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

Fang jiaheng, Yan yuqing, Yu jianing, Zhang yifan

May 26, 2025

1 Introduction

We will show 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 χ^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.

First, let us review the classical Gibbs sampling method.

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

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

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

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

where $X^{[-A]} = \{x(i) : i \in A^c\}$ denotes components of X excluding those components indicated in the set A, and $\pi(\ |\ )$ denotes all the conditional distributions generated by the target distribution π. 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 updatings, 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 π 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.

2 Definition and Basics

Definition 2.1:

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 2.2:

Two operators F and B where F stands for 'forward' and B stands for 'backward' are defined as

$$Ft(X_1) \overset{\text{def}}{=} E\{t(X_2)|X_1\} = \int t(Y)K(Y|X_1)\,dY,$$

$$Bt(X_2) \overset{\text{def}}{=} E\{t(X_1)|X_2\} = \int t(X)\frac{K(X_2|X)}{\pi(X_2)}\pi(X)\,dX.$$

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

$$F^n t(X_0) = E\{t(X_n)|X_0\} \quad \text{and} \quad B^n t(X_n) = E\{t(X_0)|X_n\}.$$

Therefore the following lemma holds.

Lemma 2.3:

For any t and s in $L_0^2(\pi)$, $\text{cov}\{t(X_n), s(X_0)\} = \langle F^n t, s\rangle = \langle F^{n-k}t, B^k s\rangle$.

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 $F = B$, and all even-lag autocovariances are non-negative, i.e.

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

Definition 2.4: The Pearson χ^2-distance from density $p(X)$ to $\pi(X)$, $d_\pi(p,\pi)$, is

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

$d_\pi(\cdot,\cdot)$ is not a true distance. It is rather a measure of discrepancy between any distribution p and the target distribution π, and it is a stronger measure than both the L^1-distance and a kind of Kullback-Leibler information distance.

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 χ^2-distance between p and π is small, so will be their total variation distance. In this sense, the convergence of the Pearson χ^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 χ^2-distance and the reversed Kullback-Leibler information distance:

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

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

3 Systematic Scan Gibbs Sampler

In this section, we will prove an important convergence theorem.

Theorem 3.1:

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 χ^2-distance from p_n to π 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.

The following conditions are needed:

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

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

(b) For the transition function K defined in equation (1),

$$\int \left\{ \frac{K(Y|X)}{\pi(Y)} \right\}^2 \pi(X)\pi(Y)\, dX\, dY < \infty.$$

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 is irreducible.

The following lemmas are useful:

Lemma 3.2:

Condition (b) implies that the forward operator F_s is compact.

Lemma 3.3:

Conditions (b) and (c) 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.

4 Random Scan Gibbs Sampler

In this section, we will obtain the covariance structure of the random scan Gibbs sampler as well as its convergence property.

First, let's write out the transition function of the random scan Gibbs sampler:

$$K(X^{(t)}|X^{(t-1)}) = \sum_{i=1}^{d} \alpha_i K_i(X^{(t)}|X^{(t-1)}) \tag{2}$$

here,

$$K_i(X^{(t)}|X^{(t-1)}) = \pi(X_i^{(t)}|X_{[-i]}^{(t-1)}) I\{X_{[-i]}^{(t)} = X_{[-i]}^{(t-1)}\}$$

The following propositions are easy to obtain:

Prop 4.1:

π is the invariant distribution of the Markov chain defined by the transition kernel (2).

Prop 4.2:

The Markov chain defined by the transition kernel (2) is reversible.

The following lemma is useful:

Lemma 4.3:

The lag-1 autocovariances of the Markov chain generated by the Random scan Gibbs sampler is non-negative. More specifically,

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

Here, $\mathbf{i}$ is the random variable representing the index to be updated.

Now we could obtain the covariance structure of the random scan Gibbs sampler:

Theorem 4.5:

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 $t(X) \in L_0^2(\pi)$, the autocovariance between $t(X_0)$ and $t(X_n)$ is a non-negative monotone decreasing function of n. It can be written as

$$\text{cov}\{t(X_0), t(X_n)\} = \text{var}\big\{E(\ldots E[E\{t(X)|\mathbf{i}, X^{[-i]}\}|X]\ldots)\big\},$$

where there are n conditional expectations taken alternately on $\{\mathbf{i}, X^{[-i]}\}$ and X.

Corollary 4.6

If $t(X) \in L_0^2(\pi)$, then $r_n(t) \leq \|F_r\|^n$, where $\|F_r\|$ is the norm of the forward operator, and $r_n(t) = \text{corr}\{t(X_0),\, t(X_n)\}$.

Now, we discuss the convergence property of the random scan Gibbs sampler.

Theorem 4.7

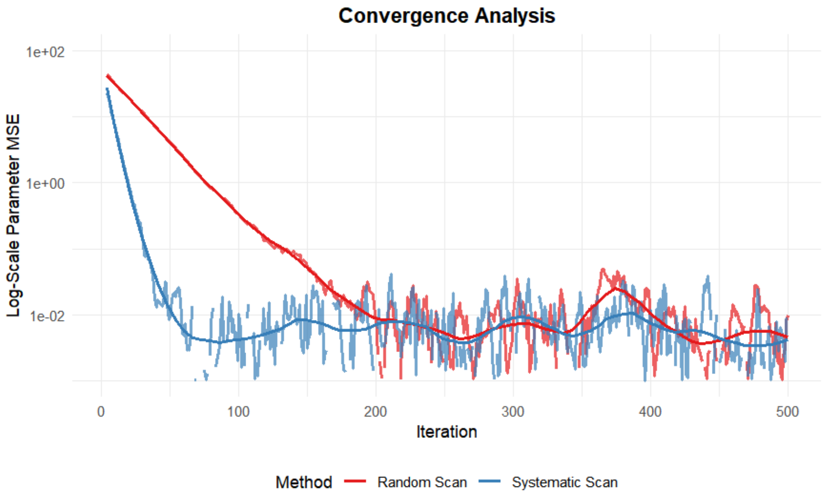
Under the conditions (b) and (c), we have $\|\mathbf{F_r}\| < 1$.

Corollary 4.8

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

5 R Implementation

Check the convergence rate using R code:


Convergence Analysis
1e+02
1e+00
1e-02
Log-Scale Parameter MSE
Iteration
Method
Random Scan
Systematic Scan