|(P_n \otimes Q_n) \wedge (Q_n \otimes P_n) - (P \otimes Q) \wedge (Q \otimes P)\| + \|P_n R_n - PR\|$$

where PR and $P_n R_n$ are $P(dx)R(x)$ and $P_n(dx)R_n(x)$ respectively. By the representation of R in (26), the second term in the right-hand side of the above is dominated by $\|P_n - P\|$ plus the first term. To find a bound of the first term, observe that for any $x_1, x_2, y_1, y_2 \in \mathbb{R}$ we have $|x_1 \wedge x_2 - y_1 \wedge y_2| \leq \sum_{i=1}^2 |x_i - y_i|$. By this inequality, $\|\mu_1 \wedge \mu_2 - \nu_1 \wedge \nu_2\| \leq \sum_{i=1}^2 \|\mu_i - \nu_i\|$. Thus we have

$$\begin{aligned}
\|P_n \otimes K_n - P \otimes K\| &\leq 2\|(P_n \otimes Q_n) \wedge (Q_n \otimes P_n) - (P \otimes Q) \wedge (Q \otimes P)\| + \|P_n - P\| \\
&\leq 4\|P_n \otimes Q_n - P \otimes Q\| + \|P_n - P\| \\
&= 4\|(P_n \otimes Q_n - P_n \otimes Q) + (P_n \otimes Q - P \otimes Q)\| + \|P_n - P\| \\
&\leq 5\|P_n - P\| + 4\|Q_n - Q\|.
\end{aligned}$$

Now we apply the above to the model (8) but we do not assume $F_t = N(t, 1)$ here. After scaling $\theta \mapsto c_n\theta$, the IMH procedure in Section 4.1 corresponds to

$$P_n(d\theta) = P^*(d\theta|x_n),\ Q_n(d\theta) = Q^*(d\theta|x_n)$$

where $P^*(d\theta|x_n)$ and $Q^*(d\theta|x_n)$ are as in Proposition A3, and (25). We already stated

$$\|P_n^*(d\theta|x_n) - P^*(d\theta|z = z_n)\| = o_{\mathbb{P}}(1),\ \|Q_n^*(d\theta|x_n) - Q^*(d\theta|z = z_n)\| = o_{\mathbb{P}}(1).$$

Although these probability measures are random (depending on x_n), by Skorohod's representation theorem, we can assume that the above convergence holds for each ω and $z_n(\omega) \to z(\omega) \sim N(0, 1)$. Also we note that $z \mapsto P^*(d\theta|z)$ and $z \mapsto Q^*(d\theta|z)$ are continuous in total variation norm, and $P^*(d\theta|z)$ and $Q^*(d\theta|z)$ are absolutely continuous for each z. Hence the Markov chain generated by the IMH procedure has local consistency by Lemmas C1, C2 for localization $\theta \mapsto c_n\theta$.

References

Bo Martin Bibby, Ib Michael Skovgaard, and Michael Sørensen (2005). Diffusion-type models with given marginal distribution and autocorrelation function. *Bernoulli*, 11 (2):191–220.

Patrick Billingsley (1999). *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition.

Persi Diaconis and Daniel Stroock (1991). Geometric bounds for eigenvalues of Markov chains. *Ann. Appl. Probab.*, 1(1):36–61.

Persi Diaconis, Kshitij Khare, and Laurent Saloff-Coste (2008). Gibbs sampling, exponential families and orthogonal polynomials. With comments and a rejoinder by the authors. *Statist. Sci.*, 23(2):151–178

Stewart N. Ethier and Thomas G. Kurtz. (1986). *Markov Processes: Characterization and Convergence* Wiley Series in Probability and Statistics. Wiley.

Sylvia Frühwirth-Schnatter (2006). *Finite mixture and Markov switching models.* Springer series in statistics. Springer, New York.

Elisabeth Gassiat (2002). Likelihood ratio inequalities with applications to various mixtures. *Ann. Inst. H. Poincaré Probab. Statist.*, 38(6):897–906.

Saul B. Gelfand and Sanjoy K Mitter (1991). Weak convergence of Markov chain sampling methods and annealing algorithms to diffusions. *J. Optim. Theory Appl.*, 68 (3):483–498.

James P. Hobert and Dobrin Marchev (2008). A theoretical comparison of the data augmentation, marginal augmentation and PX-DA algorithms. *Ann. Statist.*, 36(2): 532–554.

Jean Jacod and Albert N. Shiryaev (2003). *Limit theorems for stochastic processes.* 2nd Edition. Grundlehren der Mathematischen Wissenschaften. Springer-Verlag, Berlin.

A. Jasra, C. C. Holmes, and D. A. Stephens (2005). Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling. *Statistical Science*, 20(1):50–67.

Kengo Kamatani (2011). Local degeneracy of Markov chain Monte Carlo methods. *arXiv:1108.2477.*

Kengo Kamatani (2013). Local consistency of Markov chain Monte Carlo methods. *Ann. Inst. Statist. Math.*, in press.

Ioannis Karatzas and Steven E. Shreve (1991). *Brownian motion and stochastic calculus.* 2nd ed edition. Number 113 in Graduate texts in mathematics. Springer-Verlag.

Jean-Michel Marin, Kerrie Mengersen, and Christian P. Robert (2005). Bayesian modelling and inference on mixtures of distributions. In D.K. Dey and C.R. Rao, editors, *Bayesian Thinking Modeling and Computation*, volume 25 of *Handbook of Statistics*, pages 459 – 507. Elsevier.

Kerrie L. Mengersen and Richard L. Tweedie (1996). Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.*, 24(1):101–121.

Sean P. Meyn and R. L. Tweedie (1994). Computable bounds for geometric convergence rates of Markov chains. *Ann. Appl. Probab.*, 4(4):981–1011.

Antonietta. Mira (1998). *Ordering, Slicing and Splitting Monte Carlo Markov Chains.* PhD thesis, University of Minnesota.

Peter H. Peskun (1973). Optimum monte-carlo sampling using markov chains. *Biometrika*, 60(3):607–612.

Christian P. Robert and George Casella (2004). *Monte Carlo Statistical Methods.* 3rd edition. Springer.

Gareth O. Roberts and Jeffrey S. Rosenthal (1998). Optimal scaling of discrete approximations to langevin diffusions. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 60(1):255–268.

Gareth O. Roberts and Jeffrey S. Rosenthal (2004). General state space markov chains and mcmc algorithms. *Probability Surveys*, 1:20–71.

Gareth O. Roberts and Richard L. Tweedie (1996). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1):95–110.

Gareth O. Roberts, Andrew Gelman, and Walter R. Gilks (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.*, 7 (1):110–120.

Jeffrey S. Rosenthal (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *J. Amer. Statist. Assoc.*, 90(430):558–566.

Jeffrey S. Rosenthal (2002). Quantitative convergence rates of markov chains: A simple account. *Electron. Commun. Probab.*, 7:no. 13, 123–128.

Chris Sherlock, Paul Fearnhead, and Gareth O. Roberts (2010). The random walk Metropolis: linking theory and practice through a case study. *Statist. Sci.*, 25(2): 172–190.

Matthew Stephens (2000). Dealing with label switching in mixture models. *J. R. Stat. Soc. Ser. B*, 62(4):795–809.

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

Luke Tierney (1998). A note on Metropolis-Hastings kernels for general state spaces. *Ann. Appl. Probab.*, 8(1):1–9.

A.W. van der Vaart (1998). *Asymptotic statistics*. Cambridge University Press, New York.

Received September 25, 2013
Revised November 25, 2013