Biometrika (1994), **81**, 1, *pp.* 27–40
Printed in Great Britain

Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes

BY JUN S. LIU

Department of Statistics, Harvard University, Cambridge, Massachussetts 02138, U.S.A.

WING HUNG WONG AND AUGUSTINE KONG

Department of Statistics, The University of Chicago, Chicago, Illinois 60637, U.S.A.

SUMMARY

We study the covariance structure of a Markov chain generated by the Gibbs sampler, with emphasis on data augmentation. When applied to a Bayesian missing data problem, the Gibbs sampler produces two natural approximations for the posterior distribution of the parameter vector: the empirical distribution based on the sampled values of the parameter vector, and a mixture of complete data posteriors. We prove that Rao–Blackwellization causes a one-lag delay for the autocovariances among dependent samples obtained from data augmentation, and consequently, the mixture approximation produces estimates with smaller variances than the empirical approximation. The covariance structure results are used to compare different augmentation schemes. It is shown that collapsing and grouping random components in a Gibbs sampler with two or three components usually result in more efficient sampling schemes.

Some key words: Data augmentation; Empirical and mixture estimators; Forward operator; Interleaving Markov property; Maximal correlation; Rao–Blackwellization.

1. INTRODUCTION

The Gibbs sampler is an iterative scheme for the approximate generation of samples from a multivariate distribution. It is related to the Metropolis algorithm (Metropolis et al., 1953; Hastings, 1970) in statistical physics, and was introduced by Geman & Geman (1984) in the context of statistical image restoration. The basic Gibbs sampler can be described as follows.

Suppose $X = \{x(1), \ldots, x(d)\}$ is a d-component random variable whose joint density function $\pi(X)$ is difficult to compute directly. Each component $x(i)$ can be multidimensional. The d conditional distributions $\pi\{x(i)|X^{[-i]}\}$, where $X^{[-i]}$ denotes $\{x(j), j \neq i\}$, however, are assumed to be easy to draw samples from. The Gibbs sampler is a stochastic relaxation technique that allows us to obtain samples from the joint density $\pi(X)$ by running a Markov chain with $\pi(X)$ as its equilibrium distribution. The chain is initiated by a draw from a starting density $p_0(X)$, or a fixed point. Each component $x(i)$ is visited and updated by a sample drawn from the conditional distribution $\pi\{x(i)|X^{[-i]}\}$. In a systematic scan, the components are visited according to a fixed order. In a random scan, the visiting order is random. As long as each component is visited infinitely often, under some mild conditions, the Markov chain will be ergodic and have $\pi(X)$ as its invariant distribution.

A systematic use of such iterative sampling schemes in parametric statistical problems was introduced by Tanner & Wong (1987) where a similar method, called data augmentation, was presented for approximate computation of posterior densities. In this paper, we use the term data augmentation to refer to the two component Gibbs sampler. Li (1988) has applied similar schemes to impute multivariate missing data. Gelfand & Smith (1990) clarify the formal connection between data augmentation and the Gibbs sampler, and raise further interesting theoretical questions like, for example, the comparisons of different estimators described below. There is by now a long list of papers in this area. Despite its popularity, some theoretical questions concerning the Gibbs sampler have not been satisfactorily resolved. Specifically, we are interested in the following two questions.

Comparisons of different estimators. Consider the case with $d = 2$ components. For simplicity, x and y are used in place of $x(1)$ and $x(2)$. It is often the case that $t(X)$, a scalar function of interest, is a function of a single component of X; that is, without loss of generality, $t(X) = t(x)$. Let (x_k, y_k) $(k = 1, 2, \ldots)$ be successive samples generated by the Gibbs sampler. Then two natural estimates of $\mu = E_\pi\{t(x)\}$ are

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} t(x_k), \tag{1}$$

$$\tilde{\mu} = \frac{1}{n}\sum_{k=1}^{n} E_\pi\{t(x)|\, y_k\}. \tag{2}$$

We call (1) the empirical estimator. The estimate (2), which is often easy to compute assuming that the conditional density $\pi(x|\, y)$ is simple, is called the mixture estimator. These two names stem from the Bayesian missing data problem setting, where x represents the parameter vector and y represents the missing data. In that case, $\hat{\mu}$ is the expectation of t under the empirical distribution based on dependent samples drawn from the true posterior density, and $\tilde{\mu}$ is a mixture of complete data posterior means. It is commonly believed that the mixture estimator is always better, i.e. has a smaller variance, than the empirical estimator. As demonstrated by Gelfand & Smith (1990), if (x_k, y_k) and (x_l, y_l) are independent for all $k \neq l$, the proof follows from the fact that $\mathrm{var}_\pi\,[E_\pi\{t(x)|\, y\}] \leqslant \mathrm{var}_\pi\,\{t(x)\}$. However their proof does not apply when (x_k, y_k) $(k = 0, 1, \ldots, n)$ are dependently drawn using the Gibbs sampler. A related question is whether it is better to use a weighted combination of the two estimates.

Comparisons of various augmentation schemes. In many applications, one has some freedom in choosing among a number of sampling schemes. To be specific, suppose there are three random variables x, y and z, each of which can either be a scalar or a vector. Of interest is some random scalar function t of x and y only; that is, $t = t(x, y)$. Consider the following three sampling schemes, each of which is iterated to obtain the required estimate.

Scheme [1]: $x|\, y, \quad y|\, x.$

Scheme [2]: $x|(y, z), \quad (y, z)|\, x.$

Scheme [3]: $x|(y, z), \quad y|(x, z), \quad z|(x, y).$

Both [1] and [2] are data augmentation schemes as there are only $d = 2$ components in the iterations. In [1], the two components being iterated are x and y with z integrated out. In [2], y and z are grouped together as a single component. Scheme [3] treats x, y

and z as three separate components. Each component is iteratively sampled conditioned on the current values of the other two. Compared with [1], an extra random variable z is introduced in Schemes [2] and [3]. This may be done because it is easier to draw from $\pi(x|y, z)$ than from $\pi(x|y)$. However, since more is imputed, we expect [2] and [3] to have slower convergence rates and higher autocovariances relative to Scheme [1]. Similarly, one may conjecture that Scheme [2] has a faster convergence rate than Scheme [3] since y and z are sampled jointly conditioned on x. In general, choosing an optimal sampling scheme requires a compromise made between the rate of convergence and the ease of computation.

By studying the covariance structure of the samples generated by the Gibbs sampler, we are able to resolve completely the first question under stationarity, and provide useful results for the second question. The paper is organized as follows: § 2 contains preliminary lemmas concerning the covariance structure of a general Markov chain; § 3 focuses on a detailed analysis of data augmentation. A rigorous proof of the superiority of the mixture estimator over the empirical estimator for data augmentation is presented in § 4. Comparisons of different sampling schemes appear in § 5. In the last section, we present an example to illustrate the main results of §§ 4 and 5.

2. Preliminaries

Throughout the paper, $X_0, X_1, \ldots$ are assumed to be consecutive samples generated by a time-homogeneous and stationary Markov chain with transition function $K(X_0, X_1)$ and equilibrium density $\pi(X)$. We use $\pi(.)$ and $\pi(.|.)$ to denote all marginal and conditional distributions under the equilibrium distribution π. Taking $X = (x, y)$ for example, $\pi(x, y)$ denotes the joint density for x and y, while $\pi(x)$ and $\pi(y)$ are the marginal densities of x and y, implying that $\pi(x, y)$, $\pi(x)$ and $\pi(y)$ are all different density functions related to each other in such a way that $\pi(x) = \int \pi(x, y)\, dy$ and $\pi(y) = \int \pi(x, y)\, dx$. Moreover, $\pi(x|y) = \pi(x, y)/\pi(y)$ and $\pi(y|x) = \pi(x, y)/\pi(x)$.

It is important to distinguish between two different kinds of expectations involved. We use $E(.)$ and $\mathrm{var}\,(.)$ to denote expectations and variances for the Markov chain where the integrations are taken over all the sample paths; and use $E_\pi(.)$ or $\mathrm{var}_\pi\,(.)$ to denote those taken under π, where the integrations are done with respect to the density π.

Let

$$L_0^2(\pi) = \{t(X): E_\pi\{t(X)\} = 0;\ \mathrm{var}_\pi\{t(X)\} < \infty\}$$

denote the space of all mean zero scalar functions of X with finite variances. It is a Hilbert space with an inner product defined by $\langle t(X), s(X)\rangle = E_\pi\{t(X)s(X)\}$. So the norm of a function $t(X)$ is $\|t\| = [E_\pi\{t^2(X)\}]^{\frac{1}{2}}$. On this space we define two operators as follows.

Definition 1. *The forward operator F and the backward operator B are:*

$$Ft(X) = E\{t(X_1)|X_0 = X\} = \int K(X, X_1)t(X_1)\, dX_1, \tag{3}$$

$$Bt(X) = E\{t(X_0)|X_1 = X\} = \int \frac{\pi(X_0)K(X_0, X)}{\pi(X)}\, t(X_0)\, dX_0. \tag{4}$$

The norm of an operator is defined as $\|F\| = \sup \|Ft\|/\|t\|$, where the supremum is taken over all $t \in L_0^2(\pi)$. The spectral radius of F is denoted as $\mathrm{spec}\,(F)$.

Clearly, F and B are adjoint linear operators on $L_0^2(\pi)$, and their norms are bounded above by one. When the state space is finite, F and B are simply the transition matrix of the Markov chain and its transpose.

From the Chapman–Kolmogorov property, it is seen that

$$F^n t(X_0) = E\{t(X_n)|X_0\}, \quad B^n t(X_n) = E\{t(X_0)|X_n\}. \tag{5}$$

Furthermore, if the chain is reversible, i.e. the detailed balance condition $\pi(X)K(X, X') = \pi(X')K(X', X)$ is satisfied, F and B are self-adjoint operators.

Note. The function space of interest might be taken to be $L^2(\pi) = \{t : \text{var}_\pi(t) < \infty\}$, which is a space of all functions of X with finite variances. So $L_0^2(\pi)$ is a subspace of it. Operators F and B exist on $L^2(\pi)$ as well. There are many references on bounding convergence rates under this setting. See Diaconis (1988) and Bhattacharya & Waymire (1990) for comprehensive discussions. Evidently, the largest eigenvalue of F, when it is restricted on $L_0^2(\pi)$, is the same as the second largest eigenvalue of the ordinary Markov transition operator acting on the space $L^2(\pi)$. By considering a more restricted space $L_0^2(\pi)$, however, we are able to eliminate the constant function as an eigenfunction for the operator, so that the comparison of different chains can be based on a comparison of norms or spectral radii of their associated forward operators.

LEMMA 2·1. *The norms and the spectral radii of operators F and B are always equal.*

Proof. A property of adjoint operators (Yosida, 1980, § VII.2). □

LEMMA 2·2. *For any functions $s(.)$ and $t(.)$ in $L_0^2(\pi)$,*

$$\text{cov}\{t(X_n), s(X_0)\} = \text{cov}\{F^k t(X), B^{n-k} s(X)\}$$

for any integer $0 \leqslant k \leqslant n$.

Proof. It follows from the Markov property that

$$\begin{aligned} E\{t(X_n)s(X_0)\} &= E[E\{t(X_n)s(X_0)|X_{n-1}\}] \\ &= E[E\{t(X_n)|X_{n-1}\}E\{s(X_0)|X_{n-1}\}] \\ &= E\{Ft(X_{n-1})B^{n-1}s(X_{n-1})\}. \end{aligned}$$

An induction argument leads to the conclusion. □

For any two random variables W and V, their maximal correlation is defined as

$$\gamma(W, V) = \sup \text{corr}\{f(W), g(V)\} = \sup_{f:\text{var}\{f(W)\}=1} (\text{var}[E\{f(W)|V\}])^{\frac{1}{2}}. \tag{6}$$

where the first supremum is taken over all nonconstant scalar functions such that $\text{var}\{f(W)\} < \infty$ and $\text{var}\{g(V)\} < \infty$. Saramanov (1958) and Lancaster (1958) provide general treatments of maximal correlations as well as conditions for $\gamma < 1$. Csàki & Fischer (1960) give examples for $\gamma = 1$. Breiman & Friedman (1985) use the concept to justify their ACE algorithm. The maximal correlation is also closely related to the concept of ρ-mixing in Markov chain theory (Bradley, 1986). Amit (1991) uses an equivalent notion to characterize the convergence rate of stochastic relaxations for some Gaussian-like distributions. The lemma below follows from (5), (6) and the definition of the norm of an operator.

LEMMA 2·3. *For the Markov chain we have considered, $\|F^n\| = \|B^n\| = \gamma(X_0, X_n)$.*

3. Data augmentation

3·1. *General description*

Data augmentation is the Bayesian analogue of the EM algorithm, and can be applied to many missing data problems. In this case, the random variable X involved in the iterative sampling scheme has two components, one of them, say x, corresponds to the parameter vector θ in the EM formulation, and the other one, y, corresponds to the missing data. So only the marginal distribution of x, instead of the joint distribution of x and y, is of interest.

With a single imputation per iteration, data augmentation is equivalent to the Gibbs sampler applied to two components. The scan used in data augmentation is a systematic one where x and y are sampled alternately. Thinking of the samples generated by data augmentation, $(x_0, y_0), (x_1, y_1), \ldots$, as consecutive states of a Markov chain, we can write down its transition function:

$$K\{(x_0, y_0), (x_1, y_1)\} = \pi(y_1 | x_1)\pi(x_1 | y_0).$$

Let F be the forward operator of this chain, F_x and F_y be the corresponding forward operators of the marginal chains $\{x_k, k = 0, 1, \ldots\}$ and $\{y_k, k = 0, 1, \ldots\}$, and K_x be the transition function for the chain $\{x_k, k = 0, 1, \ldots\}$. Clearly,

$$K_x(x_0, x_1) = \int \pi(x_1 | y_0)\pi(y_0 | x_0)\, dy_0.$$

We assume that the Markov chain induced by data augmentation is always in stationarity, which is equivalent to assuming that the distribution of x_0 is the marginal distribution of x under $\pi(x, y)$. We also use $\gamma_\pi(x, y)$ to denote the maximal correlation between x and y whose joint distribution is $\pi(x, y)$; that is, $\gamma_\pi(x, y) = \gamma(x, y)$.

3·2. *Covariance structure*

LEMMA 3·1. *The marginal chains $\{x_k, k = 0, 1, \ldots\}$ and $\{y_k, k = 0, 1, \ldots\}$ constructed by data augmentation are reversible Markov chains.*

Proof. We check the detailed balance condition directly:

$$\begin{aligned}
\pi(x_0)K_x(x_0, x_1) &= \int \pi(x_1 | y_0)\pi(y_0 | x_0)\pi(x_0)\, dy_0 \\
&= \int \pi(x_1 | y_0)\pi(x_0, y_0)\, dy_0 = \int \pi(x_0 | y_0)\pi(x_1, y_0)\, dy_0 \\
&= \int \pi(x_0 | y_0)\pi(y_0 | x_1)\pi(x_1)\, dy_0 = \pi(x_1)K_x(x_1, x_0).
\end{aligned}$$

The first and last equalities follow from the fact

$$K_x(x_0, x_1) = \int \pi(x_1 | y)\pi(y | x_0)\, dy. \qquad \square$$

LEMMA 3·2. *For $t(.) \in L_0^2(\pi)$, the one-lag autocovariance of $t(x)$ is nonnegative and equals*

$$\operatorname{cov}\{t(x_0), t(x_1)\} = \operatorname{var}_\pi [E_\pi \{t(x) | y\}].$$

A similar identity holds for the one-lag autocovariance of $s(y)$.

Proof. We have that

$$\text{cov}\,\{t(x_0), t(x_1)\} = E_\pi[E\{t(x_0)t(x_1)|\, y_0\}] = E_\pi([E_\pi\{t(x)|\, y\}]^2).$$

The second equality follows because x_0 and x_1 are conditionally independent and identically distributed given y_0. □

The above proof uses an important property of the marginal chains, i.e. the interleaving Markov property defined as follows.

DEFINITION 2. *A pair of stationary Markov chains* $\{x_k, k=0, 1, \ldots\}$ *and* $\{y_k, k=0, 1, \ldots\}$ *are said to be conjugate to each other with the interleaving Markov property if*:

(*a*) x_k *and* x_{k+1} *are conditionally independent given* y_k, *for* $k=0, 1, \ldots$;

(*b*) y_k *and* y_{k+1} *are conditionally independent given* x_{k+1}, *for* $k=0, 1, \ldots$;

(*c*) (x_k, y_k), (x_{k+1}, y_k) *and* (x_{k+1}, y_{k+1}) *are identically distributed, for* $k=0, 1, \ldots$.

The interleaving Markov property automatically implies the reversibility of both chains, which can be proved the same way as Lemma 3·1. The following lemma is also immediate.

LEMMA 3·3. *The marginal chains* $\{x_k, k=0, 1, \ldots\}$ *and* $\{y_k, k=0, 1, \ldots\}$ *constructed by data augmentation are conjugate to each other with the interleaving Markov property.*

THEOREM 3·1. *If a stationary Markov chain* $\{x_k, k=0, 1, \ldots\}$ *has the interleaving Markov property with its conjugate chain* $\{y_k, k=0, 1, \ldots\}$, *then Rao–Blackwellization causes a one-lag delay for autocovariances, i.e., for any integer* $m>0$ *and function* $t(x) \in L_0^2(\pi)$,

$$\text{cov}\,\{t(x_0), t(x_m)\} = \text{cov}\,\{s^*(y_0), s^*(y_{m-1})\},$$

where $s^*(y) = E_\pi\{t(x)|\, y\}$. *Hence for any square integrable functions* $t(.)$ *of* x *and* $s(.)$ *of* y, *their n-lag autocovariances are nonnegative, monotone decreasing with n, and have the following expressions*:

$$\text{cov}\,\{t(x_0), t(x_n)\} = \text{var}_\pi\,\{E_\pi(\ldots E_\pi[E_\pi\{t(x)|\, y\}|\, x]|\ldots)\}, \tag{7}$$

$$\text{cov}\,\{s(y_0), s(y_n)\} = \text{var}_\pi\,\{E_\pi(\ldots E_\pi[E_\pi\{s(y)|x\}|\, y]|\ldots)\}, \tag{8}$$

where the right-hand sides of both (7) *and* (8) *have n expectation signs conditioned alternately on x and y.*

Proof. By applying Lemma 3·3, using the Markov property, and noting that

$$E\{t(x_0)|\, y_0\} = s^*(y_0), \quad E\{t(x_m)|\, y_{m-1}\} = s^*(y_{m-1}),$$

we get

$$\begin{aligned} E\{t(x_0)t(x_m)\} &= E[E\{t(x_0)t(x_m)|\, y_0, x_m\}] = E[E\{t(x_0)|\, y_0\}t(x_m)] \\ &= E[E\{s^*(y_0)t(x_m)|\, y_0, y_{m-1}\}] = E[s^*(y_0)E\{t(x_m)|\, y_{m-1}\}] \\ &= E\{s^*(y_0)s^*(y_{m-1})\}. \end{aligned}$$

To prove the second part, we apply induction to both $E\{t(x_0)t(x_n)\}$ and $E\{s(y_0)s(y_n)\}$ simultaneously. When $n=1$, the result is true from Lemma 3·2. Assume the result is true for $n=m-1$. Then by the one-lag delay phenomenon and the induction assumption on

(8), it follows that

$$E\{t(x_0)t(x_m)\} = E\{s^*(y_0)s^*(y_{m-1})\}$$

$$= E_\pi[\{E_\pi(\ldots E_\pi[E_\pi\{s^*(y)|x\}|y]|\ldots)\}^2] \tag{9}$$

$$= E_\pi([E_\pi\{\ldots E_\pi(E_\pi[E_\pi\{t(x)|y\}|x]|y)|\ldots\}]^2), \tag{10}$$

where (9) has $m-1$ expectation signs while (10) has m. The formula for $E\{s(y_0)s(y_m)\}$ can be obtained in the same way by applying the induction assumption on (7). The monotone decreasing property of the autocovariances follows the inequality $\text{var}_\pi\{g(x)\} \geqslant \text{var}_\pi[E_\pi\{g(x)|y\}]$, for any g. □

The next result illustrates the relations between the joint chain and the marginal chains in terms of the norms and spectral radii of the corresponding forward operators. These norms and spectral radii are all related to the maximal correlation $\gamma_\pi(x, y)$ between x and y.

THEOREM 3·2. *Let $X_k = (x_k, y_k)$ $(k = 0, 1, \ldots)$ be consecutive samples generated from data augmentation with $X_0 \sim \pi$, then $\gamma(X_0, X_1) = \gamma_\pi(x, y)$ and $\gamma(x_0, x_1) = \{\gamma_\pi(x, y)\}^2$, where $\gamma(.,.)$ is defined in* (6). *Hence $\|F\|^2 = \|F_x\| = \|F_y\|$. However, the spectral radii of F, F_x and F_y are all the same and equal to $\{\gamma_\pi(x, y)\}^2$.*

Proof. Since (x_0, y_0), (x_1, y_0), (x_1, y_1) are all identically distributed with distribution π, it is immediate from (6) that $\gamma(X_0, X_1) \geqslant \gamma_\pi(x, y)$. For any function $g(X)$ with $\text{var}_\pi(g) = 1$, however,

$$E\{g(X_1)|X_0\} = E\{g(x_1, y_1)|y_0\} = E[E_\pi\{g(x_1, y_1)|x_1\}|y_0] = E_\pi[E_\pi\{g(X)|x\}|y = y_0].$$

Thus if we write $g^*(x) = E_\pi\{g(X)|x\}$,

$$\text{var}[E\{g(X_1)|X_0\}] = \text{var}_\pi[E_\pi\{g^*(x)|y\}] \leqslant \{\gamma_\pi(x, y)\}^2 \text{var}_\pi\{g^*(x)\} \leqslant \{\gamma_\pi(x, y)\}^2,$$

which implies that $\{\gamma(X_0, X_1)\}^2 \leqslant \{\gamma_\pi(x, y)\}^2$. Hence $\gamma(X_0, X_1) = \gamma_\pi(x, y)$.

Applying (6) to the samples obtained from data augmentation, we have

$$\{\gamma(x_0, x_1)\}^2 = \sup_{t:\text{var}_\pi(t)=1} \text{var}_\pi[E\{t(x_1)|x_0\}] = \sup_{t:\text{var}_\pi(t)=1} \text{var}_\pi(E_\pi[E_\pi\{t(x)|y\}|x]).$$

Since

$$\{\gamma_\pi(x, y)\}^2 = \sup_{t:\text{var}_\pi(t)=1} \text{var}_\pi[E_\pi\{t(x)|y\}] = \sup_{s:\text{var}_\pi(s)=1} \text{var}_\pi[E_\pi\{s(y)|x\}],$$

it is easy to see that $\{\gamma(x_0, x_1)\}^2 \leqslant \{\gamma_\pi(x, y)\}^4$. On the other hand, since

$$\gamma(x_0, x_1) \geqslant \sup_{t:\text{var}_\pi(t)=1} \text{cov}\{t(x_0)t(x_1)\} = \sup_{t:\text{var}_\pi(t)=1} \text{var}_\pi[E_\pi\{t(x)|y\}] = \{\gamma_\pi(x, y)\}^2,$$

we obtain $\gamma(x_0, x_1) = \{\gamma_\pi(x, y)\}^2$. The same is true for $\gamma(y_0, y_1)$. The identities for the norms of the three operators follow from Lemma 2·3.

To prove the identities for the spectral radii, we notice that for any function $g(X)$, where $X = (x, y)$,

$$Fg(X_0) = E\{g(X_1)|X_0\} = E\{g(X_1)|y_0\}.$$

Hence $F^n g(X) = F_y^{n-1}(Fg)(y)$. It is easy to see that $\text{spec}(F) = \text{spec}(F_y)$ from the identity $\text{spec}(F) = \lim_{n\to\infty} \|F^n\|^{1/n}$ (Yosida, 1980, § VIII.2).

Since the marginal chain $\{x_k, k = 0, 1, \ldots\}$ is reversible, its forward operator F_x is self-adjoint. Yosida (1980, § VII.2) says that, for self-adjoint operators $\|F_x^n\| = \|F_x\|^n$; his

§ VIII.2 has that $\lim_n \|F_x^n\|^{1/n}$ converges to the spectral radius of F_x. Hence $\|F_x^n\| = \|F_x\|^n = \{\gamma_\pi(x, y)\}^{2n}$, and $\{\gamma_\pi(x, y)\}^2$ is the spectral radius of F_x. □

3·3. *Upper and lower bounds for autocovariances*

Using Theorem 3·1, we are able to bound the autocovariances, which are defined as

$$\{\sigma_n(t)\}^2 = \text{cov}\,\{t(x_0), t(x_n)\} \quad (n = 0, 1, \ldots), \tag{11}$$

by geometric sequences.

LEMMA 3·4. *The autocovariances are log-convex in lags, i.e.* $\{\sigma_n(t)\}^2 \leqslant \sigma_j(t)\sigma_{2n-j}(t)$.

Proof. Define $G_0 = t(x)$, $G_1 = E_\pi\{t(x)|y\}$, and recursively $G_{2k} = E_\pi(G_{2k-1}|x)$, $G_{2k+1} = E_\pi(G_{2k}|y), \ldots$. Then

$$E_\pi(G_0 G_{2k}) = E_\pi[G_0(x)E_\pi\{G_{2k-1}(y)|x\}] = E_\pi\{G_0(x)G_{2k-1}(y)\}$$
$$= E_\pi[E_\pi\{G_0(x)G_{2k-1}(y)|y\}] = E_\pi\{G_1(y)G_{2k-1}(y)\}.$$

By repeating the above procedure we obtain $E_\pi(G_{2k-j}G_j) = E_\pi(G_k^2)$, for any nonnegative integers k and $j < 2k$. Therefore Theorem 3·1 and Hölder's inequality together give the assertion. □

THEOREM 3·3. *For any* $t \in L_0^2(\pi)$ *with* $\text{var}_\pi(t) = 1$, *we have the following bounds for the autocovariances*:

$$\{\sigma_1(t)\}^{2n} \leqslant \text{cov}\,\{t(x_0), t(x_n)\} \leqslant \{\gamma_\pi(x, y)\}^{2n}.$$

The bounds are sharp.

Proof. By Lemma 3·4 and an induction argument, we easily get the first inequality. By Theorem 3·2 and Lemma 2·2, the second inequality follows from

$$\{\sigma_n(t)\}^2 = \text{cov}\,\{t(x_0), t(x_n)\} = \langle F_x^n t, t\rangle \leqslant \|F_x^n\| = \{\gamma_\pi(x, y)\}^{2n},$$

where $\text{var}_\pi(t) = 1$ is assumed. The sharpness of the bounds is given by the following example. □

Example 1: *The sharpness of the upper and lower bounds.* Let (x, y) be distributed as bivariate normal with conditional distributions

$$x|y \sim N(\rho y, 1 - \rho^2), \quad y|x \sim N(\rho x, 1 - \rho^2).$$

Iterative draws alternating between x and y are performed. If the function of interest is $t(x) = x$, $\sigma_1(t)$ is ρ. Direct calculation shows that

$$\{\sigma_n(t)\}^2 = E(x_0, x_n) = E_\pi\{(E_\pi[\ldots E_\pi\{E_\pi(x|y)|x\}\ldots])^2\} = \rho^{2n} = \{\sigma_1(t)\}^{2n}.$$

Thus the lower bound is attained in this bivariate normal example.

On the other hand, Lancaster (1958) demonstrates that, for bivariate normal distributions, the maximal correlation between x and y is the absolute value of their correlation coefficient, ρ. Hence the upper bound is also attained.

4. MIXTURE AND EMPIRICAL ESTIMATORS

In practice, a choice has to be made among different estimators: the empirical estimator, the mixture estimator, or a combination of both. If the samples are drawn independently,

the superiority of the mixture estimator is obvious. The same conclusion holds in the situation of data augmentation where the draws are dependent.

THEOREM 4·1. *If a stationary Markov chain* $\{x_k, k=0,1,\ldots\}$ *has the interleaving Markov property with the conjugate chain* $\{y_k, k=0,1,\ldots\}$, *then for any function* $t(.)$ *of* x *we have*

$$\text{var}\,\{t(x_1)+\ldots+t(x_n)\} \geqslant \text{var}\,[E_\pi\{t(x)|y_1\}+\ldots+E_\pi\{t(x)|y_n\}]. \tag{12}$$

Therefore, in data augmentation, the mixture estimator (2) *is always better than the empirical estimator* (1). *The reduction of variance can be expressed as a linear function of the autocovariances* $\text{cov}\,\{t(x_0), t(x_m)\}$ $(m=0,\ldots,n)$.

Proof. Using Theorem 3·1 and the stationarity of both $\{x_k\}$ and $\{y_k\}$, we have

$$\text{cov}\,\{t(x_k), t(x_{k+m})\} = \text{var}_\pi\,\{E_\pi(\ldots E_\pi[E_\pi\{t(x)|y\}|x]|\ldots)\},$$

where there are m expectation signs. Correspondingly, if we write $E_\pi\{t(x)|y\}$ as $s^*(y)$, then

$$\text{cov}\,\{s^*(y_k), s^*(y_{k+m})\} = \text{cov}\,\{t(x_k), t(x_{k+m+1})\}. \tag{13}$$

Since $\text{cov}\,\{t(x_0), t(x_m)\}$ is a monotone decreasing function of m, it is obvious that

$$\text{cov}\,\{t(x_k), t(x_{k+m})\} \geqslant \text{cov}\,\{s^*(y_k), s^*(y_{k+m})\}.$$

This implies that any term in the quadratic expansion of

$$\text{var}\,[E_\pi\{t(x)|y_1\}+\ldots+E_\pi\{t(x)|y_n\}]$$

is smaller than or equal to the corresponding term in the expansion of

$$\text{var}\,\{t(x_1)+\ldots+t(x_n)\}.$$

Thus the superiority of (2) over (1) follows.

Using the notation in (11) of § 3·3 and omitting the argument of the σ, we have

$$n^2\,\text{var}\,(\hat{\mu}) = n\sigma_0^2 + 2(n-1)\sigma_1^2 + \ldots + 2\sigma_{n-1}^2,$$
$$n^2\,\text{var}\,(\tilde{\mu}) = n\sigma_1^2 + 2(n-1)\sigma_2^2 + \ldots + 2\sigma_n^2.$$

The reduction of the variance by using the mixture estimate is

$$\text{var}\,(\hat{\mu}) - \text{var}\,(\tilde{\mu}) = \frac{1}{n^2}\{n(\sigma_0^2-\sigma_1^2) + 2(n-1)(\sigma_1^2-\sigma_2^2) + \ldots + 2(\sigma_{n-1}^2-\sigma_n^2)\}.$$

In a sense, Rao–Blackwellization causes a one-lag delay for dependent samples, and will be especially beneficial when the chain mixes fast with respect to t. □

Taking any linear combination of the two estimators is not profitable.

LEMMA 4·1. *Let* $s^*(y) = E_\pi\{t(x)|y\}$. *Then for any* $m \geqslant 1$, *we have*

$$\text{cov}\,\{t(x_0), s^*(y_m)\} = \text{cov}\,\{s^*(y_0), s^*(y_m)\} = \sigma_{m+1}^2,$$
$$\text{cov}\,\{t(x_m), s^*(y_0)\} = \text{cov}\,\{s^*(y_0), s^*(y_{m-1})\} = \sigma_m^2.$$

The first equality also holds for $m=0$.

Proof. Noting that $E\{t(x_0)|y_0\} = s^*(y_0)$, we have

$$E\{t(x_0)s^*(y_m)\} = E[E\{t(x_0)s^*(y_m)|y_0, y_m\}]$$
$$= E[E\{t(x_0)|y_0\}s^*(y_m)] = E\{s^*(y_0)s^*(y_m)\}.$$

For cov $\{t(x_m), s^*(y_0)\}$, we notice that

$$E\{t(x_m)|y_{m-1}\} = E_\pi\{t(x)|y_{m-1}\} = s^*(y_{m-1}).$$

Thus

$$E\{t(x_m)s^*(y_0)\} = E[E\{t(x_m)s^*(y_0)|y_0, y_{m-1}\}]$$
$$= E[E\{t(x_m)|y_{m-1}\}s^*(y_0)] = E\{s^*(y_{m-1})s^*(y_0)\}.$$

The result then follows from (13). □

THEOREM 4·2. *If w_1 and w_2 are nonnegative weights with $w_1 + w_2 = 1$,*

$$\text{var}\,(w_1\hat{\mu} + w_2\tilde{\mu}) \geqslant \text{var}\,(\tilde{\mu}).$$

Proof. Writing $s^*(y) = E_\pi\{t(x)|y\}$ and applying Lemma 4·1, we obtain that

$$E\{t(x_k)\tilde{\mu}\} = \frac{1}{n}\sum_{m=1}^{n} \text{cov}\,\{t(x_k), s^*(y_m)\}$$
$$= \sigma^2_{k-1} + \sigma^2_{k-2} + \ldots + \sigma^2_1 + \sigma^2_1 + \ldots + \sigma^2_{n-k+1}.$$

Thus the covariance between $\hat{\mu}$ and $\tilde{\mu}$ is

$$\text{cov}\,(\hat{\mu}, \tilde{\mu}) = \frac{1}{n^2}\{n\sigma^2_1 + (n-1)(\sigma^2_1 + \sigma^2_2) + (n-2)(\sigma^2_2 + \sigma^2_3) + \ldots + (\sigma^2_{n-1} + \sigma^2_n)\}$$

$$\geqslant \frac{1}{n^2}\{n\sigma^2_1 + 2(n-1)\sigma^2_2 + \ldots + 2\sigma^2_n\} = \text{var}\,(\tilde{\mu}).$$

Therefore

$$\text{var}\,(w_1\hat{\mu} + w_2\tilde{\mu}) = w_1^2\,\text{var}\,(\hat{\mu}) + 2w_1w_2\,\text{cov}\,(\hat{\mu}, \tilde{\mu}) + w_2^2\,\text{var}\,(\tilde{\mu}) \geqslant \text{var}\,(\tilde{\mu}).$$ □

When $t(.)$ is not restricted to a function of one component, however, the mixture estimator can have larger variance.

Example 2. Let (x, y) be distributed as bivariate normal with conditional distributions

$$x|y \sim N(\rho y, 1-\rho^2), \quad y|x \sim N(\rho x, 1-\rho^2).$$

Suppose we are interested in estimating the mean of $t(x, y) = x - by$, where b is some constant, by Gibbs sampling. Let $X_1 = (x_1, y_1)$ and $X_2 = (x_2, y_2)$ be consecutive samples from the chain, and $s^*(y) = E_\pi\{t(X)|y\} \equiv (\rho - b)y$. Then

$$\text{var}\,\{t(X_1) + t(X_2)\} = 2 + 2b^2 - 6\rho b + 2(1+b^2)\rho^2 - 2b\rho^3,$$

$$\text{var}\,\{s^*(y_1) + s^*(y_2)\} = (\rho - b)^2(2 + 2\rho^2).$$

The difference between the two variances is

$$\text{var}\,\{t(X_1) + t(X_2)\} - \text{var}\,\{s^*(y_1) + s^*(y_2)\} = 2(1-\rho^2)(1 + \rho^2 - b\rho).$$

When $b \geqslant 2/\rho$, the variance of the empirical estimator of $E_\pi\{t(X)\}$ using two samples is less than that of the mixture estimator. □

5. Comparison of schemes corresponding to partitioning

In this section we discuss the problem of comparing Schemes [1], [2] and [3] introduced in § 1. Relative to [3], we will refer to Scheme [1] as collapsing, and Scheme [2] as grouping. We emphasize again that the distribution $\pi(x, y)$ is the same for all the three schemes, and the distribution $\pi(x, y, z)$ is the same for Schemes [2] and [3]. As mentioned before, the notation is simplified by writing $\pi(x, y) = \int \pi(x, y, z)\, dz$ in the above.

The transition functions corresponding to collapsing, grouping and the ordinary systematic scan are respectively:

Scheme [1],

$$K_1\{(x_0, y_0), (x_1, y_1)\} = \pi(x_1 | y_0)\pi(y_1 | x_1);$$

Scheme [2],

$$K_2\{(x_0, y_0, z_0), (x_1, y_1, z_1)\} = \pi(x_1 | y_0, z_0)\pi(y_1, z_1 | x_1);$$

Scheme [3],

$$K_3\{(x_0, y_0, z_0), (x_1, y_1, z_1)\} = \pi(x_1 | y_0, z_0)\pi(y_1 | x_1, z_0)\pi(z_1 | x_1, y_1).$$

The corresponding forward operators are denoted by F_1, F_2 and F_3.

Theorem 5·1. *The norms of the three corresponding forward operators have the following ordering*:

$$\|F_1\| \leqslant \|F_2\| \leqslant \|F_3\|.$$

Furthermore, the spectral radius of Scheme [1] *is less than or equal to that of Scheme* [2].

Proof. For any function $t(x)$ of only one component, we notice that

$$E_\pi\{t(x) | y\} = E_\pi[E_\pi\{t(x) | (y, z)\} | y].$$

Therefore

$$\mathrm{var}_\pi\,[E_\pi\{t(x) | y\}] \leqslant \mathrm{var}_\pi\,[E_\pi\{t(x) | (y, z)\}],$$

which implies that the maximal correlation between x and y is always smaller than that between x and (y, z). Hence the first inequality follows from Lemma 2·3 and Theorem 3·2. The statement about their spectral radii follows from Theorem 3·2.

For the second inequality, it is enough to prove that the backward operators satisfy $\|B_2\| \leqslant \|B_3\|$. The reason for considering the backward operators is that we can handle the reversed Schemes [2] and [3] easily. The transition functions for the reversed Schemes [2] and [3] are

$$K_2^*\{(x_0, y_0, z_0), (x_1, y_1, z_1)\} = \pi(y_1, z_1 | x_0)\pi(x_1 | y_1, z_1),$$

$$K_3^*\{(x_0, y_0, z_0), (x_1, y_1, z_1)\} = \pi(z_1 | y_0, z_0)\pi(y_1 | x_0, z_1)\pi(x_1 | y_1, z_1).$$

Hence, a relation between K_2^* and K_3^* is

$$\begin{aligned} K_2^*\{(x_0, y_0, z_0), (x_1, y_1, z_1)\} &= \pi(z_1 | x_0)\pi(y_1 | x_0, z_1)\pi(x_1 | y_1, z_1) \\ &= \int \pi(z_1 | x_0, y_0)\pi(y_1 | x_0, z_1)\pi(x_1 | y_1, z_1)\pi(y_0 | x_0)\, dy_0 \\ &= E_\pi[K_3^*\{(x_0, y_0, z_0), (x_1, y_1, z_1)\} | x_0]. \end{aligned}$$

For any square integrable function of three components $s(x, y, z)$,

$$\begin{aligned}\|B_2 s\|^2 &= E_\pi([E_2\{s(x_1, y_1, z_1)|x_0, y_0, z_0\}]^2)\\ &= E_\pi\{(E_\pi[E_3\{s(x_1, y_1, z_1)|x_0, y_0, z_0\}|x_0])^2\}\\ &\leqslant E_\pi([E_3\{s(x_1, y_1, z_1)|x_0, y_0, z_0\}]^2) = \|B_3 s\|^2,\end{aligned}$$

where E_l denotes the expectation taken under transition K_l^*, for $l = 2, 3$. Hence $\|F_2\| = \|B_2\| \leqslant \|B_3\| = \|F_3\|$. This completes the proof. □

By Lemma 2·3 and Theorem 3·2, the norm of the operator corresponding to Scheme [1] is the maximal correlation between x and y. Similarly for Scheme [2], the norm of F_2 is the maximal correlation between x and (y, z). It is clear that adding one more variable will increase the maximal correlation. Accordingly, the following suggestions can be given: if any pair of the three components are highly dependent, integrating one of them out is the best strategy; or, when integration is difficult, grouping these two together can also increase the convergence rate and reduce the autocovariances.

Although none of the three chains for comparison is reversible, the comparison between Schemes [1] and [2] is exact in a minimax sense; i.e. the autocovariances of the worst functional under Scheme [1] decay faster than those of Scheme [2]. A reason for such an exact comparison is that the marginal chains induced by [1] and [2] are reversible. Of course, it is possible to find functions orthogonal to the eigenspace generated by the largest eigenvalue so that, for these deliberately selected functions, Scheme [2] gives a better estimate than Scheme [1]. Since [3] is not reversible, and its marginal chains are not even Markov, its comparison with [2] is less clear, as the following example demonstrates.

Example 3. Let $(x, y, z) \sim N(0, \Sigma)$ with

$$\Sigma = \begin{pmatrix} 1 & 0{\cdot}5 & 0{\cdot}5 \\ 0{\cdot}5 & 1 & \rho \\ 0{\cdot}5 & \rho & 1 \end{pmatrix}.$$

By Theorem 3·2, spec (F_2) is the square of the maximal correlation between x and (y, z):

$$\text{spec}\,(F_2) = \frac{1}{2(1+\rho)}.$$

Following Amit (1991), we denote Σ^{-1} by Q, where $Q = (q_{ij})$. Then the spectral radius of F_3 is the largest norm of the eigenvalues of

$$(I - D_1 Q)(I - D_2 Q)(I - D_3 Q),$$

where D_i is the matrix with all entries zero except that the ith entry of the diagonal is q_{ii}^{-1}. In particular, when $\rho = 0$, the spectral radius of F_3 is $\frac{1}{3} < 0{\cdot}5 = \text{spec}\,(F_3)$. However,

$$\|F_2\| = 0{\cdot}707 < 0{\cdot}72 = \|F_3\|$$

still holds. Further computation shows that spec $(F_2) >$ spec (F_3) when $\rho < 0{\cdot}25$.

If one symmetrizes chain [3] which physically corresponds to a sampler iterating one pass forward and another backward, the corresponding symmetrized forward operator is

B_3F_3, which is self-adjoint. Since $\|B_3F_3\|$ equals $\|F_3\|^2$, and also equals the spectral radius of B_3F_3, our results show that the symmetrized Scheme [3] is worse than the symmetrized Scheme [2] in mixing rates.

6. An example

To conclude, one more example is presented to illustrate that the choice of iterating scheme can depend on whether the mixture estimate is used. This ties together the two main results presented in this paper.

Let $(x, y, z) \sim N(0, \Sigma)$ with

$$\Sigma = \begin{pmatrix} 1 & \sqrt{0{\cdot}1} & \sqrt{0{\cdot}8} \\ \sqrt{0{\cdot}1} & 1 & 0 \\ \sqrt{0{\cdot}8} & 0 & 1 \end{pmatrix}.$$

Suppose we want to estimate $\mu = E_\pi(x)$ using Gibbs sampling and may either apply Scheme [1] or Scheme [2] defined in § 5. The two relevant maximal correlations are

$$\gamma_\pi(x, y) = \sqrt{0{\cdot}1}, \quad \gamma_\pi\{x, (y, z)\} = (0{\cdot}1 + 0{\cdot}8)^{\frac{1}{2}} = \sqrt{0{\cdot}9}.$$

From Theorem 3·2 and Lemma 2·3, the mixing rates of the two schemes are $\|F_{1x}\| = 0{\cdot}1$ and $\|F_{2x}\| = 0{\cdot}9$. Let $(x_{k[1]}, y_{k[1]})$ $(k = 1, \ldots, n)$ be samples generated from Scheme [1], and $(x_{k[2]}, y_{k[2]}, z_{k[2]})$ $(k = 1, \ldots, n)$, be samples generated using Scheme [2]. Let

$$\hat{\mu}_{[1]} = \frac{1}{n}\sum_{k=1}^{n} x_{k[1]}, \quad \hat{\mu}_{[2]} = \frac{1}{n}\sum_{k=1}^{n} x_{k[2]}$$

be the two corresponding empirical estimates of $E_\pi(x)$. Then, for large n,

$$n \operatorname{var}(\hat{\mu}_{[1]}) \doteq 1 + 2\sum_{k=1}^{\infty} (0{\cdot}1)^k = 1{\cdot}222,$$

$$n \operatorname{var}(\hat{\mu}_{[2]}) \doteq 1 + 2\sum_{k=1}^{\infty} (0{\cdot}9)^k = 19.$$

This implies that $\hat{\mu}_{[1]}$ is about $19/1{\cdot}222 = 15{\cdot}5$ times more efficient than $\hat{\mu}_{[2]}$. However, suppose that, for each iteration, running Scheme [1] is 20 times as expensive as running Scheme [2], then, using the empirical estimates for comparisons, Scheme [2] is actually more efficient than Scheme [1]. Now, consider the mixture estimates

$$\tilde{\mu}_{[1]} = \frac{1}{n}\sum_{k=1}^{n} E_\pi(x \mid y_{k[1]}), \quad \tilde{\mu}_{[2]} = \frac{1}{n}\sum_{k=1}^{n} E_\pi(x \mid y_{k[2]}, z_{k[2]}).$$

Applying Theorem 4·1, we get

$$n \operatorname{var}(\tilde{\mu}_{[1]}) \doteq 0{\cdot}1 + 2\sum_{k=2}^{\infty} (0{\cdot}1)^k = 1{\cdot}222 \times 0{\cdot}1 = 0{\cdot}1222,$$

$$n \operatorname{var}(\hat{\mu}_{[2]}) \doteq 0{\cdot}9 + 2\sum_{k=2}^{\infty} (0{\cdot}9)^k = 19 \times 0{\cdot}9 = 17{\cdot}1.$$

For the same n, $\tilde{\mu}_{[1]}$ is about $17{\cdot}1/0{\cdot}1222 = 140$ times more efficient than $\tilde{\mu}_{[2]}$. So,

Scheme [1] is to be preferred even if it is 20 times more expensive than Scheme [2] per iteration. The key here is that using the mixture estimate instead of the empirical estimate leads to a substantial improvement in efficiency with Scheme [1], but not with Scheme [2]. Although this example is artificial, the point it illustrates is valid in a very general setting. A fast mixing scheme gains an extra factor in efficiency if the mixture estimate can be easily computed.

ACKNOWLEDGEMENT

The authors are grateful for the helpful comments of the referees, the two associate editors, and the editor.

REFERENCES

AMIT, Y. (1991). On rates of convergence of stochastic relaxation for Gaussian and non-Gaussian distributions. *J. Mult. Anal.* **38**, 82–99.

BHATTACHARYA, R. N. & WAYMIRE, E. C. (1990). *Stochastic Processes with Applications*. New York: Wiley.

BRADLEY, R. (1986). Basic properties of strong mixing conditions. In *Dependence in Probability and Statistics*, Ed. E. Eberlein and M. S. Taqqu, pp. 165–92. Boston: Birkhaüsser.

BREIMAN, L. & FRIEDMAN, J. H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *J. Am. Statist. Assoc.* **80**, 580–619.

CSÀKI, P. & FISCHER, J. H. (1960). Contributions to the problem of maximal correlation. *Magy. Tud. Akad., Budapest, Mat. Kotato Intezet, Kozl.* **5**, 325–37.

DIACONIS, P. (1988). *Group Representations in Probability and Statistics*. Hayward: IMS.

GELFAND, A. E. & SMITH, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *J. Am. Statist. Assoc.* **85**, 398–409.

GEMAN, S. & GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intel.* **6**, 721–41.

HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

LANCASTER, H. O. (1958). The structure of bivariate distributions. *Ann. Math. Statist.* **29**, 719–36.

LI, K. H. (1988). Imputation using Markov chains. *J. Statist. Computat. Simul.* **30**, 57–79.

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. & TELLER, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–91.

SARMANOV, O. V. (1958). The maximal correlation coefficient. *Dokl. Akad. Nauk USSR* **120**, 715–8.

TANNER, M. A. & WONG, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Am. Statist. Assoc.* **52**, 528–50.

YOSIDA, K. (1980). *Functional Analysis*, 6th ed. New York: Springer-Verlag.

[*Received April* 1991. *Revised July* 1993]