数理解析研究所講究録
第1860巻 2013年 140-145

Note on asymptotic properties of probit Gibbs sampler

鎌谷 研吾

KENGO KAMATANI [*]

大阪大学大学院基礎工学研究科

GRADUATE SCHOOL OF ENGINEERING SCIENCE, OSAKA UNIVERSITY [†]

1 Introduction

The Markov chain Monte Carlo (MCMC) method is an efficient tool for the approximation of an integral with respect to a particular type of probability measure. The strategy has been developed in the past 50 years and it becomes one of the most popular method in the Bayesian statistics. See Robert and Casella [9] for a recent review.

Let $p(dx|\theta)=p(x|\theta)dx$ be a probability measure with the prior distribution $p(d\theta)$. The posterior distribution $p(d\theta|x)$ is proportional to $p(x|\theta)p(d\theta)$ under an observation x. Sometimes the posterior distribution does not have a closed form that usually requires some kind of approximation. Present paper deal with the so-called data augmentation (DA) procedure. This procedure uses the so-called augment data model $p(dxdy|\theta)$ that satisfies $\int_{Y}p(dxdy|\theta)=p(dx|\theta)$. The DA procedure iterates the following;

$$\textbf{simulate } y\sim p(dy|x,\theta),\ \textbf{simulate } \theta\sim p(d\theta|x,y) \tag{1}$$

where $p(d\theta|x,y)$ is the posterior distribution of $p(dxdy|\theta)$ with the prior $p(d\theta)$. This procedure results in a Markov chain $\theta_{0},\theta_{1},\ldots$ with invariant distribution $p(d\theta|x)$ (see Tierney [11], Gilks et al. [3], Roberts and Rosenthal [10]; the sequence of y is omitted here). Moreover under mild conditions, the value $I_{m}=m^{-1}\sum_{i=0}^{m-1}\varphi(\theta_{i})$ converges to $I=\int\varphi(\theta)p(d\theta|x)$ as $m\rightarrow\infty$ for any function φ and for each observation x, so we can use I_{m} as an approximation of I.

Sometimes the convergence of I_{m} to I is very slow and there have been a lot of efforts for the analysis of the sufficient number of iteration (see ex. Roberts and Rosenthal [10], Diaconis et al. [2]). In the current paper we review one of these approaches, "the large sample" approach by Kamatani [5, 6, 7] (we call the other, "fixed sample size" approaches).

Consider the following simple model;

$$x\sim \mathrm{Bernoulli}(\Phi(\theta)) \tag{2}$$

where Φ is the cumulative distribution function of the standard normal distribution $N(0,1)$. Let $x_{n}=\{x^{1},\ldots,x^{n}\}$ be an i.i.d. sample from this model and let $N(0,1)$ be the prior distribution. Consider two

[*] Supported in part by Grant-in-Aid for Young Scientists (B) 24740062.

[†] kamatani@sigmath.es.osaka-u.ac.jp

choices of the augmented data models and construct two DA procedures corresponding to these models;

$$y \sim N(0,1), \ x = 1_{\{y \leq \theta\}} \tag{3}$$

and

$$y \sim N(-\theta,1), \ x = 1_{\{y \leq 0\}}. \tag{4}$$

Though two DA procedures have similar steps and both of which have geometric ergodicity, the performances of them are quite different. This difference is strongly related to their model regularity. The former model has the parameter-depending support, and the latter is a regular model. It is quite usual that the regular and non-regular models have different asymptotic properties (see ex. Akahira and Takeuchi [1]). In the present paper we analyze the large sample size approach through a very simple binomial model (2) with the augmented data model (3).

2 Key properties

Let $x_n = \{x^1, \ldots, x^n\}$ be an independent observation from $p(dx|\theta)$ and $\theta \sim p(d\theta)$ (assume the subjective Bayes setting). Let $\hat{\theta}_n$ be the maximum likelihood estimator. Let $I_n := \int \psi(\sqrt{n}(\theta - \hat{\theta}_n))p(d\theta|x_n)$ and $I_{n,m} := m^{-1}\sum_{i=0}^{m-1}\psi(\sqrt{n}(\theta_i - \hat{\theta}_n))$. Write $\mathbb{P}_n$ for the underlying probability measure. It would be helpful if we have for any $m_n \to \infty$ and any bounded and continuous function ψ as $n \to \infty$

$$I_n - I_{n,m_n} = o_{\mathbb{P}_n}(1). \tag{5}$$

This property was called the local consistency of the MCMC procedure in Kamatani [7]. This property is always satisfied for the DA procedure under regularity conditions. Assume the following regularity conditions.

1. $\{p(dx|\theta)\}$ is quadratic mean differentiable.
2. The Fisher information matrix of $\{p(dx|\theta)\}$ is non-singular.
3. $\{p(dx|\theta)\}$ has a uniformly consistent test.
4. The prior p has a continuous, positive and bounded density.
5. $\{p(dx|\theta)\}$ is identifiable.

Theorem 1 (Kamatani [7])
Assume the above conditions and assume $\theta_0 \sim p(d\theta|x_n)$. Then the DA procedure has the local consistency.

This fact illustrates that if an MCMC have poor performances, then the model may not satisfy the above conditions. For the analysis of such poor MCMC procedures, the local degeneracy was defined in Kamatani [6]. The MCMC procedure is called locally degenerate if for any $m \in \mathbf{N}$ and any continuous and bounded function ψ as $n \to \infty$

$$I_{n,1} - I_{n,m} = o_{\mathbb{P}_n}(1). \tag{6}$$

This means that the MCMC procedure using m iteration does not provide helpful approximations of I than that using only one iteration. If

$$\theta_0 \sim p(d\theta|x_n) \tag{7}$$

and if the posterior distribution has $\sqrt{n}$-consistency, this is equivalent to

$$\sqrt{n}|\theta_1 - \theta_0| = o_{\mathbb{P}_n}(1). \tag{8}$$

We call (7) stationarity condition and assume it throughout in this paper. This assumption is impractical but if we take a good initial guess θ_0 the same results hold (see Kamatani [7] for details).

The following order is useful to calculate the severity of the degeneracy.

Definition 2 (Order of degeneracy)
For an increasing sequence of positive number d_n, if

$$d_n\sqrt{n}|\theta_1 - \theta_0| = O_{\mathbb{P}_n}(1) \tag{9}$$

then $\{d_n\}$ is called the order of the local degeneracy of the MCMC procedure.

This property is easy to check in practice. However it is rather indirect approach for the analysis of (5). The following more direct approach due to Kamatani [5] may be appealing.

Definition 3 (Order of weak consistency)
For an increasing sequence of positive number d'_n, if (5) holds for any m_n such that $m_n/d'_n \to \infty$, then $\{d'_n\}$ is called the rate of local weak consistency of the MCMC procedure.

The sequence d'_n corresponds to the sufficient number of iteration of the MCMC procedure. For local consistent MCMC procedure, $d'_n \equiv 1$, that is the best possible order.

The next section provides an example for the analysis of the order of the local degeneracy and the order of the local weak consistency and their relation.

3 An application

Consider a reparametrization $\theta \mapsto \Phi^{-1}(\theta)$. Then the model (2) becomes extremely simple;

$$x \sim \text{Bernoulli}(\theta) \tag{10}$$

with the prior distribution $U(0,1)$. Let $x_n = \{x^1, \dots, x^n\}$ be an i.i.d. copy from the above model. Write $n_i = \sum_{j=1}^n 1_{\{x^j = i\}}$ $(i = 0, 1)$. Note that the Bayes estimator $\hat{\theta}_n = (n_1 + 1)/(n+2)$ is a sufficient statistic. The DA procedure becomes

$$\text{for } n = 1, 2, \dots, \textbf{ simulate } y^i \sim \begin{cases} U(0,\theta] & \text{if } x^i = 1 \\ U(\theta, 1) & \text{if } x^i = 0 \end{cases}, \text{ then } \textbf{simulate } \theta \sim U[y_*, y^*) \tag{11}$$

where $y^* := \min_{i;x^i=0} y^i$ and $y_* := \max_{i;x^i=1} y^i$.

First we remark a fixed sample size property. Let $\|\nu\| = 2\sup_{A\in\mathcal{E}} |\nu(A)|$ where ν is a signed measure on a measurable space $(E, \mathcal{E})$. For transition kernels $S(x, dy)$ and $T(x, dy)$ on $(E, \mathcal{E})$, let $(ST)(x, dz) = \int_{y\in E} S(x, dy)T(y, dz)$. Write $S^1 = S$ and $S^n = S^{n-1}S$ $(n \geq 2)$. Assume that there exists the invariant probability measure Π for a transition kernel S. If there exists $R < \infty$ and $\rho \in (0, 1)$ such that

$$\|S^n(x, \cdot) - \Pi\| \leq R\rho^n \ (x \in X) \tag{12}$$

then this Markov chain is called uniformly ergodic. If $X_1, X_2 \ldots,$ is a Markov chain with the transition kernel S that have uniformly ergodicity, then $n^{-1}\sum_{i=1}^{n} f(X_i) \to 0$ almost surely and the central limit theorem holds for $n^{-1/2}\sum_{i=1}^{n} f(X_i)$ for any $f \in L^2(\Pi)$ and $\Pi(f) = 0$.

Uniform ergodicity can be checked by the Döblin condition.

Proposition 4

Assume $n_0, n_1 \geq 1$. Then the Markov chain defined by the DA procedure is uniformly ergodic.

Proof Write θ for the current value and θ' for the next value of one iteration of this DA procedure. By definition, when $\theta \in (0,1)$,

$$\mathbb{P}_n(\frac{y^* - \theta}{1-\theta} > t | x_n, \theta) = (1-t)^{n_0}, \mathbb{P}_n(\frac{\theta - y^*}{\theta} > s | x_n, \theta) = (1-s)^{n_1}. \tag{13}$$

Hence the transition kernel of this DA procedure $k(\theta, \theta')d\theta'$ is

$$\begin{aligned} k(\theta,\theta') &= \int_{s,t\in[0,1]} \frac{1_{[\theta(1-s),\theta+(1-\theta)t]}(\theta')}{(1-\theta)t+\theta s} \frac{(1-s)^{n_1-1}}{n_1} \frac{(1-t)^{n_0-1}}{n_0} dsdt \\ &\geq \left(\frac{\theta'}{\theta}\right)^{n_1} \left(\frac{1-\theta'}{1-\theta}\right)^{n_0} \geq \theta'^{n_1}(1-\theta')^{n_0} \ (\theta, \theta' \in (0,1)) \end{aligned}$$

where we used $(1-\theta)t + \theta s \leq 1$ in the first inequality. Hence the Markov chain is uniformly ergodic by Theorem 16.2.4 of [8]. ∎

We have the following large sample property.

Proposition 5

Assume the stationarity condition. For the model (10), the DA procedure is locally degenerate of the order $d_n = \sqrt{n}$.

Proof It is clear since $n^{1/2}|\theta' - \theta| \leq n^{1/2}(y^* - y_*) = O_{\mathbb{P}_n}(n^{-1/2})$. ∎

Also we have the following.

Proposition 6

Assume the stationarity condition. For the model (10), the DA procedure is locally weak consistent of the order $d'_n = n$.

Proof Since $\theta' \sim U[y_*, y^*)$,

$$\begin{aligned} \mathbb{E}_n[\exp(\lambda(\theta' - \theta))|x_n, y_n, \theta] &= \frac{\exp(\lambda(y^* - \theta)) - \exp(-\lambda(\theta - y_*))}{\lambda(y^* - \theta) - (-\lambda(\theta - y_*))} \\ &= \sum_{k=1}^{\infty} \frac{1}{k!} \frac{(\lambda(y^* - \theta))^k - (-\lambda(\theta - y_*))^k}{\lambda(y^* - \theta) - (-\lambda(\theta - y_*))} \\ &= \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{k!} \sum_{i=0}^{k-1} (y^* - \theta)^{k-i-1}(-(\theta - y_*))^i. \end{aligned} \tag{14}$$

Next we integrate out y_* and y^* in the above using (13). For $k \geq 0$,

$$\mathbb{E}_n[(y^* - \theta)^k | x_n, \theta] = (1-\theta)^k \binom{k+n_0}{n_0}^{-1}, \ \mathbb{E}_n[(\theta - y_*)^k | x_n, \theta] = \theta^k \binom{k+n_1}{n_1}^{-1}.$$

Let $h_n = \sqrt{n}(\theta - \hat{\theta}_n)$. Now for a random variable $X_n = f_n(h_n, \hat{\theta}_n)$, we denote $X_n = O(n^\alpha)$ if $\limsup_{n\to\infty} \sup_{(h,\theta)\in K} |n^{-\alpha} f_n(h,\theta)| < \infty$ for any compact set $K \subset \mathbf{R} \times (0,1)$. By considering Taylor's expansion of the left hand side of (14), we obtain

$$\mu_k := \mathbb{E}_n[n^{k/2}(\theta^* - \theta)^k | x_n, \theta] = \frac{n^{k/2}}{k+1} \sum_{i=0}^{k} \binom{k-i+n_0}{n_0}^{-1} \binom{i+n_1}{n_1}^{-1} (-\theta)^i (1-\theta)^{k-i}.$$

By simple algebra,

$$\mu_1 = -\frac{1}{2n} \frac{h_n}{\hat{\theta}_n(1-\hat{\theta}_n)} + O(n^{3/2}),\ \mu_2 = n^{-1} + O(n^{-3/2}),\ \mu_4 = O(n^{-4}).$$

By choosing suitable probability space $(\Omega, \mathcal{F}, \mathbb{P})$, without loss of generality we may assume $\hat{\theta}_n(\omega)$ tends to $\hat{\theta}_n(\omega) \in (0,1)$ for all ω (Skorohod's representation theorem). Write $\mathcal{G}$ for the σ-algebra generated by $\{\hat{\theta}_n, n = 1, 2, \ldots\}$. Consider a stochastic process $h_n(t) = n^{1/2}(\theta([nt]) - \hat{\theta}_n)$. By checking conditions in Theorem 9.4.21 of Jacod and Shiryaev [4], we can show that the law of $\{h_n(t); t \geq 0\}$ tends $\mathcal{G}$-stably to that of the following Ornstein—Uhlenbeck process $\{h(t); t \geq 0\}$;

$$dh(t) = -\frac{h(t)}{2\hat{\theta}(1-\hat{\theta})} dt + dW(t); h(0) \sim N(0, 1/\hat{\theta}(1-\hat{\theta})) \tag{15}$$

where W is the standard Wiener process and $\hat{\theta} \sim U[0,1]$ and both of which are independent. Then the claim follows by Kamatani [5]. ∎

The convergence to the diffusion process (15) illustrates the difference between the order of local degeneracy and that of local weak consistency. Essentially, these orders are defined to be

$$\sum_{i=1}^{d_n} |\Delta_i h_n| \approx d_n / n^{1/2} = O_{\mathbb{P}}(1), \text{ or } (\sum_{i=1}^{d'_n} |\Delta_i h_n|^2)^{1/2} \approx \sqrt{d'_n / n} = O_{\mathbb{P}}(1) \tag{16}$$

with respectively, where $\Delta_i h_n = h_n(i/n) - h_n((i-1)/n)$. Therefore it is natural that the order of local degeneracy is smaller than that of the local weak consistency.

On the other hand, by Theorem 1, the DA procedure using the augmented data model (4) is locally consistent that can not be locally degenerate. Therefore the DA procedure using (3) should be much worse than that using (4).

References

[1] M. Akahira and K. Takeuchi. *Non-regular statistical estimation.* Lecture notes in statistics. Springer, 1995.

[2] P. Diaconis, K. Khare, and L. Saloff-Coste. Gibbs Sampling, Exponential Families and Orthogonal Polynomials (with discussion). *Statistical Science*, 23(2):151–200, 2008.

[3] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice.* CRC Press, 1996.

[4] J. Jacod and A. N. Shiryaev. *Limit theorems for stochastic processes.* Grundlehren der Mathematischen Wissenschaften. Springer-Verlag, Berlin, 2nd edition, 2003.

[5] K. Kamatani. Large sample scaling limit to a diffusion process of Markov chain Monte Carlo methods. *arXiv:1103.5679*, 2011.

[6] K. Kamatani. Local degeneracy of Markov chain Monte Carlo methods. *arXiv:1108.2477*, 2011.

[7] K. Kamatani. Local consistency of Markov chain Monte Carlo methods. *Annals of the Institute of Statistical Mathematics*, (doi:10.1007/s10463-013-0403-3), 2013.

[8] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability.* Springer, 1993.

[9] C. Robert and G. Casella. A Short History of Markov Chain Monte Carlo - Subjective Recollections from Incomplete Data. *Statistical Science*, pages 1–21, 2011.

[10] G. O. Roberts and J. S. Rosenthal. General state space markov chains and mcmc algorithms. *Probability Surveys*, 1:20–71, 2004.

[11] L. Tierney. Markov Chains for Exploring Posterior Distributions (with discussion). *The Annals of Statistics*, 22(4):1701–1762, 1994.