ASYMPTOTIC PROPERTIES OF MONTE CARLO STRATEGIES FOR A CUMULATIVE LINK MODEL

Kengo Kamatani*

For a cumulative link model in the Bayesian context, the posterior distribution cannot be obtained in closed form, and we have to resort to an approximation method. A simple data-augmentation strategy is widely used for that purpose but is known to work poorly. The marginal augmentation procedure and the parameter-expanded data-augmentation procedure are considered to be remedies, but such strategies are still not free from poor convergence. In this paper, we propose a kind of the hybrid Markov chain Monte Carlo strategy. To evaluate the efficiency, a local non-degeneracy is introduced, and we also provide a numerical simulation to show the effect.

Key words and phrases: Bayesian inference, Markov chain, Monte Carlo method, probit model.

1. Introduction

For ordinal response data, models such as the cumulative logit and probit models have been developed in the past few decades. These models were proposed by Aitchison and Silvey (1957) and became popular due to a seminal paper by McCullagh (1980). They have been applied to many areas, such as in the social sciences (Samejima (1969)) and in the action of drugs (Sheiner *et al.* (1997)). See Liu and Agresti (2005) for a general review of ordinal response modeling. The cumulative probit model for a c -category ordinal response variable y with an explanatory variable $x \in \mathbb{R}^p$ has the form

$$(1.1) \quad P(y \leq j \mid x) = \Phi(\alpha_j + \beta^t x) \quad (j = 1, \dots, c),$$

where β^t is a transpose of the vector $\beta \in \mathbb{R}^p$ and Φ is the cumulative distribution function of the standard normal distribution. Its parameter is $\theta = (\alpha, \beta)$ where $\alpha = (\alpha_2, \dots, \alpha_{c-1})$ satisfies a constraint condition $0 = \alpha_1 < \dots < \alpha_{c-1} < \alpha_c = +\infty$. When the parameters are inferred from the likelihood approach, there are computational difficulties that usually require some kind of approximation, such as the Gauss-Hermite quadrature or the Monte Carlo expectation-maximization method. In this paper, we focus on the Bayesian approach. As in the likelihood approach, the posterior distribution is complicated. The data-augmentation (DA) procedure by Albert and Chib (1993) is easy to implement, and so it is commonly used for this approximation. However, it has a slow mixing property, which results in inference bias. To illustrate this, we consider the following simplified model.

Received August 28, 2012. Revised November 24, 2012. Accepted February 25, 2014.

*Graduate School of Engineering Science, Osaka University, 1-3 Machikaneyama-cho, Toyonaka-si, Osaka 560-0043, Japan.

Let y be a two-category variable with the following law:

$$(1.2) \quad P(y = 1) = \Phi(\theta), \quad P(y = 2) = 1 - \Phi(\theta),$$

with the prior $N(0, 1)$ for θ . Consider the so-called augmented-data model that is required by the DA procedure:

$$(1.3) \quad z \sim N(0, 1), \quad y = 1 + 1_{\{z > \theta\}}.$$

The DA procedure results in a sequence $\theta_0, \theta_1, \dots$, and by using this sequence, it is possible to approximate integrals with respect to the posterior distribution. For example, to obtain the posterior mean of θ , approximate it by $M^{-1} \sum_{m=0}^{M-1} \theta_m$. Figure 1 shows the results of the DA procedure for sample sizes $n = 10^2$ and 10^3 , with four runs for each. Although the sample size $n = 10^3$ is unrealistic, it makes the problem clear. It seems to have a very small mixing property, and this causes poor convergence. Note that this DA procedure has uniform ergodicity.

Figure 1. Trajectory of the data-augmentation strategy for $n = 100$ (left) and $n = 1000$ (right). The dashed lines are the 10th, 50th and 90th percentile of the posterior distribution.

Since this poor behavior is well known, there are some techniques for improvements. These include the Metropolis-within-Gibbs procedure (Cowles (1996)) and the parameter-expanded data-augmented (PX-DA) procedure (Liu and Sabatti (2000)). However, the former requires a careful choice of tuning parameters, and the latter works well only for a small number of categories. The aim of this paper is to find an efficient procedure without this poor behavior. With this in mind, we propose the hybrid PX-DA Metropolis-Hastings (HPX-DA-MH) procedure, which is a hybrid of the Markov chain Monte Carlo (MCMC) procedure (see Section 2.4 of Tierney (1994)) and hence possesses a Metropolis step. It is worth pointing out that our Metropolis step does not require any tuning parameters.

The efficiency of a strategy is determined by numerical simulations and evaluation of the local properties (Kamatani (2014a, b)). Global properties, such as the ergodicity of a Markov chain, may not provide much information since the DA for the above toy example has good ergodicity; it is uniformly ergodic, although the convergence is quite slow. To check the local properties, we introduce the local non-degeneracy, which is easy to check. Our strategy has been

shown to have the local non-degeneracy, even though the usual DA has the local degeneracy (Kamatani (2014b)).

The rest of the paper is organized as follows. In Section 2, we briefly discuss traditional MCMC procedures such as the DA procedure, the Metropolis-Hastings (MH) procedure, and the hybrid procedure. These procedures are then used to construct our strategy. We also introduce the local non-degeneracy in this section by means of the dispersion property, which was considered by Itô (2004). In Section 3, the HPX-DA-MH strategy is introduced and analyzed for an ordinal probit model. The results of numerical simulations are presented in Section 4.

2. Some properties of degeneracy

In this section, for any two measurable spaces $(\mathbf{A}, \mathcal{A})$ and $(\mathbf{B}, \mathcal{B})$, we write $P(B \mid a)$ ($B \in \mathcal{B}, a \in \mathbf{A}$) for a (probability) transition kernel from $\mathbf{A}$ to $\mathbf{B}$, that is, $P(\cdot \mid a)$ is a probability measure on $\mathbf{B}$ for any $a \in \mathbf{A}$, and $a \mapsto P(B \mid a)$ is $\mathcal{A}$ -measurable for any $B \in \mathcal{B}$.

2.1. The dispersion

For a probability measure μ on $\mathbb{R}^d$, the dispersion $\delta(\mu)$ is defined by

$$\delta(\mu) = -\log \int_{\mathbb{R}^d \times \mathbb{R}^d} \exp \left(-\sum_{j=1}^d |x_j - y_j| \right) \mu(dx) \mu(dy)$$

for $x = (x_1, \dots, x_d)$ and $y = (y_1, \dots, y_d)$. It is a measure of variation of μ ; $\delta(\mu) = 0$ if and only if $\mu = \delta_x$ for some $x \in \mathbb{R}^d$. Also $\delta(\mu_n) \rightarrow 0$ if and only if $\mu_n(\cdot - x_n) \rightarrow \delta_0$ in the Prohorov metric for some $x_n \in \mathbb{R}^d$ ($n \in \mathbb{N}$). See Itô (2004) for the details. Dispersion is the key for the definition of the degeneracy. The following is another representation of the dispersion.

LEMMA 2.1 (Representation of the dispersion).

$$\delta(\mu) = -\log \int_{\mathbb{R}^d} |\varphi(\xi; \mu)|^2 \prod_{j=1}^d \frac{d\xi_j}{\pi(1 + \xi_j^2)}$$

where $\xi = (\xi_1, \dots, \xi_d)$ and $\varphi(\xi; \mu) = \int \exp(i\xi^t x) \mu(dx)$ is the characteristic function of μ .

PROOF. The characteristic function of the Cauchy distribution is

$$\exp(-|u|) = \int_{\xi \in \mathbb{R}} \exp(iu\xi) \frac{d\xi}{\pi(1 + \xi^2)}.$$

Hence we have

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \exp \left(-\sum_{j=1}^d |x_j - y_j| \right) \mu(dx) \mu(dy)$$

$$\begin{aligned}
&= \int_{\mathbb{R}^d \times \mathbb{R}^d} \prod_{j=1}^d \exp(-|x_j - y_j|) \mu(dx) \mu(dy) \\
&= \int_{\mathbb{R}^d \times \mathbb{R}^d} \left\{ \int_{\xi \in \mathbb{R}^d} \prod_{j=1}^d \exp(i(x_j - y_j)\xi_j) \frac{d\xi_j}{\pi(1 + \xi_j^2)} \right\} \mu(dx) \mu(dy) \\
&= \int_{\mathbb{R}^d} \varphi(\xi; \mu) \varphi(-\xi; \mu) \prod_{j=1}^d \frac{d\xi_j}{\pi(1 + \xi_j^2)} = \int_{\mathbb{R}^d} |\varphi(\xi; \mu)|^2 \prod_{j=1}^d \frac{d\xi_j}{\pi(1 + \xi_j^2)}
\end{aligned}$$

where we used Fubini's theorem in the third equation. $\square$

By this, δ is continuous with respect to the Prohorov metric and it is shift invariant, that is, $\delta(\mu) = \delta(\mu(\cdot - x))$ for $x \in \mathbb{R}^d$. Let X on $(\Omega, \mathcal{F}, \mathbb{P})$ have a distribution μ . We write $\delta(X)$ for $\delta(\mu)$. We also write $\delta(\mathbb{P}(X \in \cdot \mid \mathcal{G}))$ by $\delta(X \mid \mathcal{G})$ where $\mathcal{G}$ is a sub σ -algebra of $\mathcal{F}$. If the sub σ -algebra is generated by Y , then we write $\delta(X \mid Y)$ instead. The following is a simple extension of Itô (2004) that will be used throughout.

LEMMA 2.2. *Let X_n and Y_n be $\mathbb{R}^d$ -valued random variables. Let $\{\mathcal{G}_n\}$ and $\mathcal{G}$ be sub σ -algebras of $\mathcal{F}$.*

- (i) *If X and Y are $\mathbb{R}^d$ -valued and conditionally independent given $\mathcal{G}$, then $\delta(X + Y \mid \mathcal{G}) \geq \delta(X \mid \mathcal{G})$.*
- (ii) *If $\delta(X_n) \rightarrow 0$, then $\delta(X_n \mid \mathcal{G}_n) = o_{\mathbb{P}}(1)$.*
- (iii) *$\delta(X_n \mid \mathcal{G}_n) \rightarrow 0$ if and only if $X_n - Y_n = o_{\mathbb{P}}(1)$ for some $\mathcal{G}_n$ -measurable random variable Y_n .*
- (iv) *$\delta(X_n \mid \mathcal{G}_n) \rightarrow 0$ and $\delta(Y_n \mid \mathcal{G}_n) \rightarrow 0$ implies $\delta(X_n + Y_n \mid \mathcal{G}_n) \rightarrow 0$.*
- (v) *Let Y_n be $\mathbb{R}$ -valued and $\mathcal{G}_n$ -measurable. If $\delta(X_n \mid \mathcal{G}_n) \rightarrow 0$ and if Y_n is tight, then $\delta(X_n Y_n \mid \mathcal{G}_n) \rightarrow 0$.*

PROOF. The first claim is obvious since

$$\begin{aligned}
|\mathbb{E}[\exp(i\xi^t(X + Y)) \mid \mathcal{G}]| &= |\mathbb{E}[\exp(i\xi^t X) \mid \mathcal{G}] \mathbb{E}[\exp(i\xi^t Y) \mid \mathcal{G}]| \\
&\leq |\mathbb{E}[\exp(i\xi^t X) \mid \mathcal{G}]|.
\end{aligned}$$

For the second claim, observe $\varphi(\xi; \mu_n) = \mathbb{E}[\exp(i\xi^t X_n) \mid \mathcal{G}_n]$ for $\mu_n(\omega, A) = \mathbb{P}(X_n \in A \mid \mathcal{G}_n)$. By Schwarz's inequality,

$$|\mathbb{E}[\exp(i\xi^t X_n)]|^2 = |\mathbb{E}[\mathbb{E}[\exp(i\xi^t X_n) \mid \mathcal{G}_n]]|^2 \leq \mathbb{E}[|\mathbb{E}[\exp(i\xi^t X_n) \mid \mathcal{G}_n]|^2].$$

Thus we obtain by Fubini's theorem,

$$\begin{aligned}
&\int |\mathbb{E}[\exp(i\xi^t X_n)]|^2 \prod_{j=1}^d \frac{d\xi_j}{\pi(1 + \xi_j^2)} \\
&\leq \mathbb{E} \left[\left\{ \int_{\xi \in \mathbb{R}^d} |\mathbb{E}[\exp(i\xi^t X_n) \mid \mathcal{G}_n]|^2 \prod_{j=1}^d \frac{d\xi_j}{\pi(1 + \xi_j^2)} \right\} \right].
\end{aligned}$$

By assumption, the left hand side tends to 1, and hence $\int |\mathbb{E}[\exp(i\xi^t X_n) | \mathcal{G}_n]|^2 \prod (\pi(1 + \xi_j^2))^{-1} d\xi$ tends in probability to 1 and hence $\delta(X_n | \mathcal{G}_n) \rightarrow 0$. Thus the second claim follows.

Next we show the third claim. Assume that $X_n - Y_n = o_{\mathbb{P}}(1)$ for some $\mathcal{G}_n$ -measurable random variable Y_n . Then for any $\xi \in \mathbb{R}^d$, by Jensen's inequality,

$$\begin{aligned} \mathbb{E}[|\mathbb{E}[\exp(i\xi^t X_n) | \mathcal{G}_n]|] &= \mathbb{E}[|\mathbb{E}[\exp(i\xi^t (X_n - Y_n)) | \mathcal{G}_n]|] \\ &\geq |\mathbb{E}[\exp(i\xi^t (X_n - Y_n))]|. \end{aligned}$$

The right hand side tends to 1. Thus, as the second claim, $\int |\mathbb{E}[\exp(i\xi^t X_n) | \mathcal{G}_n]|^2 \prod (\pi(1 + \xi_j^2))^{-1} d\xi$ tends in probability 1 and hence $\delta(X_n | \mathcal{G}_n) \rightarrow 0$. On the other hand, assume $\delta(X_n | \mathcal{G}_n) \rightarrow 0$. Then we can find $\mathcal{G}_n$ -measurable random variables $Y_n = (Y_{n1}, \dots, Y_{nd})$ so that

$$\begin{aligned} \int \exp\left(-\sum_{j=1}^d |x_j - Y_{nj}|\right) \mu_n(\omega, dx) \\ \geq \int \exp\left(-\sum_{j=1}^d |x_j - y_j|\right) \mu_n(\omega, dx) \mu_n(\omega, dy). \end{aligned}$$

Since the left hand side tends to 1 in probability, $X_n - Y_n$ tends in probability to 0.

The forth claim is an easy corollary of the third claim. Indeed, by assumption, there exists $\mathcal{G}_n$ -measurable random variables X'_n and Y'_n such that $X_n - X'_n = o_{\mathbb{P}}(1)$ and $Y_n - Y'_n = o_{\mathbb{P}}(1)$ and hence $(X_n + Y_n) - (X'_n + Y'_n) = o_{\mathbb{P}}(1)$.

Finally, if Y_n is tight, then $Y_n(X_n - X'_n) = o_{\mathbb{P}}(1)$ by Slutsky's lemma. Since $Y_n X'_n$ is $\mathcal{G}_n$ -measurable, the claim follows by the third claim. $\square$

As an example, we consider the dispersion of the gamma distribution. Let $\text{Gamma}(\alpha, \beta)$ be the gamma distribution with density function $\beta^\alpha x^{\alpha-1} \exp(-\beta x) / \Gamma(\alpha)$.

PROPOSITION 2.1. *Let $\mu_n(\omega, \cdot) = \text{Gamma}(\alpha_n, \beta_n)$ where $\alpha_n \rightarrow \infty$ is a deterministic sequence but $\{\beta_n\}$ are random variables. Then $\delta(\mu_n) = o_{\mathbb{P}}(1)$ if and only if $\alpha_n / \beta_n^2 = o_{\mathbb{P}}(1)$.*

PROOF. By the estimate $\log(1 + x) \leq x$, we have

$$(2.1) \quad |\varphi(\xi; \mu_n)|^2 = |1 - i\xi/\beta_n|^{-2\alpha_n} = (1 + \xi^2/\beta_n^2)^{-\alpha_n} \geq \exp(-\xi^2 \alpha_n / \beta_n^2)$$

that yields the sufficiency of the condition, since by Fubini's theorem and the dominated convergence theorem, we have

$$\mathbb{E} \left[\left\{ \int |\varphi(\xi; \mu_n)|^2 \frac{d\xi}{\pi(1 + \xi^2)} \right\} \right] \geq \int \mathbb{E}[\exp(-\xi^2 \alpha_n / \beta_n^2)] \frac{d\xi}{\pi(1 + \xi^2)} \rightarrow 1.$$

Now assume $\alpha_n/\beta_n^2 \neq o_{\mathbb{P}}(1)$. By choosing subsequence of $\{\alpha_n, \beta_n\}_n$, we may assume $\mathbb{P}(\alpha_n/\beta_n^2 \geq \epsilon) \rightarrow \eta > 0$ for some $\epsilon, \eta > 0$. If $\alpha_n/\beta_n^2 \geq \epsilon$, then the left-hand of (2.1) is dominated by $(1 + \xi^2\epsilon/\alpha_n)^{-\alpha_n}$. Hence

$$\begin{aligned} & \mathbb{E} \left[\left\{ \int |\varphi(\xi; \mu_n)|^2 \frac{d\xi}{\pi(1 + \xi^2)} \right\} \right] \\ & \leq 1 - \mathbb{P}(\alpha_n/\beta_n^2 \geq \epsilon) \left(1 - \int (1 + \xi^2\epsilon/\alpha_n)^{-\alpha_n} \frac{d\xi}{\pi(1 + \xi^2)} \right) \\ & \rightarrow 1 - \eta \left(1 - \int e^{-\xi^2\epsilon} \frac{d\xi}{\pi(1 + \xi^2)} \right) < 1. \end{aligned}$$

Hence $\int |\varphi(\xi; \mu_n)|^2 d\xi/\pi(1 + \xi^2) \neq 1 + o_{\mathbb{P}}(1)$ and hence $\delta(\mu_n) \neq o_{\mathbb{P}}(1)$. $\square$

2.2. The degeneracy

Let $\{(X_n, \mathcal{X}_n, P_n)\}$ be a sequence of probability spaces for the “real” observations and write x_n for an element of X_n . Let an open subset $\Theta \subset \mathbb{R}^d$ be the parameter space and let $P_n(d\theta | x_n)$, the “posterior distribution”, be the transition kernel from X_n to Θ . The canonical product $\Theta^{\mathbb{N}}$ is the space for Monte Carlo simulations. After we obtain an observation x_n from X_n , we perform a simulation that results in a sequence $\theta_{\infty} = (\theta_0, \theta_1, \dots)$ as an element of $\Theta^{\mathbb{N}}$. After this simulation, we will perform inference and estimation. For example, the empirical mean $M^{-1} \sum_{m=0}^{M-1} \theta_m$ is commonly used for an estimate of $\int_{\Theta} \theta P_n(d\theta | x_n)$.

To introduce an efficiency for the Monte Carlo simulation, we extend the probability space. Let $\Omega_n = X_n \times \Theta^{\mathbb{N}}$ and $\mathcal{F}_n$ be the corresponding product σ -algebra. Each element of Ω_n is denoted by $\omega = (x_n, \theta_{\infty}) = (x_n, \theta_0, \theta_1, \dots)$. A probability measure $\mathbb{P}_n$ is assumed to be the following form:

$$\mathbb{P}_n(d\omega) = P_n(dx_n)P_n(d\theta_{\infty} | x_n).$$

We assume the stationarity condition; the sequence θ_{∞} is a stationary process under $P_n(d\theta_{\infty} | x_n)$ for any x_n , and its invariant distribution is $P_n(d\theta | x_n)$. Note that it is possible to apply our results to a non-stationarity process, though we do not discuss it in this paper. See Kamatani (2014a) for the details.

Let us define efficiency and inefficiency. Write $C_0(\Theta)$ for the space of the $\mathbb{R}$ -valued continuous functions with a compact support on Θ . In Kamatani (2014a), the following property is called the consistency of $\{P_n(d\theta_{\infty} | x_n)\}_n$; for any $M_n \rightarrow \infty$ and for any $\psi \in C_0(\Theta)$,

$$(2.2) \quad \frac{1}{M_n} \sum_{m=0}^{M_n-1} \psi(\theta_m) - \int \psi(\theta) P_n(d\theta | x_n) = o_{\mathbb{P}_n}(1).$$

In practice, sometimes the evaluation of the above integral is difficult, but that of the empirical mean $\frac{1}{M_n} \sum_{m=0}^{M_n-1} \psi(\theta_m)$ is easy. The property (2.2) says that this easy calculation (the first term) provides a good approximation of the second term, which we want to calculate. However, in the present study, we are

interested in another property, the degeneracy, proposed by Kamatani (2014b); for any $M \in \mathbb{N}$ and $\psi \in C_0(\Theta)$,

$$(2.3) \quad \frac{1}{M} \sum_{m=0}^{M-1} \psi(\theta_m) - \psi(\theta_0) = o_{\mathbb{P}_n}(1).$$

This property says that the approximation using m sequences is no more helpful than that using only one sequence. As we see later, unfortunately, many MCMC algorithms satisfy this condition and so it has great importance for the study of (2.3).

The following two propositions were proved in Kamatani (2014b) with slightly different notation. The first property states that the degeneracy is merely a relation of each adjacent pair in θ_∞ .

PROPOSITION 2.2. *Assume $\mathbb{E}_n[P_n(d\theta \mid x_n)] =: P_n(d\theta)$ is tight. Then the degeneracy is equivalent to $\theta_1 - \theta_0 = o_{\mathbb{P}_n}(1)$, and it is also equivalent to*

$$(2.4) \quad \delta(\theta_1 - \theta_0) = o_{\mathbb{P}_n}(1).$$

PROOF. According to Kamatani (2014b), the degeneracy is equivalent to $\theta_1 - \theta_0 = o_{\mathbb{P}_n}(1)$. Therefore necessity of (2.4) follows by Lemma 2.2 (iii) where $\mathcal{G}_n = \{\emptyset, \Omega\}$ for this case. On the other hand, assume (2.4). Then $\theta_1 - \theta_0 = u_n + o_{\mathbb{P}_n}(1)$ for some sequence $\{u_n\}$ of Θ . By tightness of θ_0 and θ_1 , the sequence $\{u_n\}$ is bounded. It is sufficient to show that $u_n \rightarrow 0$. If it is not, we may assume $u_n \rightarrow v \neq 0$ for some $v \in \Theta$ by the Bolzano-Weierstrass theorem, and we can also assume $P_n \rightarrow P$ for a probability measure P by tightness. By stationarity, $P_n(d\theta_0) = P_n(d\theta_1)$. However, since $\theta_1 = \theta_0 + u_n + o_{\mathbb{P}_n}(1)$, we have $P = P(\cdot - v)$ but it can not be true since P is a probability measure. (To see this, observe $P(A) = P(A - v) = P(A - 2v) = \dots$) $\square$

The second proposition proves that essentially, the consistency and degeneracy are mutually exclusive; if both hold, the dispersion of the posterior distribution tends to 0 in probability. Recall that $\delta(\theta_0 \mid x_n)$ is the dispersion of the law $P_n(d\theta_0 \mid x_n)$.

PROPOSITION 2.3 (Consistency and degeneracy are mutually exclusive). *Assume the same condition as Lemma 2.2. If both the consistency and the degeneracy hold, then*

$$(2.5) \quad \delta(\theta_0 \mid x_n) = o_{\mathbb{P}_n}(1).$$

PROOF. For any $\psi \in C_0(\Theta)$, we can choose $M_n \rightarrow \infty$ so that (2.3) holds for M_n in place of M . By (2.2),

$$\psi(\theta_0) = \int \psi(\theta) P_n(d\theta \mid x_n) + o_{\mathbb{P}_n}(1).$$

We show that we can take $\psi(\theta) = \exp(i\xi^t\theta) = \cos(\xi^t\theta) + i\sin(\xi^t\theta)$, though this $\psi(\theta)$ is not in $C_0(\Theta)$. To see this, for any $\epsilon > 0$, choose a compact set K such that $\limsup_{n \rightarrow \infty} P_n(K^c) < \epsilon/2$ and $\rho \in C_0(\Theta)$ such that $\rho(\theta) = 1$ on $\theta \in K$ and $|\rho(\theta)| \leq 1$ ($\theta \in \Theta$). Then for $\psi_K(\theta) := \rho(\theta)\psi(\theta) \in C_0(\Theta)$,

$$\begin{aligned} \limsup_{n \rightarrow \infty} \mathbb{E}_n \left[\left| \psi(\theta_0) - \int \psi(\theta) P_n(d\theta | x_n) \right| \right] \\ \leq \epsilon + \limsup_{n \rightarrow \infty} \mathbb{E}_n \left[\left| \psi_K(\theta_0) - \int \psi_K(\theta) P_n(d\theta | x_n) \right| \right] = \epsilon \end{aligned}$$

and hence we can take $\psi(\theta) = \exp(i\xi^t\theta)$. Thus

$$|\mathbb{E}_n[\exp(i\xi^t\theta) | x_n]| = \left| \int \psi(\theta) P_n(d\theta | x_n) \right| = |\psi(\theta_0)| + o_{\mathbb{P}_n}(1) = 1 + o_{\mathbb{P}_n}(1).$$

Hence $\mathbb{E}[\int |\mathbb{E}_n[\exp(i\xi^t\theta) | x_n]|^2 \prod_{j=1}^d (\pi(1 + \xi_j^2))^{-1} d\xi] \rightarrow 1$ by Fubini's theorem. This proves the claim. $\square$

2.3. Degeneracy for some MCMC procedures

We consider the degeneracy properties for three kinds of MCMC procedures; the DA procedure, MH procedure and hybrid procedure. First we consider the DA procedure. Let $(X_n, \mathcal{X}_n, P_n)$ be a probability space and let $(Y_n, \mathcal{Y}_n)$ be a measurable space. This space is for the latent variables. For the observation $x_n \in X_n$, the latent variable $y_n \in Y_n$ and the parameter $\theta \in \Theta \subset \mathbb{R}^d$, we introduce five transition kernels in one equation by

$$\begin{aligned} (2.6) \quad P_n(d\theta dy_n | x_n) &= P_n(d\theta | x_n, y_n) P_n(dy_n | x_n) \\ &= P_n(dy_n | \theta, x_n) P_n(d\theta | x_n). \end{aligned}$$

Write $P_n(dx_n dy_n) = P_n(dy_n | x_n) P_n(dx_n)$. Let θ_0 be any element of Θ , called the initial guess, not a “true parameter”. For an observation x_n with the initial guess $\theta = \theta_0$, consider the iteration of the following:

$$\begin{cases} y_n \sim P_n(dy_n | \theta, x_n), \\ \theta \sim P_n(d\theta | x_n, y_n) \end{cases}$$

where x_n is fixed throughout but θ and y_n are updated simultaneously. Iteration of the above results in a sequence

$$\theta_\infty^* = (\theta_0, y_n^1, \theta_1, y_n^2, \dots)$$

and write its law by $P_n(d\theta_\infty^* | x_n)$. Accordingly, we extend the probability space $(X_n, \mathcal{X}_n, P_n)$ to $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ where

$$\mathbb{P}_n(d\omega) = P_n(dx_n) P_n(d\theta_\infty^* | x_n), \quad \omega = (x_n, \theta_\infty^*) \in X_n \times (\Theta \times Y_n)^{\mathbb{N}} = \Omega_n.$$

By (2.6), if $\theta_0 \sim P_n(d\theta \mid x_n)$, then $\theta_\infty = (\theta_0, \theta_1, \dots)$ is a $P_n(d\theta_\infty^* \mid x_n)$ -stationary process. To show this, by Markov property, it is sufficient to check that the law of θ_1 given x_n is $P_n(d\theta \mid x_n)$. For this,

$$(2.7) \quad \begin{aligned} P_n(d\theta_1 \mid x_n, y_n) P_n(dy_n \mid \theta_0, x_n) P_n(d\theta_0 \mid x_n) \\ = P_n(d\theta_0 \mid x_n, y_n) P_n(dy_n \mid \theta_1, x_n) P_n(d\theta_1 \mid x_n) \end{aligned}$$

by integrating out θ_0 and y_n . In this paper, we call $\{P_n(d\theta_\infty \mid x_n)\}$, which is the conditional distribution of θ_∞ given x_n , the data augmentation (DA) procedure. The following sharpens the results of Lemma 2.2. We will write y_n for y_n^1 . Recall that $\delta(\theta_1 \mid x_n, y_n)$ is the dispersion of the probability measure $P_n(d\theta_1 \mid x_n, y_n)$.

LEMMA 2.3 (Degeneracy for DA procedure (Kamatani (2014b))). *Assume the same condition as Lemma 2.2. Then for the DA procedure, the degeneracy is equivalent to*

$$(2.8) \quad \delta(\theta_1 \mid x_n, y_n) = o_{\mathbb{P}_n}(1).$$

PROOF. If (2.8), there exists $\mathcal{X}_n \times \mathcal{Y}_n$ -measurable u_n such that $\theta_1 - u_n = o_{\mathbb{P}_n}(1)$. By (2.7), $\theta_0 - u_n$ should have the same law as $\theta_1 - u_n$, and hence $o_{\mathbb{P}_n}(1)$. Thus $\theta_1 - \theta_0 = (\theta_1 - u_n) - (\theta_0 - u_n) = o_{\mathbb{P}_n}(1)$, and hence the degeneracy follows by (2.2). Necessity of (2.8) is clear by $\delta(\theta_1 \mid x_n) = o_{\mathbb{P}_n}(1)$ with Lemma 2.2 (ii). $\square$

Second, we consider the degeneracy of the Metropolis-Hastings (MH) procedure on $(X_n, \mathcal{X}_n, P_n)$. Recall that the MH procedure is a kind of MCMC procedure that has the accept-reject step; for $u_1, u_2, \dots$ from uniform distribution $U[0, 1]$, set

$$\theta_{m+1} = \begin{cases} \tilde{\theta}_{m+1} & \text{if } u_{m+1} \leq \alpha(\theta_m, \tilde{\theta}_{m+1} \mid x_n) \\ \theta_m & \text{if } u_{m+1} > \alpha(\theta_m, \tilde{\theta}_{m+1} \mid x_n) \end{cases}$$

where a function $(\theta_1, \theta_2, x_n) \mapsto \alpha(\theta_1, \theta_2 \mid x_n) \in [0, 1]$ is designed to hold the so-called balanced condition:

$$(2.9) \quad \begin{aligned} E_n[\alpha(\theta_m, \tilde{\theta}_{m+1} \mid x_n), \{\tilde{\theta}_{m+1} \in A, \theta_m \in B\} \mid x_n] \\ = E_n[\alpha(\tilde{\theta}_{m+1}, \theta_m \mid x_n), \{\theta_m \in A, \tilde{\theta}_{m+1} \in B\} \mid x_n]. \end{aligned}$$

We call $\tilde{\theta}_{m+1}$ the propose of an iteration from θ_m . As in the DA procedure, consider $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ where

$$\omega = (x_n, \theta_\infty^*) = (x_n, \theta_0, u_1, \tilde{\theta}_1, \theta_1, u_2, \dots) \in X_n \times (\Theta \times [0, 1] \times \Theta)^{\mathbb{N}} = \Omega_n$$

where $\mathcal{F}_n$ is the corresponding product σ -algebra. We assume (2.9) for $\mathbb{P}_n(d\omega) = P_n(dx_n)P_n(d\theta_\infty^* \mid x_n)$. Also the stationarity condition is assumed, that is, the law of θ_∞ given x_n is stationary.

LEMMA 2.4 (Degeneracy for MH procedure). *If $|\tilde{\theta}_1 - \theta_0| = o_{\mathbb{P}_n}(1)$, then the MH procedure is degenerate. On the other hand, if*

$$\lim_{\epsilon \rightarrow 0} \limsup_{n \rightarrow \infty} \mathbb{P}_n(\alpha(\theta_0, \tilde{\theta}_1 \mid x_n) \leq \epsilon) = 0$$

and $|\tilde{\theta}_1 - \theta_0| \neq o_{\mathbb{P}_n}(1)$, then the MH procedure is not degenerate.

PROOF. The first claim is clear since $|\theta_1 - \theta_0| \leq |\tilde{\theta}_1 - \theta_0|$. For the second, since $\alpha \geq \epsilon(1 - 1_{\{\alpha \leq \epsilon\}})$,

$$\begin{aligned} \mathbb{E}_n[\min\{1, c_n|\theta_1 - \theta_0|\}] &= \mathbb{E}_n[\min\{1, c_n|\tilde{\theta}_1 - \theta_0|\}\alpha(\tilde{\theta}_0, \theta_1)] \\ &\geq \epsilon(\mathbb{E}_n[\min\{1, c_n|\tilde{\theta}_1 - \theta_0|\}] - \mathbb{P}_n(\alpha(\theta_0, \tilde{\theta}_1) \leq \epsilon)). \end{aligned}$$

Choose a subsequence $\{n_i\}$ and $\delta > 0$ such that $\lim_{i \rightarrow \infty} \mathbb{E}_{n_i}[\min\{1, c_{n_i}|\tilde{\theta}_1 - \theta_0|\}] = \delta$ and find $\epsilon > 0$ to be $\limsup_{n \rightarrow \infty} \mathbb{P}_n(\alpha(\theta_0, \tilde{\theta}_1) \leq \epsilon) < \delta/2$. Then $\lim_{i \rightarrow \infty} \mathbb{E}_{n_i}[\min\{1, c_{n_i}|\tilde{\theta}_1 - \theta_0|\}] \geq \epsilon\delta/2$ proves the latter claim. $\square$

Third, we consider a hybrid scheme on $(X_n, \mathcal{X}_n, P_n)$. Let $\Omega_n = X_n \times \Theta^{\mathbb{N}}$ and let $\mathcal{F}_n$ be the corresponding product σ -algebra. Suppose that there are sequence of probability measures

$$\mathbb{P}_{n,k}(d\omega) = P_n(dx_n)P_{n,k}(d\theta_\infty \mid x_n) \quad (k = 1, \dots, K)$$

on $(\Omega_n, \mathcal{F}_n)$ but θ_∞ is $P_{n,k}(d\theta_\infty \mid x_n)$ -stationary for each $k = 1, \dots, K$ with the same stationary distribution $P_n(d\theta \mid x_n)$. We assume that $\mathbb{P}_n(d\omega) = P_n(dx_n)P_n(d\theta_\infty \mid x_n)$ is a mixture of these probability measures in the following sense:

$$(2.10) \quad P_n((\theta_0, \theta_1) \in A \mid x_n) = \frac{1}{K} \sum_{k=1}^K P_{n,k}((\theta_0, \theta_1) \in A \mid x_n), \quad (A \in \mathcal{B}(\Theta)^{\otimes 2}).$$

The following lemma states that the hybrid scheme is not worse than each of the components $k = 1, \dots, K$ in the sense of the degeneracy.

LEMMA 2.5 (Degeneracy for hybrid procedure). *The sequence $\{P_n(d\theta_\infty \mid x_n)\}_n$ is degenerate if and only if $\{P_{n,k}(d\theta_\infty \mid x_n)\}_n$ are degenerate for all $k = 1, \dots, K$.*

PROOF. Degeneracy is equivalent to $\theta_1 - \theta_0 = o_{\mathbb{P}_n}(1)$. Thus the following proves the claim:

$$\mathbb{P}_n(|\theta_1 - \theta_0| > \epsilon) = \frac{1}{K} \sum_{k=1}^K \mathbb{P}_{n,k}(|\theta_1 - \theta_0| > \epsilon). \quad \square$$

2.4. Localization

Let $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ as in Subsection 2.2. Let $C^1(\Theta)$ be the space of the $\mathbb{R}$ -valued continuously differentiable functions. As in Doob (1949), under fairly general conditions, the posterior distribution has the consistency property, that is, (2.5) holds as the sample size n grows. For this fact, the consistency and degeneracy defined in Subsection 2.2 are useless in a Bayesian context, so it is reasonable to consider those properties after scaling and mapping

$$\theta \mapsto n^{1/2}(\varphi(\theta) - \varphi(u_n)) =: \vartheta$$

so as

$$(2.11) \quad \delta(\vartheta \mid x_n) \neq o_{\mathbb{P}_n}(1)$$

where $\varphi \in C^1(\Theta)$ and u_n is a particular $\mathcal{X}_n$ -measurable estimator such as the maximum likelihood estimator. We say that $\{P_n(d\theta_\infty \mid x_n)\}_n$ has the φ -local consistency/degeneracy if $\{P_n(d\vartheta_\infty \in \cdot \mid x_n)\}$ is consistent/degenerate, where

$$\vartheta_m := n^{1/2}(\varphi(\theta_m) - \varphi(u_n)), \quad m = 0, 1, \dots, \quad \text{and} \quad \vartheta_\infty := (\vartheta_0, \vartheta_1, \dots).$$

Since ϑ_∞ is still a stationary process, we can apply Lemmas 2.2–2.5. If φ is the identity map (in this case, φ becomes $\mathbb{R}^d$ -valued, though), we simply call them local consistency/degeneracy. Another remark is that the local consistency/degeneracy property is preserved by the map φ . See Kamatani (2014b) for a proof.

LEMMA 2.6 (Local consistency/degeneracy is preserved by the map). *Assume the law of u_n and $\tilde{P}_n(A) := \mathbb{E}_n[P_n(\sqrt{n}(\theta - u_n) \in A \mid x_n)]$ are tight. Then if local consistency/degeneracy holds, it has the φ -local consistency/degeneracy.*

In this paper we define the following, as a property of MCMC, which is “not so bad”. The “good” property, local consistency is a sufficient condition by Lemma 2.6. We will show that our proposed algorithm satisfies this condition.

DEFINITION 2.4.1 (Local non-degeneracy). *The sequence $\{P_n(d\theta_\infty \mid x_n)\}$ is said to have the local non-degeneracy if it has the φ -local degeneracy only if $\delta(n^{1/2}\varphi(\theta_0) \mid x_n) = o_{\mathbb{P}_n}(1)$ for $\varphi \in C^1(\Theta)$.*

We briefly mention the behavior of the MCMC procedure under the local non-degeneracy condition. Auto-correlation plots are commonly used as a measure of performance of the MCMC procedure. If the plots drop towards zero quickly, it is interpreted that the MCMC procedure works well. However, under suitable conditions, if the local degeneracy holds, the auto-correlation does not drop at all since each θ_i produces almost the same value. On the other hand, under the local non-degeneracy condition, this cannot happen. In this sense, roughly, the local non-degeneracy is a sign of a good convergence.

By Lemma 2.6, the local consistency is a sufficient condition for the local non-degeneracy, and hence it is always preferred. However sometimes it requires tedious work to check (2.2). For such a case, the non-degeneracy property is appealing.

2.5. Examples

Let Θ be a parametric space. As in (2.7), several transition kernels will be introduced. We further introduce a transition kernel $P_n(dx_n dy_n \mid \theta) = P_n(dx_n \mid \theta)P_n(dy_n \mid \theta, x_n)$.

Example 2.1 (DA procedure for the probit model). Consider the models (1.2) and (1.3) in the introduction. The DA procedure is the iteration of the following two steps; given $\theta = \theta_0$,

$$\begin{cases} \text{simulate } y^i \sim \begin{cases} N(0, 1, -\infty, \theta) & \text{if } x^i = 1 \\ N(0, 1, \theta, \infty) & \text{if } x^i = 2 \end{cases} & \text{for } i = 1, \dots, n, \\ \text{then simulate } \theta \sim N\left(0, 1, \max_{i; x^i=1} y^i, \min_{i; x^i=2} y^i\right) \end{cases}$$

where $N(0, 1, a, b)$ is the standard normal distribution truncated to $[a, b]$. It is easy to show the Bernstein-von Mises's theorem for (1.2). Hence (2.11) holds by Lemma 2.2 (iii). For $u_n^\dagger = \max_{i; x^i=1} y^i$, (2.8) is satisfied since

$$(2.12) \quad n|\theta_1 - u_n^\dagger| \leq n \left(\min_{i; x^i=1} y^i - \theta_0 \right) + n \left(\theta_0 - \max_{i; x^i=2} y^i \right) = O_{P_n(dx_n dy_n | \theta_0)}(1)$$

for any θ_0 ; to see this, take $M = -\log(1 - \epsilon)/\phi(\theta_0)$ for $\epsilon > 0$, then by $n_1/n \rightarrow \Phi(\theta_0)$ for $n_1 = \sum_{i=1}^n 1_{\{x^i=1\}}$,

$$\begin{aligned} P_n \left(n \left(\min_{i; x^i=1} y^i - \theta_0 \right) > -M \mid \theta_0 \right) &= E_n \left[\left(\frac{\Phi \left(\theta_0 - \frac{M}{n} \right)}{\Phi(\theta_0)} \right)^{n_1} \mid \theta_0 \right] \\ &\rightarrow e^{-\phi(\theta_0)M} = 1 - \epsilon. \end{aligned}$$

Similar convergence holds for $n(\theta_0 - \max_{i; x^i=2} y^i)$. Thus by integrating out $\theta_0 \sim N(0, 1)$ together with Lemma 2.3, the DA procedure has the local degeneracy.

There are several procedures that improve the performance of the DA procedure. In this paper, we focus on the so-called parameter-expanded data-augmentation (PX-DA) procedure. It introduces one working parameter γ . We slightly extend the probability space $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ considered in the previous example to handle the sequence of γ . For simplicity, however, we still use the same notation.

Example 2.2 (PX-DA procedure for the probit model). Consider the model (1.2). Then the PX-DA procedure becomes

$$\begin{cases} \text{simulate } \tilde{\theta}_{m+1} \text{ from } \theta_m \text{ by the DA procedure} \\ \text{simulate } \gamma_{m+1}^2 \sim \text{Gamma} \left(\frac{n+1}{2}, \frac{\sum_{i=1}^n (y^i)^2 + \tilde{\theta}_{m+1}^2}{2} \right), \\ \text{set } \theta_{m+1} \leftarrow \gamma_{m+1} \tilde{\theta}_{m+1}. \end{cases}$$

We show that the PX-DA procedure does not suffer from the local degeneracy. Suppose for contradiction that it has the local degeneracy. Then $n^{1/2}(\theta_1 - \theta_0) = o_{\mathbb{P}_n}(1)$, so $n^{1/2}(\theta_1^2 - \theta_0^2) = o_{\mathbb{P}_n}(1)$ by Lemma 2.2 (v). By conditioning on $\tilde{\theta}_1, \theta_0, y_n$, we have

$$\delta(n^{1/2}\gamma_1^2\tilde{\theta}_1^2 \mid \tilde{\theta}_1, y_n) = o_{\mathbb{P}_n}(1)$$

by Lemma 2.2 (ii). Recall that $\tilde{\theta}_1$ follows $P_n(d\theta \mid x_n)$ and hence by integrating out x_n , it follows the prior distribution $N(0, 1)$. Hence $\tilde{\theta}_1^{-2}$ is tight, and we have $\delta(n^{1/2}\gamma_1^2 \mid \tilde{\theta}_1, y_n) = o_{\mathbb{P}_n}(1)$ by Lemma 2.2 (v). However it is a contradiction of Proposition 2.1. Indeed, we can apply the proposition to $\alpha_n = (n+1)/2$ and $\beta_n = (\sum_{i=1}^n (y^i)^2 + \tilde{\theta}_1^2)/2\sqrt{n} = \sqrt{n}(1 + o_{\mathbb{P}_n}(1))/2$.

The PX-DA procedure does not always work well. Consider

$$P_\theta(x=1) = \Phi(\theta_1), \quad P_\theta(x=2) = \Phi(\theta_2) - \Phi(\theta_1), \quad P_\theta(x=3) = 1 - \Phi(\theta_2).$$

Write $\theta_m = (\theta_{m1}, \theta_{m2})$.

Example 2.3 (Degenerate PX-DA procedure). For the above model, the DA procedure is

$$\text{simulate } y^i \sim \begin{cases} N(0, 1, -\infty, \theta_{m1}) & \text{if } x^i = 1 \\ N(0, 1, \theta_{m1}, \theta_{m2}) & \text{if } x^i = 2 \\ N(0, 1, \theta_{m2}, \infty) & \text{if } x^i = 3 \end{cases},$$

$$\text{then simulate } \theta_{m+1,j} \sim N\left(0, 1, \max_{i:x^i=j} y^i, \min_{i:x^i=j+1} y^i\right) \quad (j=1, 2)$$

The PX-DA procedure adds a simulation of γ^2 to the DA procedure, and the output becomes $\gamma\theta$ instead of θ . The DA produces have the local degeneracy by the same argument as in Example 2.1; on the other hand, the PX-DA might look fine in simulation. However, $\varphi(\theta_m) = \theta_{m1}/\theta_{m2}$ cancels out the benefit of this additional step, and thus the PX-DA procedure has the φ -local degeneracy.

3. Application to a cumulative probit model

3.1. The model and setting

In Section 2, x was the observation, and y was the latent variable. In this section, the observations are both $x \in X = \mathbb{R}^p$ and $y \in Y = \{1, \dots, c\}$, and the latent variable is $z \in Z = \mathbb{R}$, and so the notation becomes slightly different. Let $M = \{P_\theta(dxdy) = P(dxdy \mid \theta)\}$ be the parametric family defined in (1.1) with $x \sim G(dx)$, and set $P_n(dx_ndy_n \mid \theta) = P_\theta^{\otimes n}(dx_ndy_n) = \prod_{i=1}^n P_\theta(dx^i dy^i)$, $x_n = (x^1, \dots, x^n)$ and $y_n = (y^1, \dots, y^n)$. We assume that G has a compact support and has the full rank, that is, there exists a linearly independent sequence $x_1, \dots, x_p \in \mathbb{R}^p$ in this support. Suppose also that the prior $P(d\theta)$ is the standard normal distribution truncated to $\Theta = \{\theta = (\alpha, \beta), 0 =: \alpha_1 < \dots < \alpha_{c-1} < \alpha_c := +\infty, \beta \in \mathbb{R}^p\}$.

Let $d = p + c - 2$ be the number of parameters. For two matrices A and B , we write $A \geq B$ if $A - B$ is positive semi-definite. Recall that from the Fisher information matrices $I(\theta)$ and $J(\theta)$ of statistics U and $t(U)$, we have the relation $I(\theta) \geq J(\theta)$ for any map t . Let $\|\nu\| = \sup_{A \in \mathcal{B}(\mathbb{R}^d)} |\nu(A)|$ be the total variation. Let $P_n(d\theta | x_n, y_n)$ be the posterior distribution, and let $P_n(dx_n dy_n) = \int P_n(dx_n dy_n | \theta) P(d\theta)$. We first obtain the following non-degeneracy of the posterior distribution.

LEMMA 3.7. *The Bernstein-von Mises's theorem holds, that is, $\|P_n(d\theta | x_n, y_n) - N(\hat{\theta}_n, n^{-1}I(\theta_0)^{-1})\| = o_{P_{\theta_0}^{\otimes n}}(1)$ for the Fisher information matrix $I(\theta)$ and an estimator $\hat{\theta}_n = \hat{\theta}_n(x_n, y_n) = (\hat{\alpha}_n, \hat{\beta}_n)$ such that $\sqrt{n}(\theta - \hat{\theta}_n)$ tends in $P_{\theta_0}^{\otimes n}$ -law to $N(0, I(\theta_0)^{-1})$. Moreover, for $\varphi \in C^1(\Theta)$, $\delta(n^{1/2}\varphi(\theta_0) | x_n y_n) = o_{P_n(dx_n, dy_n)}(1)$ if and only if φ is a constant.*

PROOF. We first check the sufficient conditions for the Bernstein-von Mises's theorem in Section 8.4 of Le Cam and Yang (2000). The existence of uniformly consistent test (A3) was checked in the appendix of Kamatani (2014b) and the quadratic mean differentiability is easy to check, so we only check the non singularity of the Fisher information matrix $I(\theta)$.

Assume now that for $j \in \{1, \dots, c-1\}$, we only observe x and 1 or 0, corresponding to the events $\{y \leq j\}$ or $\{y > j\}$. For this restricted model, let $I_j(\theta)$ be the Fisher information matrix. Then $I(\theta) \geq I_j(\theta)$. It is straightforward to check that $I_1(\theta)$ is non singular with respect to the parameter β , and $I_j(\theta)$ is also non singular with respect to α_j for $j = 2, \dots, c-1$. Hence $I(\theta)$ should be non singular, and the Bernstein-von Mises's theorem holds.

Next we show the latter claim. The sufficiency of $\varphi = \text{const.}$ is clear. Assume now that φ is not a constant. Then the derivative $\varphi'(\theta) \neq 0$ in an open subset U of Θ . Hence by the Bernstein-von Mises's theorem, the law of $n^{1/2}\varphi(\theta_0)$ tends to a normal distribution with variance $\Sigma(\theta) := \varphi'(\theta)^t I(\theta)^{-1} \varphi'(\theta) > 0$ under $P_n(dx_n dy_n | \theta)$. Therefore $\delta(n^{1/2}\varphi(\theta_0) | x_n, y_n)$ tends in probability to $\delta(N(0, \Sigma(\theta))) > 0$ under $P_n(dx_n dy_n | \theta)$ on $\theta \in U$. Therefore, integrating out θ , we have $\delta(n^{1/2}\varphi(\theta_0) | x_n, y_n) \neq o_{P_n(dx_n, dy_n)}(1)$. Hence the claim follows. $\square$

The DA procedure implicitly uses the following model:

$$z \sim N(-\beta^t x, 1), \quad x \sim G, \quad y = j \quad \text{if } z \in (\alpha_{j-1}, \alpha_j].$$

Construction of the DA procedure is straightforward, and hence we omit it to save the space; see Kamatani (2014b), Liu and Sabatti (2000) for the details. The DA procedure resulted in a sequence $(\theta_0, \theta_1, \dots)$ where

$$\theta_m = (\alpha_m, \beta_m), \quad \alpha_m = (\alpha_{mj}; j \in J)$$

where $J = \{2, \dots, c-1\}$. We check local degeneracy of the DA procedure.

LEMMA 3.8. *For $\varphi \in C^1(\Theta)$, the DA procedure has the φ -local degeneracy if and only if $\partial\varphi/\partial\beta \equiv 0$.*

PROOF. First we note tightness of $\sqrt{n}(\theta_1 - \theta_0)$. By Lemma 3.7 with stationarity, both $\sqrt{n}(\theta_0 - \hat{\theta}_n)$ and $\sqrt{n}(\theta_1 - \hat{\theta}_n)$ are $\mathbb{P}_n$ -tight, and so $\sqrt{n}(\theta_1 - \theta_0) = \sqrt{n}(\theta_1 - \hat{\theta}_n) + \sqrt{n}(\hat{\theta}_n - \theta_0)$ is $\mathbb{P}_n$ -tight.

By Lemma 2.2, the φ -local degeneracy is equivalent to $\delta(n^{1/2}(\varphi(\theta_1) - \varphi(\theta_0))) = o_{\mathbb{P}_n}(1)$. Hence if $\partial\varphi/\partial\beta = 0$, by Taylor's expansion, together with tightness of $\sqrt{n}(\theta_1 - \theta_0)$ and Lemma 2.2 (iv), it is also equivalent to

$$(3.1) \quad \delta \left(n^{1/2} \frac{\partial\varphi(\theta_0)}{\partial\alpha} (\alpha_1 - \alpha_0) \right) = o_{\mathbb{P}_n}(1).$$

However, by the same argument as in (2.12), $\delta(n^{1/2}(\alpha_1 - \alpha_0)) = o_{P_n(dx_n, dy_n | \theta_0)}(1)$ and so $\delta(n^{1/2} \partial_\alpha \varphi(\theta_0) (\alpha_1 - \alpha_0)) = o_{P_n(dx_n, dy_n | \theta_0)}(1)$ by Lemma 2.2 (v). By integrating out θ_0 , (3.1) follows, and hence it has the φ -local degeneracy.

Suppose now that $\partial\varphi/\partial\beta \neq 0$ on an open set U . If it has the φ -local degeneracy, then by Lemma 2.2 (i) and (ii) with Taylor's expansion, we have

$$\begin{aligned} o_{\mathbb{P}_n}(1) &= \delta \left(n^{1/2} \frac{\partial\varphi(\theta_0)}{\partial\beta} (\beta_1 - \beta_0) + n^{1/2} \frac{\partial\varphi(\theta_0)}{\partial\alpha} (\alpha_1 - \alpha_0) \mid x_n, y_n, z_n, \theta_0 \right) \\ &\geq \delta \left(n^{1/2} \frac{\partial\varphi(\theta_0)}{\partial\beta} (\beta_1 - \beta_0) \mid x_n, y_n, z_n, \theta_0 \right). \end{aligned}$$

The conditional distribution of $n^{1/2}\beta_1$ follows the normal distribution with variance $n(I_p + \sum_{i=1}^n (x^i)^t x^i)^{-1}$ where I_p is the identity matrix. This variance converges to a non-degenerate matrix $\Sigma = (\int x^t x G(dx))^{-1}$ in probability under $P_n(dx_n, dy_n, dz_n \mid \theta_0)$. Hence $\delta(n^{1/2}(\beta_1 - \beta_0) \mid x_n, y_n, z_n, \theta_0)$ converges to $\delta(N(0, \Sigma)) > 0$ for $\theta_0 \in U$. This contradicts Lemma 2.2 (v), and hence the claim follows. $\square$

3.2. Construction of the HPX-DA-MH procedure

By Lemma 3.8, DA can be locally degenerate. The PX-DA procedure adds a simple step to the DA procedure:

$$\left\{ \begin{array}{l} \text{simulate } \theta_{m+1}^* \text{ from } \theta_m \text{ by using DA procedure} \\ \text{simulate } \gamma_{m+1}^2 \sim \text{Gamma} \left(\frac{n+d}{2}, \right. \\ \quad \left. \frac{1}{2} \left(\sum_{i=1}^n (z^i + (\beta_{m+1}^*)^t x^i)^2 + |\beta_{m+1}^*|^2 + |\alpha_{m+1}^*|^2 \right) \right) \\ \text{set } \theta_{m+1} = \gamma_{m+1} \theta_{m+1}^* \end{array} \right.$$

By Kamatani (2014b), we already know that the PX-DA procedure can not escape from local degeneracy. We define its extension, the HPX-DA-MH procedure. For the definition, we first prepare sub models of M .

For $k \in J = \{2, \dots, c-1\}$, let

$$\alpha^k = (\alpha_j; j \in J_k), \quad \theta^k := (\alpha^k, \beta)$$

where $J_k = J \setminus \{k\}$. We will use “ k ” as the map $\theta \mapsto \theta^k$. Let $y_k = y - 1_{\{y > k\}}$, $y_k^i = y^i - 1_{\{y^i > k\}}$ and $y_{n,k} = \{y_k^i\}_{i=1,\dots,n}$. Using this observation (x, y_k) , we define a sub model. We denote the property of this sub model without proof.

LEMMA 3.9 (Projection into $c - 1$ category model). *If $(x, y) \sim P_\theta \in M$, then (x, y_k) follows the $c - 1$ -category cumulative probit model with parameter θ^k . The prior distribution is still the standard normal distribution with constraint condition $0 = \alpha_1 < \dots < \alpha_{k-1} < \alpha_{k+1} < \dots < \alpha_{c-1} < \alpha_c = +\infty$.*

For each $k \in J$, the totality of this $c - 1$ category model is denoted by $M_k = \{P_{\theta^k}\}$ and the prior in Lemma 3.9 is written as $P^k(d\theta^k)$. For example, if $c = 4$ and $k = 2$, P_{θ^2} is the 3-category model:

$$\begin{aligned} P_{\theta^2}(y_2 \leq 1 \mid x) &= \Phi(\alpha_1 + \beta^t x), \\ P_{\theta^2}(y_2 \leq 2 \mid x) &= \Phi(\alpha_3 + \beta^t x), \quad x \sim G(dx), \quad \theta^2 = (\alpha_1, \alpha_3, \beta). \end{aligned}$$

As described in Lemma 3.9, M_k can be embedded into M by a projection $P_\theta \mapsto P_{\theta^k}$ induced by $(x, y) \mapsto (x, y_k)$. Let

$$r_k(\theta \mid x_n, y_n) = \frac{dP_{\theta}^{\otimes n}}{dP_{\theta^k}^{\otimes n}} := \frac{\prod_{i=1}^n \Phi(\alpha_{y^i} + \beta^t x^i) - \Phi(\alpha_{y^i-1} + \beta^t x^i)}{\prod_{i=1}^n \Phi(\alpha_{y_k^i} + \beta^t x^i) - \Phi(\alpha_{y_k^i-1} + \beta^t x^i)},$$

where $\otimes n$ indicate the direct product, for example, $P_{\theta}^{\otimes n}(dx_n dy_n) = P_n(dx_n dy_n \mid \theta)$. Set the so-called acceptance ratio $\alpha_k(\theta, \tilde{\theta} \mid x_n, y_n) = \min\{1, r_k(\tilde{\theta} \mid x_n, y_n) / r_k(\theta \mid x_n, y_n)\}$. Call the following, the PX-DA-MH procedure for model M_k ; if the current parameter is $\theta_m = (\alpha_m, \beta_m)$ where $\alpha_m = (\alpha_{mj}; j \in J)$ then

$$\left\{ \begin{array}{l} \textbf{simulate } \tilde{\theta}_{m+1}^k \text{ from } \theta_m^k \text{ by using PX-DA procedure for the model } M_k \\ \textbf{set } \tilde{\theta}_{m+1} \text{ by attaching } \alpha_{mk} \text{ to } \tilde{\theta}_{m+1}^k \\ \textbf{set } \theta_{m+1} = \begin{cases} \tilde{\theta}_{m+1} & \text{with probability } \alpha_k(\theta_m, \tilde{\theta}_{m+1} \mid x_n, y_n) \\ \theta_m & \text{with probability } 1 - \alpha_k(\theta_m, \tilde{\theta}_{m+1} \mid x_n, y_n) \end{cases} \end{array} \right.$$

The HPX-DA-MH procedure is defined by the following iteration:

choose k uniformly from $\{2, \dots, c\}$

if $k < c$, **simulate** θ_{m+1} from θ_m by using the PX-DA-MH procedure for the model M_k

if $k = c$, **simulate** θ_{m+1} from θ_m by using the PX-DA procedure for the model M

The φ -local degeneracy of the PX-DA procedure is due to the linearity of $\theta \mapsto \gamma\theta$. On the other hand, since the HPX-DA-MH fixes α_k in each iteration, this linear relation vanishes. Hence it does produce a Markov chain with a good mixing property.

Remark 1. The PX-DA-MH procedure for model M_k does not update α_k .

Remark 2. For the Metropolis step, the procedure may propose a parameter that does not satisfy $\alpha_2 < \dots < \alpha_{c-1}$. This may occur before a sufficient number of iterations, when some categories have small counts. However, the number of incorrect proposals is usually very small (lower than 0.5% for each of our experiments).

3.3. Local non-degeneracy of the HPX-DA-MH procedure

Let $P_n(d\theta_\infty | x_n, y_n)$ be the law of a stationary Markov chain defined by the HPX-DA-MH procedure. Let $P_{n,k}(d\theta_\infty | x_n, y_n)$ be that defined by the PX-DA-MH procedure for the model M_k if $k \in J$, or the PX-DA procedure if $k = c$. Then $P_n(d\theta_\infty | x_n, y_n)$ is the mixture of $P_{n,k}(d\theta_\infty | x_n, y_n)$ in the sense of (2.10). Write $\omega = (x_n, y_n, \theta_\infty)$ and set

$$\begin{aligned}\mathbb{P}_n(d\omega) &= P_n(dx_n dy_n) P_n(d\theta_\infty | x_n, y_n), \\ \mathbb{P}_{n,k}(d\omega) &= P_n(dx_n dy_n) P_{n,k}(d\theta_\infty | x_n, y_n).\end{aligned}$$

For $k \in J$, write $\tilde{\theta}_1^k = \gamma_1 \theta_1^{k*}$ for the proposed value of the MH procedure, and $\theta_1^{k*} = (\alpha_1^{k*}, \beta_1^{k*})$ is the output of the DA procedure. To apply Lemma 2.4 to the HPX-DA-MH procedure, we prove the following.

LEMMA 3.10 (Tightness of some variables). *For $k \in J$, the following holds:*

- (i) $n^{1/2}|\theta_1^{k*} - \theta_0^k| = O_{\mathbb{P}_{n,k}}(1)$.
- (ii) $n^{1/2}|\alpha_1^{k*} - \alpha_0^k| = o_{\mathbb{P}_{n,k}}(1)$.
- (iii) $\sqrt{n}(\gamma_1 - 1)$ is $\mathbb{P}_{n,k}$ -tight, and hence $n^{1/2}|\tilde{\theta}_1^k - \theta_0^k|$ is $\mathbb{P}_{n,k}$ -tight.
- (iv) $\alpha_k(\theta_0, \tilde{\theta}_1 | x_n, y_n)^{-1}$ is $\mathbb{P}_{n,k}$ -tight.

PROOF. We omit “ k ” from $\tilde{\theta}_1^k$, θ_1^{k*} and θ_0^k for notational simplicity.

- (i) By Lemma 3.9, in fact θ_0 and θ_1^* are the outputs from the DA procedure for the model M_k with the prior $P^k(d\theta^k)$. Apply Lemma 3.7 to the model M_k with P^k . Then there exists $\hat{\theta}_n = \hat{\theta}_n(x_n, y_n, k)$ such that $\sqrt{n}(\theta_0 - \hat{\theta}_n) = O_{\mathbb{P}_{n,k}}(1)$. Thanks to the stationarity of the DA procedure, $\sqrt{n}(\theta_1^* - \hat{\theta}_n)$ has the same law, and hence it is also tight. Hence $\sqrt{n}(\theta_1^* - \theta_0) = \sqrt{n}(\theta_1^* - \hat{\theta}_n) + \sqrt{n}(\hat{\theta}_n - \theta_0)$ is $\mathbb{P}_{n,k}$ -tight.
- (ii) This proof is essentially the same as Example 2.1. We omit it.
- (iii) By (2.7), conditioned on θ_1^* , the law of z^i is $N(-(\beta_1^*)^t x_i, 1)$ and hence the marginal distribution of $z^i + (\beta_1^*)^t x_i$ is $N(0, 1)$. Thus $\sum_{i=1}^n (z^i + (\beta_{m+1}^*)^t x_i)^2 / n = 1 + O_{\mathbb{P}_{n,k}}(n^{-1/2})$ and hence $\gamma_1^2 \sim \text{Gamma}(\alpha_n, \beta_n)$ where $\alpha_n = n(1 + o(1))/2$ and $\beta_n = n(1 + O_{\mathbb{P}_{n,k}}(n^{-1/2}))/2$. Then its mean $\alpha_n/\beta_n = 1 + O_{\mathbb{P}_{n,k}}(n^{-1/2})$ and the variance $\alpha_n/\beta_n^2 = O_{\mathbb{P}_{n,k}}(1/n)$. Hence the first claim, the tightness of $\sqrt{n}(\gamma_1 - 1)$ follows. The latter claim is clear since $\tilde{\theta}_1 - \theta_0 = (\gamma_1 - 1)\theta_1^* + (\theta_1^* - \theta_0)$.
- (iv) Let $p_\theta(x, y)$ be the density of the model $P_\theta \in M$. It is not difficult to check that for any θ_0 and for any θ_1, θ_2 in a neighborhood of θ_0 , there exists a

bounded function l such that

$$|\log p_{\theta_1}(x, y) - \log p_{\theta_2}(x, y)| \leq l(x, y)|\theta_1 - \theta_2|.$$

Then by using Lemma 19.31 of van der Vaart (1998), the log likelihood ratio, $\sup_{\theta^*, \sqrt{n}|\theta - \theta^*| < M} |\log(dP_{\theta^*}^{\otimes n}/dP_{\theta}^{\otimes n})|$ is $P_{\theta}^{\otimes n}$ -tight for any $M > 0$. The same property holds for $P_{\theta^k} \in M_k$. Since $\alpha_k(\theta, \tilde{\theta} \mid x_n, y_n)$ is defined by using these likelihood ratios, we have

$$\sup_{\theta^*, n^{1/2}|\theta - \theta^*| < M} \alpha_k(\theta, \theta^* \mid x_n, y_n)^{-1} = O_{P_{\theta}^{\otimes n}}(1)$$

for any $M > 0$ and θ . Then to prove the tightness of $\alpha_k(\theta_0, \tilde{\theta}_1 \mid x_n, y_n)^{-1}$, it is sufficient to show $\mathbb{P}_{n,k}$ -tightness of $n^{1/2}|\tilde{\theta}_1 - \theta_0|$ which was already proved. $\square$

PROPOSITION 3.4. *The HPX-DA-MH procedure has the local non-degeneracy.*

PROOF. As in the previous proof, we omit “ k ” for simplicity. We shall obtain a contradiction by assuming the HPX-DA-MH procedure has the φ -local degeneracy. By Lemma 2.5, there exists $\varphi \in C^1(\Theta)$ such that

$$n^{1/2}(\varphi(\theta_1) - \varphi(\theta_0)) = o_{\mathbb{P}_{n,c}}(1).$$

Conditioning on x_n, y_n, z_n and θ_0 , we have $\delta(n^{1/2}\varphi(\theta_1) \mid x_n, y_n, z_n) = o_{\mathbb{P}_{n,c}}(1)$ by Lemma 2.2 (ii). This means $\delta(n^{1/2}\varphi(\theta_0) \mid x_n, y_n, z_n) = o_{\mathbb{P}_{n,c}}(1)$ by (2.7) and hence the DA procedure has the φ -local degeneracy by Lemma 2.3. Thus $\partial_{\beta}\varphi \equiv 0$ by Lemma 3.8.

By Lemma 3.10 (ii), $n^{1/2}(\varphi(\theta_1^*) - \varphi(\theta_0)) = o_{\mathbb{P}_{n,k}}(1)$. Together with this fact and Lemma 2.4, Lemma 2.5 and Lemma 3.10 (iv) we have $n^{1/2}(\varphi(\tilde{\theta}_1) - \varphi(\theta_1^*)) = o_{\mathbb{P}_{n,k}}(1)$. Since $\tilde{\theta}_1 = \gamma_1\theta_1^*$, by Taylor’s expansion we have $n^{1/2}(\gamma_1 - 1)(\sum_{j \in J_k} \partial_{\alpha_j}\varphi(\theta_1^*)\alpha_{1j}^*) = o_{\mathbb{P}_n}(1)$ where $\theta_1^* = (\alpha_{1j}^*, \beta_1^*; j \in J_k)$. By tightness of $\gamma_1 + 1$ we have $n^{1/2}(\gamma_1^2 - 1)(\sum_{j \in J_k} \partial_{\alpha_j}\varphi(\theta_1^*)\alpha_{1j}^*) = o_{\mathbb{P}_n}(1)$. Together with Lemma 2.2 (ii),

$$\delta\left(n^{1/2}\gamma_1^2\left(\sum_{j \in J_k} \partial_{\alpha_j}\varphi(\theta_1^*)\alpha_{1j}^*\right) \mid x_n, y_n, z_n, \theta_1^*\right) = o_{\mathbb{P}_{n,k}}(1).$$

If φ is not a constant, there exists $\theta \in \Theta$ and $k \in J$ such that $\sum_{j \in J_k} \partial_{\alpha_j}\varphi(\theta)\alpha_j \neq 0$. Therefore, $P^k(\Theta_k) > 0$ for some k where $\Theta_k = \{\theta; \sum_{j \in J_k} \partial_{\alpha_j}\varphi(\theta)\alpha_j > \epsilon\}$. Together with the stationarity and the property of the DA, the marginal distribution of θ_1^* is P^k . Thus conditioned on θ_1^* by Lemma 2.2 (v),

$$\delta(n^{1/2}\gamma_1^2 \mid x_n, y_n, z_n, \theta_1^*) = o_{\mathbb{P}_{n,k}(\cdot|\theta_1^*)}(1)$$

for $\theta_1^* \in \Theta_k$ for P^k -a.s. However since the conditional distribution of $n^{1/2}\gamma_1^2$ is the Gamma distribution with parameter $\alpha_n = n(1 + o(1))/2$ and $\beta_n = \frac{\sqrt{n}}{2}(1 + o_{\mathbb{P}_{n,k}(\cdot|\theta_1^*)}(1))$, it cannot happen by Proposition 2.1. Hence the claim follows. $\square$

4. Numerical results

We consider a (relatively) large number of categories, $c = 5$. In this case, the DA and PX-DA procedures have the φ -local degeneracy for some φ , and the HPX-DA-MH procedure has the local non-degeneracy. In practice, the sample size varies from less than 100 to greater than 10^4 . Larger sample sizes result in poor performances for both the DA and the PX-DA procedures. First we show the global behavior for a large sample size. We next show the local behavior for a large sample size, and finally we consider a realistic sample size.

Figures 2 and 3 show the trajectories of α and β , respectively, for the three

Figure 2. Trajectory of α for the DA procedure (left), the PX-DA procedure (middle), and the HPX-DA-MH procedure (right), for $n = 1000$. The dashed lines are the true values.

Figure 3. Trajectory of β for the DA procedure (left), the PX-DA procedure (middle), and the HPX-DA-MH procedure (right), for $n = 1000$. The dashed line is the true value.

MCMC procedures, with $n = 10^3$ and $c = 5$. To see the global behavior, we take an initial guess relatively far from the true value. In this figure, the DA procedure shows poorer mixing than do the other two procedures. Also, the HPX-DA-MH procedure seems better than the PX-DA procedure.

Next we consider the local behaviors. The number of iterations was $m = 10^5$ and the burn-in time is $m/2$. The thinning rate is 5, that is, every fifth iteration is corrected. The auto correlation function (acf, Fig. 4) shows the differences

Figure 4. Auto correlation function (acf) of α_4 for the DA procedure (left), the PX-DA procedure (middle) and the HPX-DA-MH procedure (right), for $n = 1000$ with burn-in time $m/2 = 50000$.

Figure 5. Auto correlation function (acf) of α_5/α_4 for the DA procedure (left), the PX-DA procedure (middle) and the HPX-DA-MH procedure (right), for $n = 1000$ with burn-in time $m/2 = 50000$.

Figure 6. Auto correlation function (acf) of α_4 for the DA procedure (left), the PX-DA procedure (middle), and the HPX-DA-MH procedure (right), for $n = 250$ with burn-in time $m/2 = 50000$.

Figure 7. Auto correlation function (acf) of α_5/α_4 for the DA procedure (left), the PX-DA procedure (middle), and the HPX-DA-MH procedure (right) for $n = 250$ with burn-in time $m/2 = 50000$.

between the three procedures clearly. The improvement of the PX-DA procedure over the DA procedure is canceled out by taking a map $\theta \mapsto \alpha_5/\alpha_4$ (Fig. 5).

Now we consider a smaller sample size of 250 for $c = 5$ with $m = 10^5$ iterations and burn-in time $m/2$. For $n = 250$, the acf of α_4 of both the PX-DA and the HPX-DA-MH procedures are good (Fig. 6). However, the acf of α_5/α_4 for the DA and PX-DA procures are similar, but that of the HPX-DA-MH procedure seems better (Fig. 7). The difference becomes smaller than that

of $n = 1000$, although the HPX-DA-MH procedure is still better than the others.

5. Discussion

The sample size is a big factor in the behaviors of MCMC procedures. For the DA procedure, the parameter α moves on the order of n^{-1} per iteration, although the posterior distribution has the order of $n^{-1/2}$. The PX-DA procedure has a similar property. Hence a larger sample size causes poorer performance for both the DA and the PX-DA procedures. Since the HPX-DA-MH procedure always has the local non-degeneracy, the difference between it and the other two procedures becomes large as n grows.

With a smaller sample size, it is difficult to choose between the PX-DA and the HPX-DA-MH procedures. From the asymptotic point of view, the HPX-DA-MH procedure is always preferred. However, since the HPX-DA-MH procedure uses a Metropolis step, one may want to avoid it when the behavior of the PX-DA procedure is acceptable. Note that when $c \leq 3$, the PX-DA procedure already has the local non-degeneracy, and there is no benefit from the HPX-DA-MH procedure from an asymptotic point of view (see Kamatani (2014b)). Also, even if $c > 4$, if $\alpha = (\alpha_2, \dots, \alpha_{c-1})$ and projections such as α/α_3 have good mixing, the PX-DA procedure may work well. Otherwise, there are advantages to using the HPX-DA-MH procedure.

The proposed procedure and its theoretical results can be easily extended to more general models. The generalization to any cumulative distribution function F instead of Φ is straightforward since our restriction was only for simplicity. It is also possible to extend it to the multinomial probit model, the multivariate probit model, and to those with random effects. We are currently working in this direction.

The notion of the local non-degeneracy is useful for comparing the different Markov chain Monte Carlo procedures. The HPX-DA-MH procedure probably has a better property, local consistency, but showing this is complex. On the other hand, the local non-degeneracy is simple and efficient. The property of the local non-degeneracy is not fully understood, and it should be further analyzed.

Acknowledgements

The author thanks an anonymous referee for constructive comments which were much help in improving this paper. This work is supported in part by Grant-in-Aid for Young Scientists (B) 24740062 and JST, CREST.

REFERENCES

- Aitchison, J. and Silvey, S. D. (1957). The generalization of probit analysis to the case of multiple responses, *Biometrika*, **44**(1), 131–140.
- Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data, *J. Am. Stat. Assoc.*, **88**(422), 669–679.
- Cowles, M. K. (1996). Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models, *Stat. Comput.*, **6**(2), 101–111.

- Doob, J. L. (1949). Application of the theory of martingales, *Le Calcul des Probabilités et ses Applications*, Colloques Internationaux du Centre National de la Recherche Scientifique, no. 13, pp. 23–27, Centre National de la Recherche Scientifique, Paris.
- Itô, K. (2004). *Stochastic Processes*, Lectures given at Aarhus University, Reprint of the 1969 original, Edited and with a foreword by Ole E. Barndorff-Nielsen and Ken-iti Sato, Springer-Verlag, Berlin, ISBN 3-540-20482-2.
- Kamatani, K. (2014a). Local consistency of Markov chain Monte Carlo methods, *Ann. Inst. Statist. Math.*, **66**(1), 63–74.
- Kamatani, K. (2014b). Local degeneracy of Markov chain Monte Carlo methods, *Accepted at ESAIM Probability and Statistics*.
- Le Cam, L. and Yang, G. L. (2000). *Asymptotics in Statistics: Some Basic Concepts*, New York, Tokyo, Springer-Verlag, 2nd edition.
- Liu, I. and Agresti, A. (2005). The analysis of ordered categorical data: An overview and a survey of recent developments, *TEST*, **14**(1), 1–73.
- Liu, J. S. and Sabatti, C. (2000). Generalised Gibbs sampler and multigrid Monte Carlo for Bayesian computation, *Biometrika*, **87**(2), 353–369.
- McCullagh, P. (1980). Regression models for ordinal data, *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **42**(2), 109–142.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores, *Psychometrika Monograph Supplement*, (17).
- Sheiner, L. B., Beal, S. L. and Dunne, A. (1997). Analysis of nonrandomly censored ordered categorical longitudinal data from analgesic trials, *J. Am. Stat. Assoc.*, **92**(440).
- Tierney, L. (1994). Markov chains for exploring posterior distributions, *Ann. Statist.*, **22**(4), 1701–1762, ISSN 0090-5364, doi:10.1214/aos/1176325750, With discussion and a rejoinder by the author.
- van der Vaart, A. W. (1998). *Asymptotic Statistics*, Cambridge, New York, Cambridge University Press.