Covariance Structure and Convergence Rate of the Gibbs Sampler with Various Scans

Jun S. Liu, Wing H. Wong, Augustine Kong


Your use of the JSTOR database indicates your acceptance of JSTOR’s Terms and Conditions of Use. A copy of JSTOR’s Terms and Conditions of Use is available at http://www.jstor.org/about/terms.html, by contacting JSTOR at jstor-info@umich.edu, or by calling JSTOR at (888)388-3574, (734)998-9101 or (FAX) (734)998-9113. No part of a JSTOR transmission may be copied, downloaded, stored, further transmitted, transferred, distributed, altered, or otherwise used, in any form or by any means, except: (1) one stored electronic and one paper copy of any article solely for your personal, non-commercial use, or (2) with prior written permission of JSTOR and the publisher of the article or other text.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

Journal of the Royal Statistical Society. Series B (Methodological) is published by Royal Statistical Society. Please contact the publisher for further permissions regarding the use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/rss.html.

Journal of the Royal Statistical Society. Series B (Methodological)
©1995 Royal Statistical Society

JSTOR and the JSTOR logo are trademarks of JSTOR, and are Registered in the U.S. Patent and Trademark Office. For more information on JSTOR contact jstor-info@umich.edu.

©2000 JSTOR
Covariance Structure and Convergence Rate of the Gibbs Sampler with Various Scans

By JUN S. LIU†, WING H. WONG and AUGUSTINE KONG

Harvard University, Cambridge, USA
University of Chicago, USA

[Received September 1991. Final revision August 1993]

SUMMARY
This paper presents results on covariance structure and convergence for the Gibbs sampler with both systematic and random scans. It is shown that, under conditions that guarantee the compactness of the Markov forward operator and irreducibility of the corresponding chain, the Gibbs sampling scheme converges geometrically in terms of Pearson \(\chi^2\)-distance. In particular, for the random scan, the autocovariance can be expressed as variances of iterative conditional expectations. As a consequence, the autocorrelations are all positive and decrease monotonically.

Keywords: AUTOCOVARIANCE; COMPACT OPERATOR; FORWARD OPERATOR; GIBBS SAMPLER; MAXIMAL CORRELATION; PEARSON \(\chi^2\)-DISTANCE

1. INTRODUCTION
As a powerful computational tool, the Gibbs sampler, whose full description can be found in Gelfand and Smith (1990), has a great impact on current statistical practice. This paper addresses a few theoretical issues concerning the algorithm. Firstly, the covariance structures of two sampling schemes are studied and related to the convergence rates. As a result, an explicit expression for the autocorrelations of a random scan is obtained. Secondly, by making use of simple inequalities that connect the convergence rate and the covariance structure, we obtain conditions for the monotone geometric convergence in terms of Pearson’s \(\chi^2\)-pseudodistance. Furthermore, we briefly explain connections between our results and maximal correlation and \(\rho\)-mixing. In Liu et al. (1994), and Liu (1994), it is shown that such connections can be used to compare various sampling schemes and estimators.

In Geman and Geman (1984), geometric convergence was obtained for a finite state space assuming that the equilibrium distribution satisfies the positivity condition (Besag, 1974). Tanner and Wong (1987) studied the method of data augmentation where only two random vectors are iteratively sampled. They proved the convergence of the scheme and also claimed a geometric rate of convergence for their method under mild conditions. However, their proof of the geometric rate of convergence contained an error. Schervish and Carlin (1993), by extending the line of arguments in Tanner and Wong (1987), provided a proof of geometric convergence for the Gibbs sampler with systematic scan. Compared with Schervish and Carlin’s (1992) assumptions, our conditions are weaker and the results are more general. The methods employed are also different. Independently, Li (1988)

†Address for correspondence: Department of Statistics, Science Center, Harvard University, 1 Oxford Street, Cambridge, MA 02138, USA.

© 1995 Royal Statistical Society

0035–9246/95/57157
implemented a variation of the Gibbs sampler and also pointed out the convergence property for a discrete state space. The recent work by Tierney (1991) provided some results on the convergence rate under the total variation distance, by making use of the Markov chain theory developed by Nummelin (1984), Asmussen (1987) and others. In contrast, our approach relies only on some elementary facts about linear functional analysis. The type of convergence in our results is also stronger.

Section 2 provides basic definitions and preliminaries. Section 3 studies the systematic scan where the updating of the variables follows a fixed deterministic order. Section 4 is on the random scan, where explicit expressions and bounds for autocovariances are given. We establish geometric convergence for both scans. Applications of our results to a Gaussian model and a covariance inference problem are given in the last section.

2. DEFINITIONS AND BASICS

Let $X_0, X_1, \ldots, X_n$ be consecutive samples taken from a time homogeneous Markov chain with equilibrium distribution $\pi(X)$ and transition function $K(X | X_0)$. Throughout the paper, $E(\ )$ and $\text{var}(\ )$ denote the expectation and variance taken under the equilibrium distribution $\pi$, while $E_p(\ )$ is the expectation taken under the probability measure $p$. We begin by establishing some necessary definitions and notation.

The Hilbert space of mean 0, square integrable, complex-valued functions of $X$ is denoted by

$$L_0^2(\pi) = \{ t(X): E\{t(X)\} = 0, \text{ and } E|t(X)|^2 < \infty \}$$

with the inner product

$$\langle t(X), s(X) \rangle = E\{t(X)\overline{s(X)}\},$$

where $\overline{s}$ denotes the complex conjugate of $s$ and $|c|$ is the modulus of a complex number $c$. The variance of a complex random variable $t(X)$ in $L_0^2(\pi)$ is then defined to be $\|t\|^2 = \langle t, t \rangle$, which is the sum of the respective variances of the real and imaginary parts of $t$. The norm of an operator $A$ on $L_0^2(\pi)$ is defined by

$$\|A\| = \sup_{t \in L_0^2(\pi), \|t\| = 1} \|At\|.$$

**Definition 1.** Two operators $F$ and $B$ where $F$ stands for 'forward' and $B$ stands for 'backward' are defined as

$$F t(X_1) \overset{def}{=} E\{t(X_2) | X_1\} = \int t(Y) K(Y | X_1) \, dY,$$  \hspace{1cm} (1)

$$B t(X_2) \overset{def}{=} E\{t(X_1) | X_2\} = \int t(X) \frac{K(X_2 | X) \pi(X)}{\pi(X_2)} \, dX.$$  \hspace{1cm} (2)

Clearly, $F$ and $B$ are operators from $L_0^2(\pi)$ to itself. It follows from their definitions that $F$ and $B$ are adjoint operators, i.e. $\langle Ft, s \rangle = \langle t, Bs \rangle$. By the Markov property, it is true that $F^n t(X_0) = E\{t(X_n) | X_0\}$ and $B^n t(X_n) = E\{t(X_0) | X_n\}$. Therefore the following lemma holds.
Lemma 1. For any $t$ and $s$ in $L_0^2(\pi)$, \( \text{cov}\{t(X_n), s(X_0)\} = \langle \mathbf{F}^n t, s \rangle = \langle \mathbf{F}^{n-k} t, \mathbf{B}^k s \rangle \).

If the underlying Markov chain is reversible, i.e. the so-called detailed balance condition $K(Y|X) \pi(X) = K(X|Y) \pi(Y)$ is satisfied, then $\mathbf{F} = \mathbf{B}$, and all even-lag autocovariances are non-negative, i.e.

$$
\text{cov}\{t(X_0), t(X_{2m})\} = E[|\mathbf{F}^m t(X)|^2] = E[|\mathbf{B}^m t(X)|^2] \geq 0.
$$

Definition 2. The Pearson $\chi^2$-distance from density $p(X)$ to $\pi(X)$, $d_\pi(p, \pi)$, is

$$
d_\pi^2(p, \pi) \overset{\text{def}}{=} \text{var}\{p(X)/\pi(X)\} = \int p^2(X)/\pi(X) \, dX - 1.
$$

$d_\pi(\ , \ )$ is not a true distance. It is rather a measure of discrepancy between any distribution $p$ and the target distribution $\pi$, and it is a stronger measure than both the $L^1$-distance and a kind of Kullback–Leibler information distance. Some detailed discussions on the comparisons are provided in remark 1 of Section 3.

3. SYSTEMATIC SCAN GIBBS SAMPLER

Let $X = \{x(1), \ldots, x(d)\}$ consist of $d$ components, each of which can be multi-dimensional. The systematic scan is defined by the updating scheme

$$
x(1) \rightarrow x(2) \rightarrow \ldots \rightarrow x(d),
$$

where $(1, \ldots, d)$ is an ordering of the components fixed in advance. We are interested in sampling from the joint density of $X = \{x(1), \ldots, x(d)\}$, to evaluate $E\{t(X)\}$ for some square integrable function $t(X)$. The Gibbs sampler provides a way to achieve this by generating a Markov chain according to the transition function

$$
K(Y|X) = \pi\{y(1) | X^{[-1]}\} \pi\{y(2) | X^{[-1,2]}\} \cdots \pi\{y(d) | Y^{[-d]}\}, \tag{3}
$$

where $X^{[-1]} = \{x(i) : i \in A^c\}$ denotes components of $X$ excluding those components indicated in the set $A$, and $\pi(\ | \ )$ denotes all the conditional distributions generated by the target distribution $\pi$. In other words, to implement this scan, we first draw $x(1)$ conditioned on the current states of the other components, then draw $x(2)$ the same way, then $x(3)$, etc., until $x(d)$. After these $d$ updates, we say that our Markov chain has moved one step. The whole process is then repeated for the next step of the Markov chain. It is easy to show that $\pi$ is an invariant distribution for the chain. An estimate of $E\{t(X)\}$ can be obtained by taking the average of $t(X_i)$ over consecutive steps of the chain, perhaps after discarding an initial segment when the chain is started far from the equilibrium. In practice, other scans are often used. For example, when the system has a nearest neighbour Markov structure, the coding set method (see Besag (1974), section 6) reduces a many component Gibbs sampler to a two-component Gibbs sampler. In the remainder of this section, $\mathbf{F}_t$ denotes the forward operator corresponding to the transition function defined in equation (3).

For the argument in this section, we assume that $p_0(\ )$ is the density of an initial $X_0$ and denote the density of $X_n$ as $p_n(\ )$, which can be obtained recursively as
\[ p_n(Y) = \int K(Y|X) p_{n-1}(X) \, dX, \quad n = 1, 2, \ldots. \]

The following conditions are needed for the results later.

(a) After a finite number \( n_0 \) of iterations the Pearson \( \chi^2 \)-distance from the density \( p_{n_0}(X) \) to the target density \( \pi \) is finite.

In particular, if the chain is started from a fixed point within the support of \( \pi \), this condition will be satisfied \( \pi \) almost surely if the following condition holds. See comment (i) below.

(b) For the transition function \( K \) defined in equation (3),
\[ \int \left\{ \frac{K(Y|X)}{\pi(Y)} \right\}^2 \pi(X) \, \pi(Y) \, dX \, dY < \infty. \tag{4} \]

Condition (b) can also be written in a simpler form as \( \int \pi^2(X, Y)/\pi(X) \, \pi(Y) \, dX \, dY < \infty \), where \( \pi(X, Y) \) is the joint distribution of the two consecutive states \( X \) and \( Y \) of the stationary Markov chain.

(c) The Markov chain defined by equation (3) is \( \pi \) irreducible.

3.1. Comments on Conditions

(i) If we start from a fixed point in the support of \( \pi \) (the case of running a single chain), then, under condition (b), condition (a) is satisfied \( \pi \) almost surely. To see this, we rewrite condition (b) as
\[ \int \left\{ K^2(Y|X_0)/\pi(Y) \right\} \pi(X_0) \, dX_0 < \infty, \]
which implies that \( \int K^2(Y|X_0)/\pi(Y) \, dY < \infty \) for \( \pi \) almost every starting point \( X_0 \).

(ii) Condition (b) is a regularity condition that has been used by many researchers, e.g. Breiman and Friedman (1985) and Schervish and Carlin (1992). It is standard but not easy to check and understand. Intuitively this condition reflects the dependence between two consecutive states \( X \) and \( Y \) and also provides some restrictions on the shape of the support region of the joint probability distribution. For example, a 'horn'-shaped region does not work. Some kinds of degeneracy of the distribution are prohibited. See Csàki and Fischer (1960) for some examples.

When there are only two components, i.e. \( X = \{x(1), x(2)\} \), condition (b) can also be written as
\[ \int \pi\{x(1)|x(2)\} \pi\{x(2)|x(1)\} \, dX < \infty. \]

(iii) Condition (c) requires us to check whether the Gibbs sampler has a non-zero probability of moving from one region with positive \( \pi \)-measure to another. As pointed out by a referee, in the presence of conditions (a) and (b), this condition is equivalent to

(d) there exists no non-constant function \( t(X) \) satisfying
\[ E\{t(X) \mid X^{(i)}\} = t(X) \quad \text{almost everywhere} \ \forall \ i. \tag{5} \]
As was illustrated by Tierney (1991), theorem 3.6 and proposition 3.9 of Nummelin (1984) show that, if the chain is \( \pi \) irreducible and has \( \pi \) as an invariant probability distribution, then it has to be recurrent. By proposition 3.13 of Nummelin (1984), functions satisfying condition (5) are known to be almost everywhere constant. However, under conditions (a), (b) and (d), geometric convergence can be obtained as shown in the proof of theorem 1 (later), and thus irreducibility is obtained as well.

In some papers (Geman and Geman, 1984; Schervish and Carlin, 1992), the positivity condition (Besag, 1974) which requires the support of the stationary distribution to be the Cartesian product of the marginal supports is used to ensure irreducibility. In general, positivity implies, and is stronger than, conditions (c) or (d) here.

### 3.2. Convergence Result

**Lemma 1.** Condition (b) implies that the forward operator \( F_s \) corresponding to the transition function (3) is compact.

**Proof.** For any \( t \in L_2^\beta(\pi) \),

\[
F_s t(X) = \int t(Y) K(Y|X) \, dY = \int t(Y) \frac{K(Y|X)}{\pi(Y)} \pi(Y) \, dY
\]

is an integral operator with the kernel \( G(X, Y) = K(Y|X)/\pi(Y) \). By condition (b), we have \( \int G^2(X, Y) \pi(X) \pi(Y) \, dX \, dY < \infty \), which is a sufficient condition for the compactness of \( F_s \) (see Yosida (1980), example 2 of section X.2).

**Lemma 2.** Conditions (b) and (d) imply that the spectral radius of \( F_s \) is strictly less than 1, i.e. the supremum modulus of the eigenvalues is smaller than 1.

**Proof.** Because the norm of \( F_s \) is less than or equal to 1, it is automatic that the spectral radius of \( F_s \) is less than or equal to 1. By the compactness of \( F_s \), the spectrum is countable with 0 as the only possible point of accumulation. Hence there is an eigenfunction corresponding to the eigenvalue \( \lambda_i \) with the largest modulus. If \( |\lambda_1| = 1 \), then

\[
F_s t(X) = \lambda_1 t(X), \quad \text{almost everywhere},
\]

which implies

\[
\text{var}\{F_s t(X)\} = \text{var}\{t(X)\}. \tag{6}
\]

It follows from equation (3) that \( F_s t(X) = E[E[\ldots E[t(X)|X^{-[1]}]|X^{-[2]}]\ldots |X^{-[d]}] \). Note that for any two complex random variables \( W \) and \( Z \)

\[
\text{var}(W) = \text{var}\{E(W|Z)\} + \text{var}\{W - E(W|Z)\}. \tag{7}
\]

(The proof of this identity is the same as that for real random variables.) Hence

\[
\text{var}\{t(X)\} \geq \text{var}\{E[t(X)|X^{-[1]}]\} \geq \ldots \geq \text{var}(E[\ldots E[t(X)|X^{-[1]}]\ldots |X^{-[d]}]).
\]

Because of equation (6), all the inequalities must be equalities. In particular, equation (7) together with the equality \( \text{var}\{t(X)\} = \text{var}\{E[t(X)|X^{-[1]}]\} \) imply that \( t(X) = E[t(X)|X^{-[1]}] \), almost everywhere. It then follows from induction that \( t(X) = \)
$E\{t(X) \mid X^{-[i]}\}$, almost everywhere, for all $i$. This contradicts condition (d) and hence all the eigenvalues must have moduli smaller than 1.

**Theorem 1.** Let the starting distribution for the systematic scan Gibbs sampler be $p_0(X)$, and let $p_n(X)$ be defined as before; then under conditions (a), (b) and (c) the Pearson $\chi^2$-distance from $p_n$ to $\pi$ is monotone decreasing at a geometric rate as $n$ increases. Furthermore, the autocorrelation between $t(X_0)$ and $t(X_n)$ converges to 0 at a geometric rate.

**Proof.** Denote by $K^{(n)}$ the $n$-step transition function, which is obtained inductively by

$$K^{(n)}(Y \mid X) = \int K(Y \mid X') K^{(n-1)}(X' \mid X) \, dX'.$$

Without loss of generality, we assume $n_0 = 0$ in condition (a). Let $g(X) = p_0(X) / \pi(X) - 1$; then for any function $t \in L_0^2(\pi)$

$$|E_{p_n}\{t(X)\} - E\{t(X)\}| = \left| \int t(Y) K^{(n)}(Y \mid X) \{p_0(X) / \pi(X) - 1\} \pi(X) \, dX \, dY \right|$$

$$= |\langle F_n^* t, g \rangle| \leq \|F_n^*\| \|t\| \|g\|. \tag{8}$$

Let $t(X) = p_n(X) / \pi(X) - 1$; then the left-hand side of these expressions is $d^2_\pi(p_n, \pi)$. Since $\|g\| = d_\pi(p_0, \pi)$, it follows that

$$d_\pi(p_n, \pi) \leq \|F_n^*\| d_\pi(p_0, \pi). \tag{9}$$

Monotonicity follows easily from noting that $\|F_n\| \leq 1$ and substituting $p_0$ by $p_{n-1}$ in the above derivation. The geometric rate follows from

$$\lim_{n \to \infty} \|F_n\|^{1/n} = r, \tag{10}$$

(see Yosida (1980), section VIII.2), where $r$ is the spectral radius of $F_n$, which is less than 1 (lemma 2). The result for autocorrelation follows from this and lemma 1.

**Remark 1.** It follows from the Cauchy–Schwarz inequality that

$$\|p - \pi\|_{L_1} \leq \int \left| \frac{p(X) - \pi(X)}{\sqrt{\pi(X)}} \right| \sqrt{\pi(X)} \, dX \leq \left\{ \int \frac{(p(X) - \pi(X))^2}{\pi(X)} \, dX \right\}^{1/2} = d_\pi(p, \pi).$$

Hence, if the Pearson $\chi^2$-distance between $p$ and $\pi$ is small, so will be their total variation distance. In this sense, the convergence of the Pearson $\chi^2$-distance is clearly stronger than the convergence of the total variation distance. It is especially useful in bounding a tail probability. For example, for any event $A$, $|p(A) - \pi(A)| \leq \|p - \pi\|_{L_1}/2$, whereas, using the Cauchy–Schwarz inequality again, $|p(A) - \pi(A)| \leq \sqrt{\pi(A)} d_\pi(p, \pi)$. Hence, if $\pi(A)$ is small, the Pearson distance can provide a much sharper bound on $p(A)$.

A similar inequality holds between the comparison of Pearson $\chi^2$-distance and the reversed Kullback–Leibler information distance:

$$d_{KL}(p, \pi) \overset{\text{def}}{=} E_p[\log\{p(X)/\pi(X)\}] \leq \int p^2(X)/\pi(X) \, dX - 1 = d^2_\pi(p, \pi).$$

The inequality follows from the fact $\log u \leq u - 1$. 

Remark 2. As demonstrated in the proof, the technique of ‘reversiblization’ adopted by both Schervish and Carlin (1992) and Fill (1991) to construct an auxiliary self-adjoint operator from the operator $F_s$ is not necessary for proving geometric convergence in our case.

Remark 3. Some connections with the classical concepts of maximal correlation or $\rho$-mixing can be drawn here. For two random variables $X$ and $Y$, which may be multidimensional, the maximal correlation $\gamma$ between them is defined as

$$\gamma(X, Y) = \sup \sup \left| \text{corr} \{f(X), g(Y)\} \right|,$$

where the supremum is taken over all functions with finite variances. If we denote the maximal correlation between $X_0$ and $X_n$ of the chain by $\gamma_n$, then an interesting relation is that $\gamma_n = \|F^n_0\|$; hence $\|F^n_0\|$ describes the strength of the Markovian dependence. This suggests that we could estimate the lag $n$ canonical correlation of the chain by using successive samples of the Gibbs sampler to obtain the convergence rate. It follows from the Markov property that $\gamma_n$ either tends to 0 exponentially fast or stays at 1 (Bradley, 1986), which in our case reflects the fact that the spectral radius of $F_s$ is either smaller than 1 or equal to 1.

Remark 4. Peligrad’s (1986) theorem 2.3 says that a stationary $\rho$-mixing sequence satisfies a central limit theorem. Also, it is known from a coupling argument that the asymptotics of a geometric $\rho$-mixing Markov chain do not depend on its starting distribution. (Imagine a Markov chain $X_0, X_1, \ldots$, started from density $\rho_0$ coupled with a chain $Y_0, Y_1, \ldots$, started from $\pi$; we see that the Pearson distance between the joint distributions of $(X_{m+1}, \ldots, X_{m+n})$ and $(Y_{m+1}, \ldots, Y_{m+n})$ are geometrically small as $m$ increases, regardless of $n$.) Thus, by proving geometric $\rho$-mixing of the chain, theorem 1 also guarantees the asymptotic normality of $\hat{t} = \{t(X_1) + \ldots + t(X_n)\}/n$ for a function $t(X)$ with finite variance.

4. RANDOM SCAN

The convergence property of the random scan in the discrete case has been derived previously, for example, by Geman and Geman (1984). However, a general result is not available. The scan can be described as follows.

At each iteration step, we independently draw an index $i$ according to a pre-assigned distribution $V = (\alpha_1, \ldots, \alpha_d)$ on the index set $I = \{1, \ldots, d\}$, then replace the value of the random variable $x(i)$ corresponding to that index by a new sample drawn from the conditional distribution $\pi(x(i) | X_{[i]}^{[-i]})$. The distribution $V$ need not be uniform, but we require that $\alpha_i > 0$ for all $i$. We can easily show that $\pi$ is invariant under this transition. Throughout the section, $F_r$ denotes the forward operator of this scan.

4.1. Covariance Structure

It is known that the Gibbs sampler with random scanning satisfies the detailed balance relation and therefore generates a reversible Markov chain. Besides the non-negative even-lag autocovariances guaranteed by the reversibility of the chain, the non-negativity of odd-lag autocovariances can be obtained as well. Furthermore,
they can be expressed as the variances of some iterative conditional expectations. To establish these properties, we first look at the lag 1 autocovariance.

**Lemma 3.** Let \( X_0 \) and \( X_1 \) be two consecutive realizations of the random scan Gibbs sampler under stationarity and \( i \) be the random variable representing which index is updated at stage 1, taking values on \( I = \{1, \ldots, d\} \) with distribution \( V \). Then for any \( t(X) \in L^2_0(\pi) \)

\[
\text{cov}\{t(X_0), t(X_1)\} = E\left[ \sum_{i=1}^{d} \alpha_i E^2\{t(X)|X^{i-1}\} \right] = E[ E^2\{t(X)|i, X^{i-1}\} ] \geq 0.
\]

Hence the corresponding forward operator is self-adjoint and positive.

**Proof.** From the definition of the scan, it is understood that

\[
F_r t(X_0) = E[ E\{t(X_1)|i, X_0\}|X_0] = \sum_{i=1}^{d} \alpha_i E\{t(X)|X^{i-1}\}. \tag{11}
\]

This expression, together with a conditional expectation argument, gives

\[
E\{t(X_0) t(X_1)\} = E( E\{t(X_0) t(X_1)|i, X_0\}|X_0) = \sum_{i=1}^{d} \alpha_i E[ E\{t(X_0) t(X_1)|i = i, X^{i-1}\} ]
\]

\[
= E\left[ \sum_{i=1}^{d} \alpha_i E^2\{t(X_1)|X^{i-1}\} \right] = E[ E^2\{t(X)|i, X^{i-1}\} ] \geq 0.
\]

The second equality follows from our understanding of the random scan; the third equality is true because, conditioned on a chosen updating index \( i = i \) and fixed values of the corresponding components, \( X_0 \) and \( X_1 \) are independent and identically distributed under stationarity. The self-adjointness of \( F_r \) follows from the reversibility of the chain; the positivity of \( F_r \) follows from \( \langle F_r, t, t \rangle = \text{cov}\{t(X_1), t(X_0)\} \geq 0 \), which is the definition of positivity.

The lemma suggests setting \( \alpha_i \) small when \( E[ E^2\{t(X)|X^{i-1}\} ] \) is large. Since \( t(X) \) has mean 0, the displayed quantity is also equal to \( \text{var}\{t(X)\} - E[\text{var}\{t(X)|X^{i-1}\}] \). Hence \( \alpha_i \) should be set small if \( E[\text{var}\{t(X)|X^{i-1}\}] \) is small, which can be understood that we should make fewer visits to a component that is less variable.

**Theorem 2.** Let \( X_0, X_1, \ldots, \) be consecutive samples generated by the random scan under stationarity, and let \( i \) be the random variable representing the random index in the updating scheme. For \( t(X) \in L^2_0(\pi) \), the autocovariance between \( t(X_0) \) and \( t(X_n) \) is a non-negative monotone decreasing function of \( n \). It can be written as

\[
\text{cov}\{t(X_0), t(X_n)\} = \text{var}\{ E(\ldots E[ E\{t(X)|i, X^{i-1}\}|X]|X) \ldots )\}, \tag{12}
\]

where there are \( n \) conditional expectations taken alternately on \( \{i, X^{i-1}\} \) and \( X \).
**Proof.** The expression is derived by repeatedly applying lemma 3 and the Markov property. The monotonicity is a simple property of conditional expectations.

**Corollary 1.** If $t(X) \in L_0^0(\pi)$, then $r_r^*(t) \leq r_n(t) \leq \|F_r\|^n$, where $\|F_r\|$ is the norm of the forward operator, and $r_n(t) = \text{corr}(t(X_0), t(X_n))$.

**Proof.** Without loss of generality, we set $\|t\| = 1$. Then, by lemma 1, $r_n(t) = \langle F_r^n t, t \rangle$, and the second inequality is immediate. Furthermore, by lemma 3, $F_r$ is positive. Thus there is a positive, self-adjoint operator $G$ such that $G^2 = F_r$. Hence $r_n(t) = \|G^n t\|^2$. Since $\|G^n t\|^2 = \langle G^n t, G^n t \rangle = \langle G^{n+k} t, G^{n-k} t \rangle \leq \|G^{n+k} t\| \|G^{n-k} t\|$, letting $n = k = 1$ gives us the inequality $r_r^*(t) \leq r_r^2(t)$, and an induction argument shows that $r_r^*(t) \leq r_n(t)$.

4.2. Geometric Convergence

The forward operator $F_r$ of the random scan is in general not guaranteed to be compact. Therefore, the method in Section 3 is not directly applicable. However, by making use of the results in Section 3, we can still prove that under the same conditions (b) and (c) the norm of $F_r$ is strictly less than 1. This is done without showing its compactness. Intuitively $F_r$ is understood as a weighted mixture of different kinds of $F_s$'s corresponding to different visiting orderings. Thus, if one of the $F_s$'s norm is less than 1, the norm of the mixture will be less than 1 as well.

**Theorem 3.** Under conditions (b) and (c), we have $\|F_r\| < 1$.

**Proof.** If we define the operators

$$A_i = E( X^{i-1} | X^{-1}) ,$$

then $F_r = \alpha_1 A_1 + \alpha_2 A_2 + \ldots + \alpha_d A_d$. Thus

$$F_r^d = \sum_{j_1, \ldots, j_d} \left( \prod_{k=1}^d \alpha_{j_k} \right) A_{j_1} \ldots A_{j_d} .$$

Condition (b) and lemma 1 imply that

$$F_s = A_d A_{d-1} \ldots A_1$$

is a compact operator; condition (c) and lemma 2 guarantee a spectral radius smaller than 1. Hence, there exists $n_0$ such that $\|F_r^{n_0}\| < 1$. Since all $A_i$s satisfy $\|A_i\| \leq 1, \forall i$,

$$\|(F_r^d)^n_0\| \leq 1 - \left( \prod_{k=1}^d \alpha_k \right)^{n_0} + \left( \prod_{k=1}^d \alpha_k \right)^n_0 \|F_r^{n_0}\| < 1. \quad (13)$$

Since $F_r$ is self-adjoint, $\|F_r^{n_0d}\| = \|F_r\|^{n_0d}$; thus the conclusion of the theorem.

**Corollary 2.** Under conditions (a), (b) and (c), the random scan Gibbs sampler converges geometrically in terms of Pearson $\chi^2$-distance, and $\text{corr}(t(X_0), t(X_n))$ decreases geometrically for any square integrable function $t$.

5. EXAMPLES

5.1. Gaussian Distribution

Suppose that $X$ has a non-degenerate multivariate normal distribution with mean
0 and covariance matrix $\Sigma_{d \times d}$. By using the inverse covariance matrix $Q = \Sigma^{-1}$, the density can be written as

$$
\pi(X) = \frac{\sqrt{\text{det}(Q)}}{\{\sqrt{(2\pi)}\}^d} \exp\left(-\frac{1}{2} X^T Q X \right).
$$

All the conditional distributions $\pi\{x(i)|X^{(i-1)}\}$ are supposed to be known to us. In image processing, these conditional distributions may depend only on the neighborhood structure of the model. The simplest such case is the 'nearest neighbour structure' in which the conditional distribution of the index $i$ is determined by the four nearest neighbours in a rectangular lattice structure.

Now we consider the ordinary systematic scan, assuming that $X$ and $Y$ are consecutive draws. Then under stationarity $(Y, X) = (y(1), \ldots, y(d), x(1), \ldots, x(d))$ is normally distributed with covariance matrix

$$
\Psi = \begin{pmatrix}
\Sigma & \Sigma_{yx} \\
\Sigma_{yx} & \Sigma
\end{pmatrix}.
$$

We show that $\Psi$ is non-singular. Suppose otherwise; then there exist non-zero vectors $a = (a_1, \ldots, a_d)$ and $b = (b_1, \ldots, b_d)$ such that $\Sigma a_i y(i) + \Sigma b_i x(i) = 0$. Since $Y$ and $X$ are non-degenerate, there exists a non-zero $a_k$ with smallest index such that

$$
-a_k y(k) = \sum_{i > k} a_i y(i) + \sum_{i < k} b_i x(i) + \sum_{i \geq k} b_i x(i).
$$

However, since given $y(k+1), \ldots, y(d), x(1), \ldots, x(k-1)$, we have that $y(k)$ is conditionally independent of $x(k), x(k+1), \ldots, x(d)$; thus

$$
E\{y(k)|y(k+1), \ldots, y(d), x(1), \ldots, x(d)\} = E\{y(k)|y(k+1), \ldots, y(d), x(1), \ldots, x(k-1)\}
$$

is a linear function of $y(k+1), \ldots, y(d), x(1), \ldots, x(k-1)$. Combining this with equation (14), we know that all $b_i$, $i \geq k$, must be 0. However, $(y(k), \ldots, y(d), x(1), \ldots, x(k-1))$ is normally distributed with a non-singular covariance matrix $\Sigma$. This leads to a contradiction.

With the non-singularity of $\Psi$, standard canonical analysis gives us the maximal correlation $\gamma$ between $X$ and $Y$, which is the maximal eigenvalue of

$$
\Sigma^{-1} \Sigma_{yx} \Sigma^{-1} \Sigma_{yx}.
$$

By remark 3 in Section 3, this value is equal to the norm of the forward operator. Eaton (1976) provided an explicit bound for $\gamma$ in terms of the maximal and minimal eigenvalues of $\Psi$:

$$
\gamma^2 \leq \frac{\lambda_1 - \lambda_{2d}}{\lambda_1 + \lambda_{2d}},
$$

where $\lambda_1$ and $\lambda_{2d}$ are respectively the largest and the smallest eigenvalues of the matrix $\Psi$.

5.2. Murray's Data

The data in Table 1 are supposed to be drawn from a bivariate normal distribution with mean $\mu_1 = \mu_2 = 0$. This artificial data set was originally produced by Murray.
TABLE 1

Observations from a bivariate normal distribution

<table>
<thead>
<tr>
<th>x_1</th>
<th>x_2</th>
<th>x_3</th>
<th>x_4</th>
<th>y_1</th>
<th>y_2</th>
<th>y_3</th>
<th>y_4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>2</td>
<td>2</td>
<td>-2</td>
<td>-2</td>
</tr>
<tr>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>y_1</td>
<td>y_2</td>
<td>-y_3</td>
<td>-y_4</td>
</tr>
</tbody>
</table>

†x_i and y_i represent the missing parts.

(1977), and later used by Tanner and Wong (1987) to illustrate their data augmentation method. Jeffreys's non-informative prior given to the covariance matrix is (Box and Tiao, 1973) \( \pi(\Sigma) \propto |\Sigma|^{-(m+1)/2} \), where \( m = 2 \) is the dimension of the distribution. Our purpose here is not to investigate the posterior distribution of the correlation \( \rho \). Instead, we are interested in whether conditions (b) and (c) can be verified.

Let the missing values be denoted by \( x_i \) and \( y_i \) as in Table 1. Let \( Z = (x_1, \ldots, x_4, y_1, \ldots, y_4) \) and

\[
\Sigma = \begin{pmatrix}
\sigma_1^2 & \rho \sigma_1 \sigma_2 \\
\rho \sigma_1 \sigma_2 & \sigma_2^2
\end{pmatrix}.
\]

A two-component Gibbs sampler (or equivalently the data augmentation) can be applied to draw \( \Sigma \) and \( Z \) iteratively. The two conditional distributions are

\[
\pi(Z|\Sigma) = \frac{1}{\{2\pi\sigma_1 \sigma_2 (1-\rho^2)\}^{4}} \exp \left[ -\frac{1}{2(1-\rho^2)} \left( \frac{\sum x_i^2}{\sigma_1^2} + \frac{\sum y_i^2}{\sigma_2^2} - \frac{4\rho \sum (x_i + y_i)}{\sigma_1 \sigma_2} \right) \right]
\]

\[
+ \frac{16\rho^2}{\sigma_2^2} + \frac{16\rho^2}{\sigma_2^2} \right],
\]

\[
\pi(\Sigma|Z) = \frac{|\Sigma|^{-15/2} |B|^{16}}{(2\pi)^{12}} \exp \left[ -\frac{1}{2(1-\rho^2)} \left( \frac{\sum x_i^2}{\sigma_1^2} + \frac{\sum y_i^2}{\sigma_2^2} - \frac{4\rho \sum (x_i + y_i)}{\sigma_1 \sigma_2} + \frac{20}{\sigma_2^2} + \frac{20}{\sigma_2^2} \right) \right],
\]

where

\[
B = \begin{pmatrix}
20 + \sum x_i^2 & 2 \sum (x_i + y_i) \\
2 \sum (x_i + y_i) & 10 + \sum y_i^2
\end{pmatrix}.
\]

Condition (c) is satisfied with this example because the densities are all positive. The main difficulty is condition (b). By comment (ii) of Section 3, we can rewrite condition (b) as

\[
\int \pi(Z|\Sigma) \pi(\Sigma|Z) \, dZ \, d\Sigma < \infty.
\]

When we multiply the two conditional distributions, it is seen that the exponential part can be rearranged to be the sum of complete squares:

\[
\frac{1}{2(1-\rho^2)} \left\{ 2 \sum x_i \left( \frac{x_i}{\sigma_1} - \frac{2\rho}{\sigma_2} \right)^2 + 2 \sum y_i \left( \frac{y_i}{\sigma_2} - \frac{2\rho}{\sigma_1} \right)^2 + \frac{20 - 16\rho^2}{\sigma_2^2} + \frac{20 - 16\rho^2}{\sigma_2^2} \right\},
\]

To check condition (b), we observe that \(|B|^{16} \) is a polynomial of \( x_i \) and \( y_i \) with a maximal degree of 12 in both \( x_i \) and \( y_i \). Hence normal moments of order at most 12 are obtained when the \( x_i \) and \( y_i \) are integrated out. The result is then a polynomial of
$1/\sigma_1, 1/\sigma_2$ which is also a rational function of $\rho$. Together with all the other coefficients, we obtain the integral

$$
\int f\left\{ \frac{1}{\sigma_1}, \frac{1}{\sigma_2}, \rho, \sqrt{1-\rho^2} \right\} \exp\left\{ -\frac{10-8\rho^2}{1-\rho^2} \left( \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \right) \right\} d\Sigma,
$$

where $f$ is a polynomial of its first three arguments and a rational function of the fourth. The integral is finite because of the exponential term. Thus we have verified condition (b). The convergence rate of data augmentation is therefore geometric in this case.

ACKNOWLEDGEMENTS

The authors thank Professor Charles Geyer for bringing up the autocovariance problem of the random scan, the Editor for many helpful comments and the Associate Editor and three referees for many invaluable and detailed suggestions which made the current version possible. Special thanks go to the Associate Editor and a referee who did two rounds of line-by-line checking and correction of the paper.

This manuscript was prepared using computer facilities supported in part by National Science Foundation grants DMS 84-04941, DMS 86-01732, DMS 87-03942 and DMS 89-05292 awarded to the Department of Statistics at the University of Chicago, and by the University of Chicago Block Fund.

REFERENCES


