ESAIM: PS 18 (2014) 713–725
DOI: 10.1051/ps/2014004

ESAIM: Probability and Statistics
www.esaim-ps.org

LOCAL DEGENERACY OF MARKOV CHAIN MONTE CARLO METHODS *

Kengo Kamatani[1]

Abstract. We study asymptotic behavior of Markov chain Monte Carlo (MCMC) procedures. Sometimes the performances of MCMC procedures are poor and there are great importance for the study of such behavior. In this paper we call degeneracy for a particular type of poor performances. We show some equivalent conditions for degeneracy. As an application, we consider the cumulative probit model. It is well known that the natural data augmentation (DA) procedure does not work well for this model and the so-called parameter-expanded data augmentation (PX-DA) procedure is considered to be a remedy for it. In the sense of degeneracy, the PX-DA procedure is better than the DA procedure. However, when the number of categories is large, both procedures are degenerate and so the PX-DA procedure may not provide good estimate for the posterior distribution.

Mathematics Subject Classification. 65C40, 62E20.

Received Januay 19, 2012. Revised May 24, 2013.

1. Introduction

This paper investigates poor behavior of Markov chain Monte Carlo (MCMC) procedures which provides good information for construction of efficient MCMC procedures. There have a vast literature related to sufficient conditions for good property, ergodicity; see reviews [16,20]. The transition kernel of an MCMC procedure is Harris recurrent and geometrically ergodic under fairly general assumptions. In practice, however, the performance can be poor even with geometric ergodicity. Therefore another approach seems to be appropriate for the study of poor behavior of MCMC procedures. This is the motivation for the present study.

Theoretical analysis for poor performance is rarely studied. However somewhat similar motivation, comparison of different MCMC procedures have been studied in the past few decades. Suppose now that $P(x,\mathrm{d}y)$ is a Markov transition kernel corresponding to an MCMC procedure. We also assume that $P(x,\mathrm{d}y)$ has the invariant probability measure Π. As an operator $f(x)\mapsto(Pf)(x)=\int_y P(x,\mathrm{d}y)f(y)$, the spectral radius of the transition kernel $P(x,\mathrm{d}y)$ is of great interest, since it determines the rate of convergence of P^n to Π. Good estimate of the spectral radius leads to good comparison of MCMC procedures. This so-called spectral approach was studied in [1] for finite state space and [22] for more general state space.

Asymptotic properties of P^n can also be studied indirectly *via* the so-called drift function $V(x)$. This approach dates back to [3] and it can calculate the convergence rate of P^n by establishing an inequality of $V(x)$ and

Keywords and phrases. Markov chain Monte Carlo, asymptotic normality, cumulative link model.

* *Supported in part by Grant-in-Aid for Young Scientists (B) 22740055 and 24740062.*

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

Article published by EDP Sciences

$PV(x) = \int P(x, \mathrm{d}y)V(y)$. It is possible to compare Markov chains by their rates of convergence. See [17, 18] for how to obtain these inequalities in practice.

These approaches are powerful, but usually it requires some technical difficulties to obtain a good comparison. There is also a beautiful indirect comparison method initially studied by [15] and developed by [14, 21]. This approach requires virtually no calculation for comparison although the conclusion of the comparison is rather weak. These approaches are summarized and further developed in [4].

In the present paper we consider degeneracy as a particular type of poor behavior. Our approach is not a comparison technique, but identify poor MCMC procedure that requires improvement. The identification of degeneracy is technically easy but the conclusion is rather strong. To describe degeneracy, consider the following well-known example:

$$P(x = 1|\theta) = \Phi(\theta),\ \ P(x = 0|\theta) = 1 - \Phi(\theta), \tag{1.1}$$

where Φ is the cumulative distribution function of the standard normal distribution. Assume we have an observation $x_n = (x^1, \dots, x^n)$ with the prior $N(0, 1)$ on θ. Define the so-called augmented data model:

$$x = \begin{cases} 1 \text{ if } y \le \theta \\ 0 \text{ if } y > \theta \end{cases} \tag{1.2}$$

where $y \sim N(0, 1)$. Then this model becomes (1.1) if we integrate out y. The data augmentation (DA) procedure based on this augmented data model is the iteration of the following steps:

$$\begin{cases} \textbf{simulate } y^i|x^i, \theta \sim \begin{cases} N(0, 1, -\infty, \theta) \text{ if } x^i = 1 \\ N(0, 1, \theta, \infty) \quad \text{if } x^i = 0 \end{cases} (i = 1, \dots, n) \\ \textbf{simulate } \theta|x_n, y_n \sim N(0, 1, \max_{i:x^i=1} y^i, \min_{i:x^i=0} y^i) \end{cases} \tag{1.3}$$

where $N(0, 1, a, b)$ is the normal distribution truncated to the interval (a, b), and $y_n = (y^i)_{i=1,\dots,n}$. This procedure generates a Markov chain having the posterior distribution as the invariant distribution.

The model (1.1) can also be constructed by introducing the following latent structure:

$$x = \begin{cases} 1 \text{ if } y \le 0 \\ 0 \text{ if } y > 0 \end{cases} \tag{1.4}$$

where $y \sim N(-\theta, 1)$. The corresponding DA procedure is the iteration of the following:

$$\begin{cases} \textbf{simulate } y^i|x^i, \theta \sim \begin{cases} N(-\theta, 1, -\infty, 0) \text{ if } x^i = 1 \\ N(-\theta, 1, 0, \infty) \quad \text{if } x^i = 0 \end{cases} (i = 1, \dots, n) \\ \textbf{simulate } \theta|x_n, y_n \sim N(-\sum_{i=1}^n y^i/(n+1), 1/(n+1)). \end{cases} \tag{1.5}$$

We obtain two DA procedures. The former has uniform ergodicity by Proposition 4 of [7], and the latter has geometric ergodicity by Theorem 1 of [19]. Despite of their similarity, the performances are quite different. Figure 1 shows trajectories of the DA procedures for $m = 200$ iterations and the sample sizes $n = 100$ (upper) and $n = 1000$ (lower). For both simulations, the true value is 0.5 and the initial value is 0.35. The DA procedure (1.3) has poor mixing property than (1.5) and it may require quite a large number of iteration until convergence. The difference between procedures becomes even larger when the sample size is larger.

The key fact is that the interval for the simulation of θ in (1.3) is very short (*cf.* p. 64 of [9]). This update produces almost the same value (write θ') as the current value (write θ) so it does cause a performance bottleneck. In [8], we define a good property, local consistency for MCMC procedures by letting the sample size $n \to \infty$. In a similar way, in the present study, we analyze such poor behavior through this limit. As a property of poor behavior, we call that an MCMC procedure has the local degeneracy if

$$\sqrt{n}|\theta' - \theta| = o_{\mathbb{P}}(1) \tag{1.6}$$

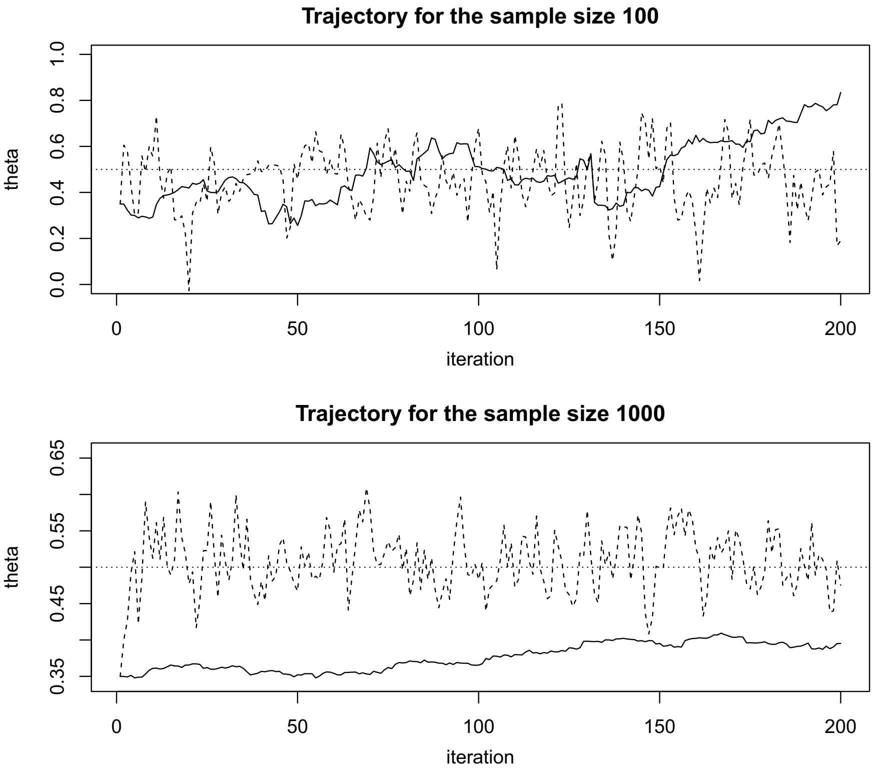


FIGURE 1. Trajectories of the DA procedures for the sample sizes $n = 100$ (upper) and $n = 1000$ (lower). The solid line is for the model (1.2) and the dashed line is that for (1.4).

where $\mathbb{P}$ is a particular probability measure. For the above two DA procedures, the former has the local degeneracy but the latter has the local consistency.

The paper is organized as follows. Section 2 is devoted to the study of degeneracy of the MCMC procedure. In the Bayesian context, we prove that degeneracy defined in (1.6) occurs only if the model has non-regularity. Therefore the performance bottleneck due to local degeneracy can be avoided by checking its model regularity. In Section 3 we apply this to the cumulative probit model. We will show that a natural MCMC procedure for this model has the same non-regularity as (1.2) so it causes a performance bottleneck. We also show that a *remedy* also suffers from the same bottleneck. Finally, some remarks are presented in Section 4.

We write $X|Y$ for the law of X conditioned on Y. We also write $X|Y \sim P(\mathrm{d}x|Y)$ if the law of X conditioned on Y is $P(\mathrm{d}x|Y)$. We assume that $P(\mathrm{d}x|Y = y)$ is a probability measure for each y and $y \mapsto P(A|Y = y)$ is measurable.

2. DEGENERACY

2.1. Definition of degeneracy

In this section, we review the (local) consistency and define degeneracy. Let $\Theta = \mathbb{R}^d$ be a parameter space and let $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ $(n = 1, 2, \ldots)$ be a sequence of probability spaces. Fix $n \in \mathbb{N}$. Suppose x_n is an observation. We are interested in the approximation of probability measure $P(\mathrm{d}\theta|x_n)$. Suppose that $\theta_\infty = (\theta_m)_{m\in\mathbb{N}_0}$ is an

$\mathbb{R}^d$-valued process such that $\theta_\infty|x_n$ is (strictly) stationary[2] and the invariant probability measure is $P(\mathrm{d}\theta|x_n)$. For $M \in \mathbb{N}$, let

$$I_M^n(f) = \frac{1}{M}\sum_{m=0}^{M-1} f(\theta_m),\ I^n(f) = \int f(\theta)P(\mathrm{d}\theta|x_n).$$

For fixed x_n, if $\theta_\infty|x_n$ is ergodic, then by Birkhoff's pointwise convergence theorem,

$$\lim_{M\to\infty} I_M^n(f)(x_n) = I^n(f)(x_n) \text{ a.s.} \tag{2.1}$$

for any $P(\mathrm{d}\theta|x_n)$-integrable function f. This, "observation wise convergence" is usually satisfied for MCMC procedures (under positive Harris recurrence. See Thm. 17.1.7 of [13]). If this convergence still holds when n and $M \to \infty$, we call the process consistent. See [8] for the details. The idea behind it is that by taking $n \to \infty$, the structure of the Markov chain becomes simpler, and the analysis becomes easier.

Definition 2.1 (Consistency)**.** The process $\theta_\infty = (\theta_m)_{m\in\mathbb{N}_0}$ is said to have the consistency if $I_{M_n}^n(f) - I^n(f) = o_{\mathbb{P}_n}(1)$ for any continuous, bounded function f for any $M_n \to \infty$.

Consistency property is usually satisfied by many Markov chains generated by Markov chain Monte Carlo procedure (see Thm. 1 of [8]). However this is not always the case. This corresponds to the non-ergodic case of Birkhoff's pointwise convergence theorem. In the following, we consider a particular poor behavior; $I_M^n(f)$ is no more helpful than $I_1^n(f)$ as an approximation of $I^n(f)$. Note that the observation wise convergence (2.1) is usually satisfied even for this case.

Definition 2.2 (Degeneracy)**.** The process $\theta_\infty = (\theta_m)_{m\in\mathbb{N}_0}$ is said to have the degeneracy if $I_M^n(f) - I_1^n(f) = o_{\mathbb{P}_n}(1)$ for any continuous, bounded function f and for any $M \in \mathbb{N}$.

We will see that essentially, the good behavior, consistency, and the bad behavior, degeneracy are exclusive. To apply consistency and degeneracy for Bayesian statistics, we need a slight modification of the above definitions. We assume Bernestein von-Mises's theorem, that is, for some $\mathbb{R}^d$-valued random variable $u_n(x_n)$, we have

$$\sqrt{n}(\theta - u_n) = O_{\mathbb{P}_n}(1) \tag{2.2}$$

where $\theta|x_n \sim P(\mathrm{d}\theta|x_n)$. For this case, it is natural to consider asymptotic properties of $(\sqrt{n}(\theta_m - u_n))_{m\in\mathbb{N}_0}$ rather than $(\theta_m)_{m\in\mathbb{N}_0}$. Moreover, we will see that even if the trajectories of $(\theta_m)_{m\in\mathbb{N}_0}$ looks fine, some projection $\varphi : \mathbb{R}^d \to \mathbb{R}^k$ $(k \le d)$ reveals its poor performance for estimation of $I^n(f)$. Thus we will study asymptotic properties of $(\sqrt{n}(\varphi(\theta_m) - u_n))_{m\in\mathbb{N}_0}$ for some $\mathbb{R}^k$-valued $u_n(x_n)$. We introduce φ-local properties.

Definition 2.3 (Local properties)**.** The process $\theta_\infty = (\theta_m)_{m\in\mathbb{N}_0}$ is said to have the φ-local consistency (resp. φ-local degeneracy) if $\tilde{\theta}_\infty = \{\sqrt{n}(\varphi(\theta_m) - u_n)\}_{m\in\mathbb{N}_0}$ satisfies consistency (resp. degeneracy). If φ is the identity map, then we call the property local consistency (resp. local degeneracy).

2.2. Properties of degeneracy

In this section, some properties of consistency and degeneracy will be studied. First we note a representation of degeneracy.

Proposition 2.4. *Assume $\mathbb{E}_n[P(\mathrm{d}\theta|x_n)] =: P_n(\mathrm{d}\theta)$ is tight. Then degeneracy is equivalent to*

$$\theta_1 - \theta_0 = o_{\mathbb{P}_n}(1). \tag{2.3}$$

[2] Stationary assumption is impractical for Markov chain generated by MCMC. However it can be weakened. See Lemma 4 of [8] and Section B of [6].

Proof. First we show that degeneracy is equivalent to

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

for any bounded continuous function f. To see this, necessity follows by $o_{\mathbb{P}_n}(1) = I_2^n(f) - I_1^n(f) = -(f(\theta_1) - f(\theta_0))/2$. Sufficiency is clear by stationarity since $I_M^n(f) - I_1^n(f)$ is a finite sum of the elements of $\{f(\theta_m) - f(\theta_{m-1})\}_{m\in\mathbb{N}}$. Hence the equivalence of degeneracy and (2.4) follows, and now we check the equivalence of (2.3) and (2.4). However, by tightness condition, the joint law of $(\theta_0, \theta_1 - \theta_0)$ is tight and hence sufficiency of (2.3) comes from continuity of $(x_1, x_2) \mapsto f(x_1) - f(x_1 + x_2)$. For the necessity of (2.3), consider $f(x) = \exp(iu^t x)$ where u^t is the transpose of the vector $u \in \mathbb{R}^d$. Then $\mathbb{E}_n[|f(\theta_1) - f(\theta_0)|] = \mathbb{E}_n[|\exp(iu^t(\theta_1 - \theta_0)) - 1|] \to 0$ and hence $\mathbb{E}_n[\exp(iu^t(\theta_1 - \theta_0))] \to 1$ for any $u \in \mathbb{R}^d$, that implies (2.3). Thus the claim follows. □

For local properties, we apply the above to $\sqrt{n}(\varphi(\theta) - u_n)$ in place of θ. Then the tightness condition becomes $\sqrt{n}(\varphi(\theta) - u_n) = O_{\mathbb{P}_n}(1)$ and (2.3) becomes $\sqrt{n}(\varphi(\theta_1) - \varphi(\theta_0)) = o_{\mathbb{P}_n}(1)$.

Another important property is that consistency and degeneracy are essentially, mutually exclusive; if the both hold, then $P(\mathrm{d}\theta|x_n)$ should converge to a Dirac measure δ_{u_n} for some random variable $u_n(x_n)$. In other words, the dispersion of $P(\mathrm{d}\theta|x_n)$ tends to 0 in probability (see [5] for the definition of the dispersion).

Proposition 2.5. *Assume the same condition as Lemma 2.4. If both consistency and degeneracy hold, then there exists $\mathbb{R}^d$-valued random variable $u_n(x_n)$ such that*

$$\int \min\{|\theta - u_n|, 1\} P(\mathrm{d}\theta|x_n) = o_{\mathbb{P}_n}(1).$$

Proof. By degeneracy, we can find $M_n \to \infty$ such that $I_{M_n}^n(f) - I_1^n(f) = o_{\mathbb{P}_n}(1)$. Thus together with consistency, we have

$$I_1^n(f) - I^n(f) = f(\theta_0) - I^n(f) = o_{\mathbb{P}_n}(1). \tag{2.5}$$

Note that $I^n(f)$ only depends on x_n. Suppose now that we have two independent draws θ and θ' from $P(\mathrm{d}\theta|x_n)$ for a fixed x_n. Then by (2.5), we have $f(\theta) - f(\theta') = (f(\theta) - I^n(f)) - (f(\theta') - I^n(f)) = o_{\mathbb{P}_n}(1)$ for any bounded continuous function f. Hence $\theta - \theta' = o_{\mathbb{P}_n}(1)$ as in the previous proposition. We can choose $u_n(x_n)$ to be measurable such that

$$\int \min\{|\theta - u_n|, 1\} P(\mathrm{d}\theta|x_n) \leq \int \min\{|\theta - \theta'|, 1\} P(\mathrm{d}\theta'|x_n) P(\mathrm{d}\theta|x_n) = o_{\mathbb{P}_n}(1). \tag{2.6}$$

Thus the claim follows. □

2.3. Degeneracy for DA procedure

For data augmentation (DA) procedure, there is another simpler equivalent condition for degeneracy. Now we consider a probability measure $P(\mathrm{d}\theta, \mathrm{d}x_n, \mathrm{d}y_n) = P(\mathrm{d}\theta, \mathrm{d}y_n|x_n)P(\mathrm{d}x_n)$ such that

$$P(\mathrm{d}\theta, \mathrm{d}y_n|x_n) = P(\mathrm{d}\theta|x_n, y_n)P(\mathrm{d}y_n|x_n) = P(\mathrm{d}y_n|\theta, x_n)P(\mathrm{d}\theta|x_n) \tag{2.7}$$

in $P(\mathrm{d}x_n)$-a.e. We define data augmentation (DA) procedure. For an observation $x_n \sim P(\mathrm{d}x_n)$, it generate Markov chain $\theta_\infty = (\theta_m)_{m\in\mathbb{N}_0}$ by iteration of

$$y_{n,m}|\theta_m, x_n \sim P(\mathrm{d}y_n|\theta_m, x_n),\ \ \theta_{m+1}|x_n, y_{n,m} \sim P(\mathrm{d}\theta|x_n, y_{n,m})$$

where $y_{n,m}$ is a working variable. This Markov chain is invariant with respect to $P(\mathrm{d}\theta|x_n)$ and we also assume stationarity of $\theta_\infty|x_n$, that is, $\theta_0|x_n \sim P(\mathrm{d}\theta|x_n)$.

Proposition 2.6. *Assume the same condition as Proposition 2.4. Then for the DA procedure, degeneracy is equivalent to the existence of* $\mathbb{R}^d$*-valued variable* $v_n(x_n, y_n)$ *such that*

$$\int \min\{|v_n - \theta|, 1\} P(\mathrm{d}\theta, \mathrm{d}x_n, \mathrm{d}y_n) = o(1). \tag{2.8}$$

Proof. By (2.7), the DA procedure has reversibility in the following sense;

$$P(\mathrm{d}\theta_1|x_n, y_{n,0})P(\mathrm{d}y_{n,0}|\theta_0, x_n)P(\mathrm{d}\theta_0|x_n) = P(\mathrm{d}\theta_0|x_n, y_{n,0})P(\mathrm{d}y_{n,0}|\theta_1, x_n)P(\mathrm{d}\theta_1|x_n).$$

Therefore, the law of $\theta_1 - v_n(x_n, y_{n,0})|x_n$ and $\theta_0 - v_n(x_n, y_{n,0})|x_n$ are the same. Hence (2.8) implies degeneracy by $\theta_1 - \theta_0 = (\theta_1 - v_n) - (\theta_0 - v_n) = o_{\mathbb{P}_n}(1)$. Now we show the necessity of (2.8). By (2.7) again, we have

$$P(\mathrm{d}\theta_1|x_n, y_n)P(\mathrm{d}y_n|\theta_0, x_n)P(\mathrm{d}\theta_0|x_n) = P(\mathrm{d}\theta_1|x_n, y_n)P(\mathrm{d}\theta_0|x_n, y_n)P(\mathrm{d}y_n|x_n)$$

Therefore if we denote $P(\mathrm{d}y_n|x_n)P(\mathrm{d}x_n)$ by $P(\mathrm{d}x_n, \mathrm{d}y_n)$, degeneracy implies

$$\int \min\{|\theta_1 - \theta_0|, 1\} P(\mathrm{d}\theta_1|x_n, y_n)P(\mathrm{d}\theta_0|x_n, y_n)P(\mathrm{d}x_n, \mathrm{d}y_n) = o(1).$$

However, as in (2.6) we can choose $v_n(x_n, y_n)$ which is measurable and satisfies

$$\int \min\{|v_n(x_n, y_n) - \theta_0|, 1\} P(\mathrm{d}\theta_0|x_n, y_n) \leq \int \min\{|\theta_1 - \theta_0|, 1\} P(\mathrm{d}\theta_1|x_n, y_n)P(\mathrm{d}\theta_0|x_n, y_n) = o_{\mathbb{P}_n}(1).$$

Hence the claim follows. □

By this proposition, if we consider local properties under $\theta \mapsto \sqrt{n}(\theta - u_n)$, local degeneracy of the DA procedure is equivalent to the existence of an estimator $v_n(x_n, y_n)$ such that $\sqrt{n}(\theta - v_n) = o_{\mathbb{P}_n}(1)$. In statistical point of view, it means that we can construct a good estimator v_n that has the convergence rate better than the usual $\sqrt{n}$-rate if we observe y_n. This implies non-regularity of the model $x_n, y_n|\theta$. Note that it is already well-known that the large amount of information of the working variable y_n speed down the convergence of the DA procedure (see for example [12]).

2.4. The first application

We defined two DA procedures in Section 1. Write Θ and $P(\mathrm{d}\theta)$ for the parameter space and the prior distribution with respectively, and assume

$$P(\mathrm{d}x_n) = \int_\Theta \prod_{i=1}^n P(\mathrm{d}x^i|\theta)P(\mathrm{d}\theta). \tag{2.9}$$

Other conditional distributions, such as the posterior distribution $P(\mathrm{d}\theta|x_n)$ are defined in the usual way from the model settings in Section 1. It is not difficult to show the existence of u_n such that $\sqrt{n}(\theta - u_n) = O_{\mathbb{P}_n}(1)$ (see Sect. 4 for the detail). For example, we can take $u_n = \Phi(n_1/n)$ where n_1 is the number of observation of $x^i = 1$.

Example 2.7. Examine asymptotic properties of two DA procedures in Section 1. First we consider the DA procedure defined in (1.3). We check

$$n(\theta - v_n) = O_{\mathbb{P}_n}(1)$$

for $v_n = \max_{i; x^i = 1} y^i$. If this holds, then it has the local degeneracy by Proposition 2.6. By Taylor's expansion, it is sufficient to check

$$n(\tilde{\theta} - \tilde{v}_n) = O_{\mathbb{P}_n}(1)$$

for $\tilde{v}_n = \Phi(v_n)$ and $\tilde{\theta} = \Phi(\theta)$. Since $\tilde{v}_n/\tilde{\theta}|\theta \sim \mathrm{Beta}(n_1, 1)$ we have

$$\begin{aligned}\mathbb{E}_n\left[\{n(\tilde{\theta}-\tilde{v}_n)\}^2\middle|\theta, x_n\right] &= n^2\tilde{\theta}^2\mathbb{V}_n\left[\frac{\tilde{v}_n}{\tilde{\theta}}\middle|\theta, x_n\right] + n^2\tilde{\theta}^2\left(\frac{n_1}{n_1+1}-1\right)^2 \\ &= n^2\tilde{\theta}^2\frac{n_1}{(n_1+1)^2(n_1+2)} + n^2\tilde{\theta}^2\left(\frac{n_1}{n_1+1}-1\right)^2.\end{aligned}$$

This value is tight since $\tilde{\theta} \in [0,1]$ and $n_1|\theta$ follows the binomial distribution and hence $n(\tilde{\theta}-\tilde{v}_n) = O_{\mathbb{P}_n}(1)$. Therefore the DA procedure defined in (1.3) has the local degeneracy by Proposition 2.6.

Second, we consider the DA procedure defined in (1.5). Suppose that θ and θ' are two independent draws from $\theta|x_n, y_n$. Then

$$\sqrt{n}(\theta-\theta') = \sqrt{n}(\theta-v_n) - \sqrt{n}(\theta'-v_n) \sim N\left(0, \frac{2n}{n+1}\right) \tag{2.10}$$

by (1.5) where $v_n = -\sum_{i=1}^n y^i/(n+1)$. However, by Propositon 2.4, if the DA procedure has local degeneracy, then $\sqrt{n}(\theta-\theta') = o_{\mathbb{P}_n}(1)$ that contradicts (2.10). Thus the DA procedure defined in (1.5) does not have the local degeneracy.

In fact, the model $P(\mathrm{d}x, \mathrm{d}y|\theta)$ in (1.4) is regular, so the DA procedure for this model has the local consistency by Theorem 1 of [8]. Hence the conclusion for (1.5) also follows from Proposition 2.5. On the other hand, the model (1.2) has the parameter-dependent support that introduces non-regularity. Thus as stated in the end of Section 2.3, local degeneracy of (1.3) is a natural consequence of this observation.

3. Application

3.1. Cumulative probit model

Cumulative probit model has a categorical variable $y \in \{1,\dots,c\}$ and an explanatory variable $x \in \mathbb{R}^p$ with the parameter $\theta = (\alpha, \beta)$ such that $\beta \in \mathbb{R}^p$ and $\alpha \in \Theta_\alpha := \{(\alpha_2,\dots,\alpha_{c-1}); 0 =: \alpha_1 < \alpha_2 < \ldots < \alpha_{c-1} < \alpha_c := +\infty\}$ such that

$$P_\theta(y \le j|x) = \Phi(\alpha_j + \beta^t x)\ (j = 1,\dots,c) \tag{3.1}$$

where β^t is the transpose of the vector β. Consider the standard normal distribution as the prior distribution for α and β truncated to $\Theta_\alpha \times \mathbb{R}^p$.

The posterior distribution is complicated, but there is a natural DA procedure; For observations $x_n = (x^i)_{i=1,\dots,n}$ and $y_n = (y^i)_{i=1,\dots,n}$, an iteration is defined by

$$\begin{cases}\textbf{simulate } z^i|x_n, y_n, \theta \sim N(-\beta^t x^i, 1, \alpha_{y^i-1}, \alpha_{y^i}) & (i = 1,\dots,n),\\ \textbf{simulate } \alpha_j|x_n, y_n, z_n \sim N(0, 1, \max_{i:y^i=j-1} z_i, \min_{i:y^i=j} z_i) & (j = 2,\dots,c-1),\\ \textbf{simulate } \beta|x_n, y_n, z_n \sim N(-(1+\sum_{i=1}^n x^i(x^i)^t)^{-1}\sum_{i=1}^n x^i z^i, (1+\sum_{i=1}^n x^i(x^i)^t)^{-1}) & \end{cases} \tag{3.2}$$

where $z_n = (z^i)_{i=1,\dots,n}$ is a working variable. This DA procedure implicitly uses the following latent structure:

$$z|x, y, \theta \sim N(-\beta^t x, 1) \text{ and } y = j \text{ if } z \in (\alpha_{j-1}, \alpha_j]. \tag{3.3}$$

Later we will see that this DA procedure work quite poorly. It is a natural consequence since the model has the parameter-dependent support.

Surprisingly, in some cases, the DA procedure can be drastically improved by adding a single working parameter. Strategies that do this include the parameter-expanded data augmentation (PX-DA) procedure proposed by [10] and the marginal augmentation procedure proposed by [11]. In the present case, we add the following step after each iteration of (3.2):

$$\begin{cases}\textbf{simulate}\ \gamma^2|\theta,x_n,y_n,z_n\sim\text{Gamma}\left(\frac{n+p+c}{2}-1,\frac{1}{2}\left(\sum_{i=1}^n(z^i+\beta^tx^i)^2+|\beta|^2+\sum_{j=2}^{c-1}\alpha_j^2\right)\right)\\ \textbf{set}\qquad\quad \theta\leftarrow\gamma\theta\end{cases}\tag{3.4}$$

where Gamma(ν,α) is the Gamma distribution with the shape parameter ν and rate parameter α. The procedure $\theta'\leftarrow\gamma\theta$ does not break stationarity, that is, if $\theta|x_n\sim P(\mathrm{d}\theta|x_n)$ and if $\gamma^2|\theta,x_n,y_n,z_n$ is generated by the above, then $\theta'=\gamma\theta|x_n\sim P(\mathrm{d}\theta|x_n)$ (see Thm. 1 of [9]).

3.2. Degeneracy results

Assume that $x_n=(x^i)_{i=1,\ldots,n}$ is an i.i.d. sample from the probability distribution $G(\mathrm{d}x)$, which has the compact support. We also assume that the expectation of $(1,x^t)\begin{pmatrix}1\\x\end{pmatrix}$ for $x\sim G(\mathrm{d}x)$ is a non-degenerate matrix. This non-degeneracy assumption is for the existence of non-degenerate Fisher information matrix (see condition R_c of [2]). Tightness condition $\sqrt{n}(\varphi(\theta)-u_n)$ for some u_n for φ defined in Propositions 3.1 and 3.2 is satisfied (see Sect. 4 for the detail).

Proposition 3.1. *The DA procedure has φ-local degeneracy for $\varphi(\theta)=(\alpha_j)_{j=2,\ldots,c-1}$ if $c>2$.*

Proof. We prove the claim by Proposition 2.6. Set $v_n=(\max_{i:y^i=2}z^i,\ldots,\max_{i:y^i=c-1}z^i)$. We only show $n(\max_{i;y^i=j}z^i-\alpha_j)=O_{\mathbb{P}_n}(1)$ for $j=2$ since the proof is the same for $j=3,\ldots,c-1$.

For $\epsilon>0$, choose a compact set $K\subset\Theta_\alpha\times\mathbb{R}^d$ so that $\int_{\theta\in K^c}P(\mathrm{d}\theta)<\epsilon/2$ where $P(\mathrm{d}\theta)$ is the prior distribution. Since the support of G is compact, we have the following bound; For any $H>0$, for some $C,c>0$

$$\frac{\Phi(\alpha_2+\beta^tx)-\Phi(\alpha_2+\beta^tx-h)}{\Phi(\alpha_2+\beta^tx)-\Phi(\alpha_1+\beta^tx)}\geq Ch,\ \Phi(\alpha_2+\beta^tx)-\Phi(\alpha_1+\beta^tx)\geq c\ (\theta\in K,x\in\operatorname{supp}G,h\in[0,H]).$$

Then for $\theta\in K$,

$$\begin{aligned}\mathbb{P}_n\left(n\left(\max_{i;y^i=2}z^i-\alpha_2\right)\leq -h|\theta,x_n,y_n\right)&=\prod_{i;y^i=2}\mathbb{P}_n(n(z^i-\alpha_2)\leq -h|\theta,x_n,y_n)\\&=\prod_{i;y^i=2}\left\{1-\frac{\Phi(\alpha_2+\beta^tx)-\Phi(\alpha_2+\beta^tx-h/n)}{\Phi(\alpha_2+\beta^tx)-\Phi(\alpha_1+\beta^tx)}\right\}\\&\leq(1-Ch/n)^{n_2}\leq\exp(-Chn_2/n)\end{aligned}$$

where n_2 is the number of elements of $y^i=2$. Since $\mathbb{P}_n(y^i=2|\theta)\geq c\ (\theta\in K)$, by the Lebesgue–Fatou lemma,

$$\limsup_{n\to\infty}\mathbb{E}_n[\exp(-Chn_2/n)|\theta]\leq\exp(-Cch).$$

Therefore if we choose $h>0$ so that $\exp(-Cch)<\epsilon/2$, we have

$$\limsup_{n\to\infty}\mathbb{P}_n\left(n\left(\max_{i;y^i=2}z^i-\alpha_2\right)\leq -h\right)\leq\int_{\theta\in K^c}P(\mathrm{d}\theta)+\exp(-Cch)<\epsilon.$$

Note here that $\max_{i;y^i=2}z^i-\alpha_2$ is always negative. Thus $n(\max_{i;y^i=j}z^i-\alpha_j)=O_{\mathbb{P}_n}(1)$ for $j=2$. In the same way, we can show it for $j\geq 3$. Hence φ-local degeneracy follows for $c>2$ by Proposition 2.6. □

Another results may be counter intuitive; Although the simulation results may look fine, the PX-DA procedure still has the local degeneracy.

Proposition 3.2. *The PX-DA procedure has the φ-local degeneracy with respect to $\varphi(\theta) = (\alpha_j/\alpha_2)_{j=3,\ldots,c-1}$ if $c > 3$.*

Proof. We apply Proposition 2.4. The function φ satisfies $\varphi(\gamma\theta) = \varphi(\theta)$. Therefore, $\varphi(\theta_1) - \varphi(\theta_0)|x_n$ has the same law under the PX-DA procedure and the DA procedure defined in (3.2). Thus, φ-local degeneracy of the PX-DA procedure is equivalent to φ-local degeneracy of the DA procedure. Let

$$v_n = \left(\frac{\max_{i:y^i=j} z^i}{\max_{i:y^i=2} z^i}\right)_{j=3,\ldots,c-1}.$$

In the previous proposition, we already observed that $n(\max_{i:y^i=j} z^i - \alpha_j)$ is tight. Therefore it is not difficult to conclude that $n(\varphi(\theta) - v_n)$ is tight. Then the DA procedure has local degeneracy by Proposition 2.6 and hence the PX-DA procedure has the φ-local degeneracy. □

3.3. Simulation

We perform simulation for the cumulative probit model for $c = 5$ and $p = 2$ for the DA and PX-DA procedures with the sample size $n = 100$, iteration $m = 10^5$ but the first $m/2 = 0.5 * 10^5$ values are eliminated as burn-in. Observations x_n and y_n are generated from (3.1) for a fixed true parameter θ^*. Trajectory of $\alpha = (\alpha_2, \alpha_3, \alpha_4)$ and $\beta = (\beta_1, \beta_2)$ are displayed in Figure 2. It shows the poor mixing property of α for the DA procedure (left side of Fig. 2). The PX-DA procedure looks better.

These results are not surprising since the poor performance of the DA procedure and the efficiency of the PX-DA procedure are well known. However the difference between the PX-DA and DA procedures can be small. In the present case, by the projection $\varphi(\theta) = \alpha_3/\alpha_2$ or α_4/α_2, we can observe that the DA and PX-DA procedures have similar poor results (Fig. 3). For further illustration of this degeneracy, consider the sample size $n = 1000$ (Figs. 4 and 5). As for the sample size $n = 100$, without the above projection, simulation results of the trajectories and its auto correlation function (acf) look much better for the PX-DA procedure. However with the projection, we can observe that the benefit of the use of the PX-DA procedure is small (Fig. 5). Thus the PX-DA is still have poor convergence property as proved in Proposition 3.2.

4. Final remark

We proposed a notion for poor performance of the MCMC procedures. This property is easy to check and efficient as studied in Section 3. In particular, the poor performance of this PX-DA procedure was not reported in elsewhere. Some theoretical properties are investigated in Section 2 and it reveals that these poor performances have close connection to non-regularity of the model.

The study of poor performance of MCMC procedures are still being developed and we hope that this paper works as a good step toward that direction. For further analysis, we are working in two directions:

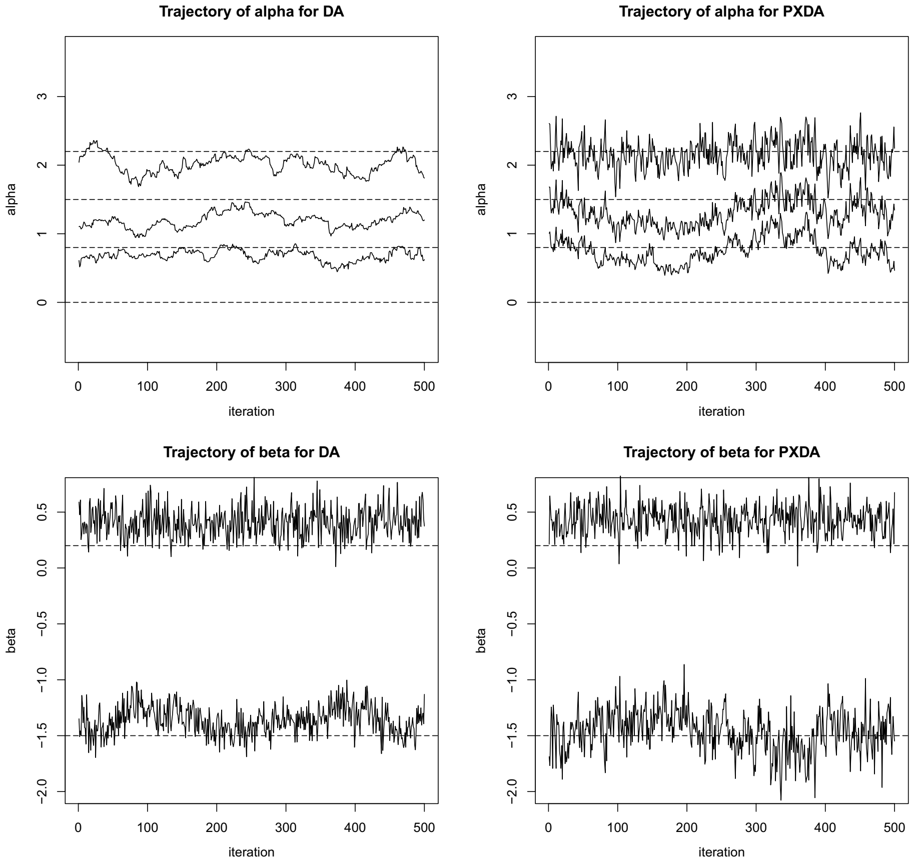
(1) The study of the rate of $M_n \to \infty$ to hold $I^n_{M_n}(f) - I^n(f) = o_{\mathbb{P}_n}(1)$, which corresponds to the analysis for the sufficient number of iteration of MCMC procedures. For the local consistency case, we can take any $M_n \to \infty$ but it is not possible for the local degeneracy case. For the latter case, we can not take "any" $M_n \to \infty$ but sometimes it is possible to find the explicit rate to hold the convergence. This direction, the analysis of weak consistency will be studied in [6].

(2) Even if degeneracy holds, usually it is possible to find the rate $r_n \to \infty$ such that

$$r_n\left(I^n_M(f) - I^n_1(f)\right) \neq o_{\mathbb{P}_n}(1).$$

Compared to the above, it is technically easier to calculate the rate r_n. This direction, the order of degeneracy provides good information for the performance bottleneck of the MCMC procedures.

Both of which defines the rate of convergence, so we can compare MCMC procedures by those rates. Also the application to the cumulative probit model is of interest. This topic will further be studied in elsewhere.

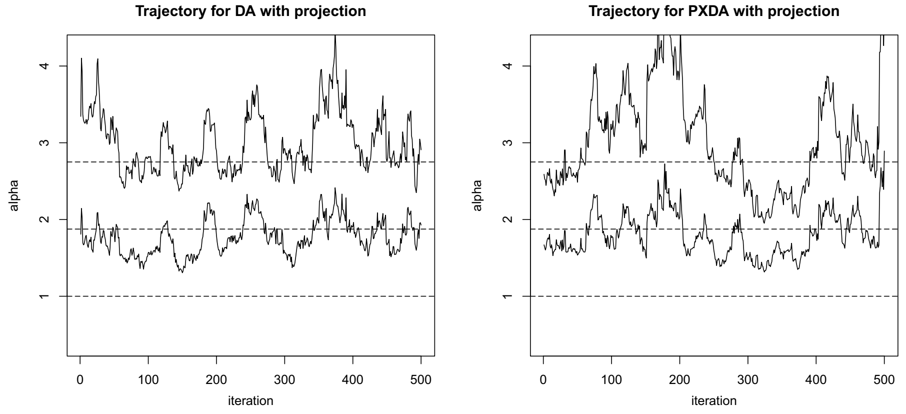

Trajectory of alpha for DA

Trajectory of alpha for PXDA

Trajectory of beta for DA

Trajectory of beta for PXDA

FIGURE 2. Trajectory for the sample size $n = 100$ for α_2 to α_4 (upper figures) and β_1, β_2 (lower figures) for the DA (left figures) and PX-DA (right figures) procedures. Horizontal lines are the true parameters θ^* including $\alpha_1 = 0$. The DA procedure has the φ-local degeneracy for $\varphi(\theta) = (\alpha_2, \ldots, \alpha_4)$.


Trajectory for DA with projection

Trajectory for PXDA with projection

FIGURE 3. The trajectories of projected sequence α_3/α_2 and α_4/α_2 for the DA procedure (left) and the PX-DA procedure (right). Horizontal lines are corresponding to the true parameter.

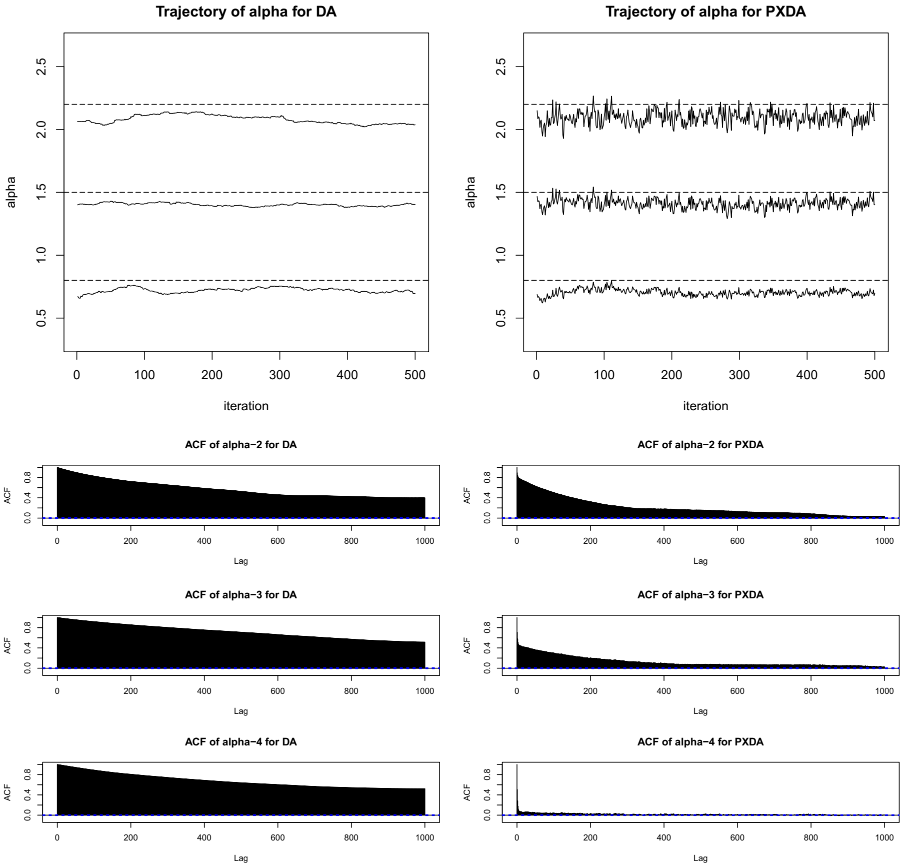


FIGURE 4. Trajectories (upper figures) and auto correlation functions (lower figures) of $\varphi(\theta) = \alpha_2, \alpha_3, \alpha_4$ for the sample size $n = 1000$. The PX-DA procedure (right figures) looks much better than the DA procedure (left figures).

APPENDIX A. TIGHTNESS CONDITION

Usually, tightness condition as in Proposition 2.4 for some u_n is not difficult to show. Consider a model $x|\theta \sim P(\mathrm{d}x|\theta)$ with prior $P(\mathrm{d}\theta)$. Let $u_n(x_n)$ be the maximum likelihood estimator for observation $x_n \sim P(\mathrm{d}x_n)$ where $P(\mathrm{d}x_n)$ is as in (2.9). Then we may assume that under regularity conditions,

$$\sqrt{n} I(\theta)^{1/2}(\theta - u_n) \Rightarrow N(0, I), \tag{A.1}$$

where $I(\theta)$ is the Fisher information matrix, and I is the identity matrix. Therefore, if

$$I(\theta)^{-1/2} \text{ is tight for } \theta \sim P(\mathrm{d}\theta) \tag{A.2}$$

then $\sqrt{n}(\theta - u_n)$ is tight. For example, if $I(\theta)$ is continuous with respect to θ and $I(\theta)$ is strictly positive, then (A.2) is satisfied.

We show this tightness condition for the model (3.1). In this case, both x_n and y_n are observed. We check the tightness of $\sqrt{n}(\varphi(\theta) - u_n)$ for some $u_n(x_n, y_n)$ for φ defined in Propositions 3.1 and 3.2. By continuity of $(\alpha_j)_{j=2,\ldots,c-1} \mapsto (\alpha_j/\alpha_2)_{j=3,\ldots,c-1}$, tightness for the latter comes from that for the former. So we only show that for the former.

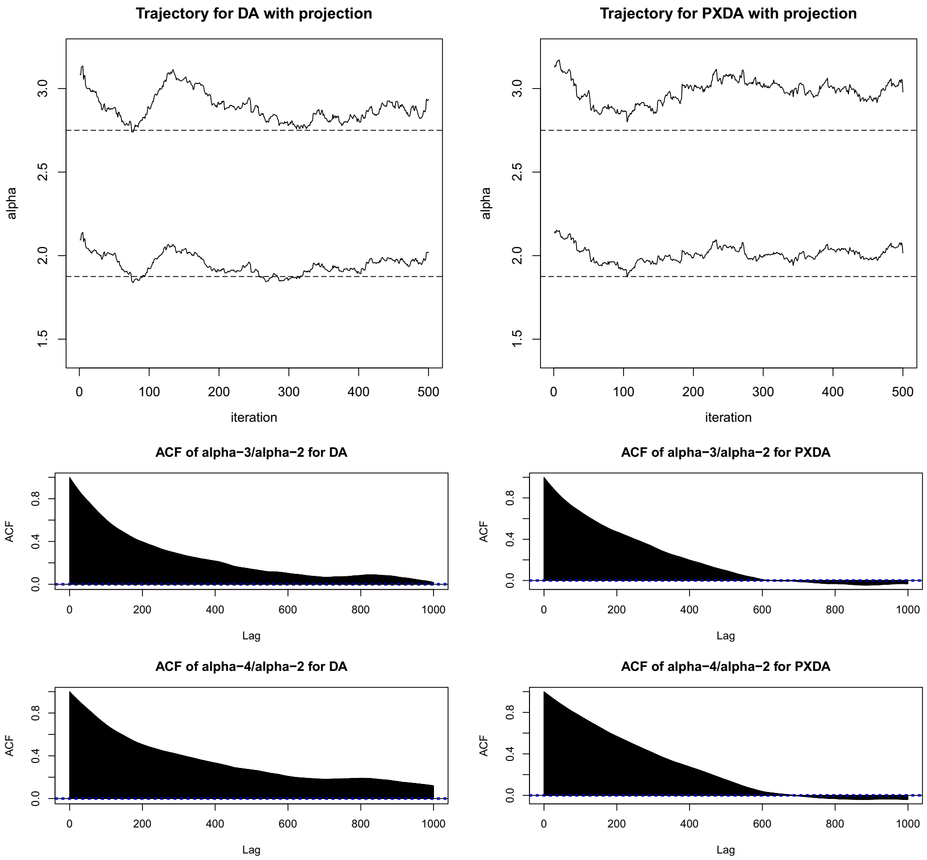


FIGURE 5. Trajectories (upper figures) and auto correlation functions (lower figures) of $\varphi(\theta) = \alpha_3/\alpha_2$ and α_4/α_2. The improvements by the PX-DA procedure (right figures) over the DA procedure (left figures) are small.

Fix $j \in \{2, \ldots, c-1\}$. Suppose that we only have a partial observation of y_n. More precisely, we only observe the event $\{y^i \le j\}$ or $\{y^i > j\}$ with x^i for $i = 1, \ldots, n$. Then this partially observed model becomes a probit model with parameter $\theta_j = (\alpha_j, \beta)$, since

$$P_\theta(y \le j|x) = \Phi(\alpha_j + \beta^t x) = \Phi\left(\theta_j^t \begin{pmatrix} 1 \\ x \end{pmatrix}\right).$$

Then we can apply Corollary 1 of [2] for this model, and hence asymptotic normality for the maximum likelihood estimator holds. Thus we obtain (A.1) and (A.2) and hence $\sqrt{n}(\alpha_j - u_{n,j}(x_n, u_n)) = O_{\mathbb{P}_n}(1)$ for some $u_{nj}(x_n, y_n)$. By showing it for each $j \in \{2, \ldots, c-1\}$, we have the claim.

Acknowledgements. The author wishes to express his thanks to Prof. Nakahiro Yoshida for his several helpful comments and suggestions. The author also thank to the Associate Editor and anonymous referee for constructive comments which helped a lot to improve the paper.

References

[1] P. Diaconis and L. Saloff-Coste, Comparison theorems for reversible markov chains. *Ann. Appl. Probab.* **696** (1993).

[2] L. Fahrmeir and H. Kaufmann, Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann. Statist.* **13** (1985) 342–368.

[3] F.G. Foster, On the stochastic matrices associated with certain queuing processes. *Ann. Math. Statist.* **24** (1953) 355–360.

[4] J.P. Hobert and D. Marchev, A theoretical comparison of the data augmentation, marginal augmentation and PX-DA algorithms. *Ann. Statist.* **36** (2008) 532–554.

[5] K. Itô, Stochastic processes. ISBN 3-540-20482-2. 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 (2004).

[6] K. Kamatani, Local weak consistency of Markov chain Monte Carlo methods with application to mixture model. *Bull. Inf. Cyber.* **45** (2013) 103–123.

[7] K. Kamatani, Note on asymptotic properties of probit gibbs sampler. *RIMS Kokyuroku* **1860** (2013) 140–146.

[8] K. Kamatani, Local consistency of Markov chain Monte Carlo methods. *Ann. Inst. Stat. Math.* **66** (2014) 63–74.

[9] J.S. Liu and C. Sabatti, Generalised Gibbs sampler and multigrid Monte Carlo for Bayesian computation. *Biometrika* **87** (2000) 353–369.

[10] Jun S. Liu and Ying Nian Wu, Parameter expansion for data augmentation. *J. Am. Stat. Assoc.* **94** (1999) 1264–1274.

[11] X.-L. Meng and David van Dyk, Seeking efficient data augmentation schemes *via* conditional and marginal augmentation. *Biometrika* **86** (1999) 301–320.

[12] Xiao-Li Meng and David A. van Dyk, Seeking efficient data augmentation schemes *via* conditional and marginal augmentation. *Biometrika* **86** (1999) 301–320.

[13] S.P. Meyn and R.L. Tweedie, Markov Chains and Stochastic Stability. Springer (1993).

[14] Antonietta. Mira, *Ordering, Slicing and Splitting Monte Carlo Markov Chains*. Ph.D. thesis, University of Minnesota (1998).

[15] P.H. Peskun, Optimum monte-carlo sampling using markov chains. *Biometrika* **60** (1973) 607–612.

[16] G.O. Roberts and J.S. Rosenthal, General state space markov chains and mcmc algorithms. *Prob. Surveys* **1** (2004) 20–71.

[17] J.S. Rosenthal. Minorization conditions and convergence rates for Markov chain Monte Carlo. *J. Am. Stat. Assoc.* **90** (1995) 558–566.

[18] J.S. Rosenthal, Quantitative convergence rates of markov chains: A simple account. *Electron. Commun. Probab.* **7** (2002) 123–128.

[19] V. Roy and J.P. Hobert, Convergence rates and asymptotic standard errors for Markov chain Monte Carlo algorithms for Bayesian probit regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** (2007) 607–623.

[20] L. Tierney, Markov chains for exploring posterior distributions. *Ann. Statist.* **22** (1994) 1701–1762.

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

[22] Wai Kong Yuen. Applications of geometric bounds to the convergence rate of Markov chains on $\mathbf{R}^n$. *Stoch. Process. Appl.* **87** 20001–23.