THE UNIVERSITY OF CHICAGO

CORRELATION STRUCTURE AND CONVERGENCE RATE OF THE GIBBS SAMPLER

A DISSERTATION SUBMITTED TO

THE FACULTY OF DIVISION OF THE PHYSICAL SCIENCES

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY

JUN LIU

CHICAGO ILLINOIS

JUNE 1991

Science Foundation Grants 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.

CONTENTS

ACKNOWLEDGEMENT ii

ABSTRACT vi

Chapter

1. INTRODUCTION AND PRELIMINARIES 1

1.1 Introduction 1

1.2 Preliminary lemmas 8

2. DATA AUGMENTATION 16

2.1 Correlation structure 17

2.2 Upper and lower bounds 21

2.3 Convergence rate problem 26

3. COMPARISONS OF ESTIMATORS AND SCHEMES 29

3.1 Mixture representation and histogram 29

3.2 Comparison of schemes corresponding to partitioning 34

4. RESULTS FOR DIFFERENT SCANS 44

4.1 Systematic scan Gibbs sampler 44

4.2 Random scan 54

4.2.1 *Correlation structure* 54
4.2.2 *Geometric convergence* 57
4.3 Other scans 59
4.3.1 *Symmetric random permutation scan (SRPS)* 59
4.3.2 *Symmetric systematic scan (SSS)* 63
5. FURTHER ANALYSIS AND EXAMPLES 66
5.1 Spectral analysis on the forward operators 66
5.2 Examples 71
6. A REVIEW OF STOCHASTIC RELAXATION TECHNIQUES 78
6.1 The Metropolis Algorithm 78
6.2 The Gibbs Sampler 86
6.3 Applications in statistics 93
BIBLIOGRAPHY 98

ABSTRACT

This thesis begins with a detailed study of the correlation structure and convergence rate of a Markov chain generated using the Gibbs sampler, a popular technique for Monte Carlo simulations from a complicated multidimensional distribution, by focusing on the special case of data augmentation which, proposed by Tanner and Wong [54], is specially used in Bayesian calculation to deal with missing data problems. It is shown in such case that the autocorrelations are non-negative and monotone decreasing as a function of lag. When applied to Bayesian missing data problem, the Gibbs sampler produces two natural estimators for the posterior distribution of the parameter vector: one is the histogram of the sampled values of the parameter vector, the other is a mixture of complete data posteriors. It is demonstrated that the mixture representation is preferable to the histogram approximation in the sense that it has a smaller variance. Some results on the geometric rate of convergence are established in such case. Several other interesting theoretical problems related to the comparisons of different estimators and different sampling schemes are addressed. It is shown that grouping variables in the process of iteration for implementing the Gibbs sampler is usually a good strategy. These results provide practical guidance for the use of the Gibbs sampler in applications.

In the latter part, the convergence rate results for the general Gibbs sampler used with various scans are derived. It is shown that under conditions which guarantee that the Markov forward operator is compact, the Gibbs sampler used with either the systematic scan or the random scan converges with a geometric rate. Here the term 'scan' refers to the order the variables are visited and updated. In particular, for the random scan, the

autocorrelations of the samples can be expressed as the variances of some iterative conditional expectations. As a consequence, the autocorrelations are all positive and decrease monotonically as a function of lag.

A review of the stochastic relaxation techniques is presented at the end of the thesis.

CHAPTER 1

INTRODUCTION AND PRELIMINARIES

1.1 Introduction

The Gibbs sampler is an iterative scheme for approximate generations of samples from a multivariate distribution. It is related to the Metropolis algorithm (Metropolis et al. [39]) in statistical physics, and was introduced by Geman and Geman [21] in the context of statistical image restoration, where they iteratively sampled each pixel value conditional on the values of neighboring pixels. The basic scheme of the Gibbs sampler can be described as follows:

Suppose $X = (x(1), \ldots, x(d))$ is a d-dimensional random variable with density function $\pi(X)$ which is difficult to compute directly. However, the d conditional distributions $\pi(x(i)|X^{-i})$, where X^{-i} denotes $\{x(j)\}_{j \neq i}$, are assumed to be simple and easy to draw from. The Gibbs sampler is a stochastic relaxation technique which allows us to obtain samples from the joint density $\pi(X)$ by running a Markov chain which has $\pi(X)$ as its equilibrium

distribution. In the later context, $\pi(A)$ may also be used to denote the equilibrium probability measure of a set A. The chain is initiated by a draw from some starting density $p_0(X)$. According to a visiting scheme, each variate $x(i)$ is visited and updated by a sample drawn from the conditional distribution $\pi(x(i)|X^{-i})$, where X^{-i} denotes the current state of the other $d-1$ variables. The visiting sequence can be either deterministic (systematic scan) or random (random scan). As long as each variable is visited infinitely often, under some mild conditions, the joint distribution of the $x's$ will converge to $\pi(X)$ as the number of visits increases.

The systematic use of such iterative sampling schemes in parametric statistical problems began with Tanner and Wong [54] where they introduced a similar method, called data augmentation, for approximate computations of posterior densities in parametric models to which the EM algorithm (Dempster, Laird, Rubin [10]) for maximum likelihood calculation is applicable, i.e., models that can be fruitfully formulated as a complete/incomplete data problem. Because of the large number of important statistical models which can be so formulated, the possibility is opened up for the application of data augmentation/Gibbs sampling type iterative schemes to these models, such as latent class models, variance component and hierarchical linear models, missing data in multivariate normal models, censored and truncated data problems, to name just a few (Tanner and Wong [54], Gelfand and Smith [19]). Simultaneously and independently, Li [34] has applied similar schemes to impute multivariate missing data. The structure and formal connection between data augmentation and Gibbs sampling algorithms was clarified by Gelfand and Smith [19], where further interesting theoretical questions, such as whether the mixture representation is superior to the histogram representation, are raised. In this paper, we use the term "data augmentation"

to refer to the two variable Gibbs sampler. There is by now a long list of papers dealing with the application of the Gibbs sampler to various problems.

We observe here that the setting for the application of the Gibbs sampler in Bayesian parametric computations ([54], [19]) is typically different from that in more traditional applications in statistical physics and image analysis (Metropolis et al [39], Geman and Geman [21]). For example, in image analysis, there is usually a large number of variables (pixel grey levels, edge elements, etc.), all of which is simple and the conditional distribution of each variable, say grey level of a pixel, given the states of its neighbors, typically has the same structure irrespective of the position of the pixel. On the other hand, in most statistical applications, we typically iterate among a few variables, each of which may be a vector with many components and the conditional distributions may have drastically different structure. Furthermore, the statistician often has some freedom in choosing the set of variables to iterate (i.e., choosing an augmentation scheme). This raises some interesting new questions which will be discussed in chapter 3.

Despite its popularity, some fundamental questions concerning the Gibbs sampler have not been satisfactorily resolved. One of the interesting questions, which is also closely tied to the convergence rate, and is perhaps more relevant in most applications, has to do with the correlation structure of the samples generated. Consider the simple case where $d = 2$ and $X = (x, y)$. Suppose a systematic scan is used so that the visiting sequence alternates between x and y. We label the successive draws of x and y by x_k and y_k, $k = 0, 1, \ldots, n$. For simplicity, suppose (x_0, y_0) are drawn from $\pi(X)$. In that case , the distribution of $X_k = (x_k, y_k)$ is $\pi(X)$ for all k. Suppose we are interested in $E_\pi[t(X)]$ for some square integrable function $t(X)$ (integrable with respect to the equilibrium measure π), a natural

and unbiased estimate of $E_\pi[t(X)]$ is

$$\hat{t}_1 = \frac{1}{n}\sum_{k=1}^{n} t(X_k).$$

The variance of this estimate depends entirely on the correlations among the $t(X_k)$'s. In particular, the efficiency of the estimate depends on how fast the autocorrelation goes to zero as the lag goes to infinity. For this reason alone, the study of the correlations among successive samples in the Gibbs sampler is of extreme importance.

The understanding of the correlation structure also allows us to answer some other interesting questions concerning the comparisons of different estimators. In the above example where $X = (x, y)$, it is often the case that $t(X)$, the function of interest, is a function of a single component of X, i.e, without loss of generality, $t(X) = t(x)$ (This is the case in data augmentation setting in Tanner and Wong [54]). In this case, there are two natural estimates of $E_\pi[t(x)]$:

$$\hat{t}_1 = \frac{1}{n}\sum_{k=1}^{n} t(x_k), \tag{1.1}$$

$$\hat{t}_2 = \frac{1}{n}\sum_{k=1}^{n} E(t(x)|y_k). \tag{1.2}$$

We call (1.2) the histogram approximation since it is based entirely on the sampled x values. The estimate (1.2), which is usually easy to compute assuming that $\pi(x|y)$ is simple, is called the mixture approximation. These two names stem from the Bayesian missing data problem setting, where x represents the parameter vector, $\hat{t}_1$ is an estimate based on dependent samples drawn from the true posterior density, and $\hat{t}_2$ is a mixture of complete data posterior means. Starting with Tanner and Wong [54], it has long been conjectured that the mixture approximation is always better, i.e, having a smaller variance, than the histogram approximation. As demonstrated by Gelfand and Smith [19], the proof

is trivial if (x_k, y_k) and (x_l, y_l) are independent for all $k \neq l$, which is equivalent to the fact that

$$\text{var}(E(t(x)|y)) \leq \text{var}(t(x)).$$

However the theoretical justification is not so straightforward when $\{(x_k, y_k), k = 1, \ldots, n\}$ are dependently drawn from the equilibrium using the Gibbs sampler. A related question is whether it is better to use a weighted combination of the two estimates since both contain information about $E_\pi(t(x))$. In this paper, we will demonstrate that for several scans, the mixture approximate is always superior and nothing can be gained by using a weighted sum of both.

The results concerning the correlation structure also provide insights into another important question, namely, the comparisons between various augmentation schemes. For example, suppose we are given the following three schemes:

$$\begin{array}{ll} \text{[i]} & x|y, \quad y|x, \\ \text{[ii]} & x|\{y,z\}, \quad \{y,z\}|x, \\ \text{[iii]} & x|\{y,z\}, \ \ y|\{x,z\}\ , \ z|\{x,y\}. \end{array}$$

Here [i] indicates the ordinary data augmentation applied to x and y by sampling x conditioned on y, y conditioned on x, and iterating between the two steps; [ii] suggests doing data augmentation on x and $\{y, z\}$ by grouping y and z together; [iii] is just the ordinary three dimensional systematic scan Gibbs sampler on $\{x, y, z\}$, in which we draw x conditioned on y, z; draw y conditioned on x, z; draw z conditioned on x, y; and then iterate. Compared with [i], an auxiliary random variable z is introduced in both schemes [ii] and [iii]. This may be done because it is easier to draw from $\pi(x|y, z)$ than $\pi(x|y)$. However, since more

variables are imputed, we expect this will lead to a slower convergence rate and higher autocorrelations. For some scans, this is shown to be indeed the case. In practice, there needs to be a compromise between the rate of convergence and the ease of imputations. Some of the results in this paper will help users in determining the optimal compromise.

In this paper, a systematic study on the correlation structure of the Gibbs sampler with various scans is conducted. It is shown that for a certain class of scans satisfying 'reversibility' and the 'interleaving Markov' property, the autocorrelations can be expressed as the variance of some iterative conditional expectations. This class includes data augmentation, the random scan and the symmetric systematic scan. With this expression we can easily prove the monotonicity and positivity of correlations between $t(X_k)$ and $t(X_l)$ for any square integrable function $t(\cdot)$ of the chain. Some results about comparing schemes are presented in which we transform the comparison of schemes into a comparison of the norms of certain operators.

Through a simple inequality, we connect the correlation problem with the convergence rate problem, and naturally derive a series of relatively thorough results on geometric convergence rate of the general Gibbs sampler with various scans under three basic conditions. The conditions will be discussed in detail from the practitioner's point of view. In Geman and Geman [21], geometric convergence rate is obtained for finite discrete state space under certain positivity condition on the equilibrium distribution. Their proof depends crucially on the assumptions of discrete finite state space and positivity, and cannot be easily generalized. Tanner and Wong [54] studied the data augmentation case where only two random vectors are iteratively sampled. They proved the convergence of the scheme and also claimed the geometric rate for their method under mild conditions. However the proof

for the geometric rate contained an error. Schervish and Carlin [50], by extending the line of arguments in Tanner and Wong [54], provided a nice proof of geometric convergence rate for the Gibbs sampler with systematic scan so that the loose point in Tanner and Wong [54] was fixed. Compared with Schervish and Carlin [50], our conditions are weaker and the results are more general. Particularly, the geometric convergence rate for the random scan Gibbs sampler with general state space, which has not been addressed in the above papers, is also obtained. The relations among the convergence rate, the maximal correlation, the ρ-mixing condition and the spectral radii of certain operators are explored.

By using spectral analysis on the forward operator of the Markov chain, we elaborate an idea of how to estimate the convergence rate for certain cases and describe a possible rule for stopping the iteration. From this point of view, we can see clearly how the choice of starting density can make a difference. This last issue was also discussed by Schervish and Carlin [50]. Examples are also presented to show the limitations of each special property we derived and that some properties are not universally true for all the scans. The applications of our theory to Gaussian distributions and the posterior calculation of covariance matrix are demonstrated too.

Before discussing the correlation and convergence problems in detail, a few useful preliminary lemmas are presented in the latter part of this chapter. Chapter two focuses on a detailed analysis of data augmentation which contains all the basic ideas and methods for dealing with the general Gibbs sampler. In chapter three, results on the comparisons of different estimators and of different schemes are obtained. Chapter four is concerned with the results on the general Gibbs sampler with different kind of scans, including the ordinary systematic scan (OSS), the random scan (RS), the random permutation scan (RPS)

and the symmetric systematic scan (SSS). Chapter five contains some refined results obtained from the spectral analysis of the self-adjoint compact operators, and two examples supplied to illustrate our theory. In the last chapter, a relatively thorough review of the stochastic relaxation techniques including the Metropolis algorithm, the Gibbs sampler and data augmentation are given. Some theoretical aspects of these methods, as well as their applications, are discussed.

1.2 Preliminary lemmas

To be more general in our discussion, we begin with some results on general Markov chains. We furnish all the proofs needed because they are simple and the notations are useful later in the thesis, though one may find similar results in different settings, e.g., Rosenblatt [47]. Suppose $X_0, X_1, X_2, \ldots, X_n$ are consecutive samples from a stationary Markov chain with the equilibrium distribution denoted by $\pi(X)$, and the transition function $K(X_1|X_0)$ governing the probability of transition from X_0 to X_1.

Definition 1 *The n-step transition function is denoted by*

$$K^n(y|x) = P(X_n = y|X_0 = x) = K * K * \cdots * K(y|x)$$

in which there are n convolutions. We will use $K^n(Y|X)$ as well.

Definition 2 *Two operators F and B, where F stands for "forward" while B stands for "backward", are defined as:*

$$F\,t(X_1) \stackrel{\text{def}}{=} \int t(Y)K(Y|X_1)dY = E(t(X_2)|X_1), \tag{1.3}$$

$$B\,t(X_2) \stackrel{\text{def}}{=} \int t(X)\frac{K(X_2|X)\pi(X)}{\pi(X_2)}dX = E(t(X_1)|X_2). \tag{1.4}$$

The Hilbert space of the centered square integrable functions of X is denoted by:

$$L_0^2(X) = \{t(X) : \int t^2(X)\pi(X)dX < \infty \quad and \quad \int t(X)\pi(X)dX = 0\}$$

with the inner product

$$\langle t(X), s(X)\rangle = E_\pi(t(X) \cdot s(X)).$$

The operators F and B map $L_0^2(X)$ to itself. From a property of Markov chain, it is seen that the two operators F and B are adjoint to each other, and

$$\begin{aligned} F^n t(X_0) &= E(t(X_n)|X_0), \\ B^n t(X_n) &= E(t(X_0)|X_n). \end{aligned}$$

Using an elementary probability inequality, it can also be proved that the norms of the two operators are all bounded above by one. With the above notations, we state and prove a fundamental lemma.

Lemma 1.2.1 *For any functions $s(\cdot)$ and $t(\cdot)$ of the random variable X, which are square integrable with respect to the stationary measure π, the covariance of $t(X_n)$ and $s(X_0)$ satisfies*

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

for any $0 \le k \le n$.

PROOF: Without loss of generality, we may assume $E_\pi t(X) = E_\pi s(X) = 0$. Since

$$\begin{aligned} E(t(X_n)s(X_0)) &= E(E(t(X_n)s(X_0)|X_{n-1})) \\ &= E(E(t(X_n)|X_{n-1}) \cdot E(s(X_0)|X_{n-1})) \\ &= E(F\, t(X_{n-1}) \cdot B^{n-1} s(X_{n-1})), \end{aligned}$$

an induction argument leads to the conclusion. □

The chain is said to be *reversible* if for any two measurable sets H and K:

$$P(X_1 \in H \ , \ X_2 \in K) = P(X_1 \in K \ , \ X_2 \in H).$$

This condition is equivalent to the *detailed balance* condition introduced by Metropolis et al [39]:

$$K(Y|X)\pi(X) = K(X|Y)\pi(Y).$$

The following lemma proves this equivalency.

Lemma 1.2.2 *The reversibility condition is equivalent to the detailed balance condition, and is also equivalent to the condition that $F = B$.*

PROOF: The equivalence between the detailed balance condition and the condition that $F = B$ is straightforward from the definition of the two operators:

$$\begin{aligned} \langle Ft(X_0), s(X_0)\rangle &= \int\int t(X_1)K(X_1|X_0)dX_1 s(X_0)\pi(X_0)dX_0 \\ &= \int t(X_1)\pi(X_1)\int s(X_0)K(X_0|X_1)dX_0 dX_1 \\ &= \langle t(X_1), Fs(X_1)\rangle = \langle t(X_0), Bs(X_0)\rangle. \end{aligned}$$

Hence the operators F and B are equal and self-adjoint. But the reversibility condition can be rephrased as

$$\int_H F\ I_K(X_1)\pi(X_1)dX_1 = \int_H B\ I_K(X_2)\pi(X_2)dX_2,$$

where $I_K(\cdot)$ is the indicator function. Therefore all three conditions are equivalent. □

An intuitive exposition of the detailed balance condition is that the probability of the system being at X and then moving to Y is the same as the system being at Y and then moving to X. With this property, an interesting corollary follows.

Corollary 1.2.1 *If the chain is reversible, then for any* $t \in L_0^2(X)$

$$\begin{aligned} cov(t(X_0), t(X_{2m})) &= E[(F^m t(X))^2] = E[(B^m t(X))^2] \\ &= E(E^2(\cdots E(E(t(X_0)|X_1)|X_2)|\cdots|X_m)) \end{aligned}$$

Hence it is a nonnegative monotone decreasing function of m. *Furthermore,*

$$|cov(t(X_0), t(X_{2m+1}))| \leq cov(t(X_0), t(X_{2m})).$$

PROOF: Since the chain is reversible, the two operators F and B are the same. Therefore the first conclusion follows from lemma 1.2.1 with $k = m$. The monotonicity of even-lag autocorrelations follows from the inequality that

$$\text{var}(E(x|y)) \leq \text{var}(x).$$

The inequality follows from Hölder's:

$$\begin{aligned} |\text{cov}((t(X_0), t(X_{2m+1}))| &= |\text{cov}(F^m t(X), F^{m+1} t(X))| \\ &\leq \sqrt{\text{var}(F^m t(X)) \cdot \text{var}(F^{m+1} t(X))} \\ &\leq \text{var}(F^m t(X)) = \text{cov}(t(X_0), t(X_{2m})). \end{aligned}$$

□

The reversible chain has been used as early as 1953 by Metropolis et al [39] which we now refer to as the *Metropolis algorithm*. Later it will be demonstrated that many scans being used with the Gibbs sampler actually lead to reversible Markov chains. The finer structure of special cases will be discussed later. We now turn to some lemmas on the convergence rate.

Definition 3 *The chain with starting distribution $P_0(dX)$ is said to be functional geometric convergent in L^2 if there exists an $\alpha < 1$ such that for any $t \in L_0^2(X)$,*

$$|\int\int t(Y)K^n(dY|X)P_0(dX) - E_\pi(t(X))| \leq c_0\alpha^n \parallel t \parallel .$$

If we use $E_n(t(X))$ to denote the expectation taken under the measure $P_n(dX)$ of the nth-step evolution from $P_0(dX)$, the above definition has a simple implication that for any $t \in L_0^2(X)$

$$|E_n(t(X)) - E_\pi(t(X))| \leq c_0\alpha^n \parallel t \parallel .$$

Taking $t = I_A - \pi(A)$, the indicator function of a measurable set A, we can get the geometric convergence of the total variation from the above definition, i.e.,

$$|P_n(A) - \pi(A)| \leq \frac{1}{2}c_0\alpha^n.$$

Hence the Hellinger distance between P_n and π (and also L^1-distance of p_n and π) goes to zero geometrically fast. An easy argument shows that the functional geometric convergence also implies the L^2 geometric convergence of the ratio $p_n(X)/\pi(X)$ where $p_n(X)$ is the nth evolved density starting from $p_0(X)$, for the reason that $E_n(t(X)) - E_\pi(t(X))$ can also be written as $\int t(X)(p_n(X)/\pi(X) - 1)\pi(X)dX$ and can be viewed as a functional of $t(X)$. The uniform boundedness of this functional gives us the L^2 integrability of p_n/π with respect to equilibrium measure π. Hence we have an statement equivalent to functional geometric convergence:

$$\parallel \frac{p_n(X)}{\pi(X)} - 1 \parallel \leq c_0\alpha^n.$$

The following condition is needed for all the main convergence results of this paper. It says that the starting density must have finite Fisher information with respect to the stationary density.

Condition (A). *The starting density $P_0(dX)$ satisfies the condition*

$$c_0^2 = \int \frac{P_0^2(dX)}{\pi(dX)} - 1 < \infty.$$

Lemma 1.2.3 *The spectral radius of the operators F and B are equal. If it is less than one, given condition (A), the chain will be functional geometric convergent.*

PROOF: Though it is well known that the adjoint operators in a Hilbert space have the same norms and spectral radii, we may want to see it go through a probabilistic argument: for any $t, s \in L_0^2(X)$ with norm 1,

$$\sup_{t,s} E(t(X_n), s(X_0)) = \sup_t \sqrt{E(E^2(t(X_n)|X_0))} = \sup_s \sqrt{E(E^2(s(X_0)|X_n))} \tag{1.5}$$

directly leads to the equality of the two norms.

Assuming that the spectral radius of F is $r < 1$, the following is standard:

$$r = \lim_{n\to\infty} \| F^n \|^{\frac{1}{n}} = \lim_{n\to\infty} \| B^n \|^{\frac{1}{n}} . \tag{1.6}$$

Hence the two operators have the same spectral radii. Furthermore, there exists a n_0 such that $\| F^{n_0} \| < 1$. By using lemma 1.2.1 and writing $g(X) = P_0(dX)/\pi(dX)$, we can convert the convergence rate problem into a covariance problem:

$$\begin{aligned} |E_n(t(X)) - E_\pi(t(X))| &= |\int t(Y) K^n(Y|X)(p_0(X) - \pi(X)) dY dX \\ &= |\mathrm{cov}_\pi(t(X_n), g(X_0))| \\ &= |\mathrm{cov}_\pi(F^n\, t(X), g(X))| \\ &\leq c_0 \| F^n \| \cdot \| t \| . \end{aligned}$$

Combining this with (1.6) gives the needed result. □

With lemma 1.2.3, we set up a duality between convergence rate and correlation structure. Our later analysis makes use of the special structure of the Gibbs sampler to bound $\| F \|$. The material here is also closely related to the ρ-mixing condition in the Markov chain literature, which says that

$$\rho_n = \sup_{t,s \in L^2(X)} \text{corr}(t(X_n), s(X_0))$$

converges to zero as n goes to infinity. We observe that this ρ_n is just the norm of F^n, and also that of B^n. Relevant results about ρ-mixing can be found in Bradley [7].

From the canonical correlation point of view, one may also think of ρ_n as the maximal correlation between X_0 and X_n. For any two random variables, the general notion is defined as

Definition 4 *If x and y are two random variables which may be multidimensional, $L^2(x) = \{f(x) : var(f(x)) < \infty\}$, $L^2(y) = \{g(y) : var(g(x)) < \infty\}$, the maximal correlation γ between x and y is defined as*

$$\gamma = \sup_{f \in L^2(x), g \in L^2(y)} corr(f(x), g(y)). \tag{1.7}$$

This concept has been developed since the 1930's. For (x, y) to be bivariate normal, several pioneer researchers Gebelein [18], Maung [37], Lancaster [33] proved, using several methods which involve applying Charlier's identity, that the maximal correlation is exactly the same as the absolute value of its ordinary correlation. Some works on general bivariate random variables were done by Lancaster [33], Csàki and Fischer [9], Renyi [44], Sarmanov [49]. More recently, Breiman and Friedman [8] proved that under certain conditions the maximal correlation between x and y is strictly less than one by using compact operator theory which

is relevant to our results. We write down the following lemma as a summary. It follows from (1.5) in the proof of lemma 1.2.3.

Lemma 1.2.4

$$\| F^n \| = \| B^n \| = \rho_n .$$

CHAPTER 2

DATA AUGMENTATION

The study of data augmentation is important. One reason is that it is the Bayesian analogue of the EM algorithm, and can be applied to many missing data problems. One of the variables, say x, usually corresponds to the parameter θ in the EM formulation so that only the marginal distribution of x, instead of the joint distribution of x and y, is of interest. Therefore the marginal chain consists of the x_k's is of special importance.

It is well-known that the data augmentation procedure is equivalent to the Gibbs sampler applied to two variables. The scan used in data augmentation is usually the systematic scan where x and y are sampled alternately. It is usually proposed as follows:

Let Ω be the sample space, x or x_k are random variables from Ω to some space (parameter space, say,) $\mathcal{X}$, y or y_k are random variables from Ω to $\mathcal{Y}$. Jointly

$$(x, y): \ \Omega \longrightarrow \mathcal{X} \times \mathcal{Y}.$$

The scheme consists of sampling x from the conditional distribution $\pi(x|y)$, sampling y from $\pi(y|x)$, and iterating between the two steps. By thinking of $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$ as

consecutive states of the chain, we can write down the transition function:

$$K((x_2, y_2)|(x_1, y_1)) = \pi(x_2|y_2)\pi(y_2|x_1).$$

A remarkable fact about it is that the marginal chains $\{x_k\}$ and $\{y_k\}$ are all reversible Markov chains although jointly $\{(x_k, y_k)\}$ is not reversible. This is not true for any random scans. The corresponding marginal forward operators are therefore self-adjoint ($F_x = B_x$). The corresponding marginal transition function can be written as

$$\begin{aligned} K_x(x_2|x_1) &= \int \pi(x_2|y_2)\pi(y_2|x_1)dy_2, \\ K_y(y_2|y_1) &= \int \pi(y_2|x_1)\pi(x_1|y_1)dx_1. \end{aligned}$$

We would like to point out that although data augmentation is mathematically a subcase of the general Gibbs sampler, the emphasis of the two are usually different. In the application of data augmentation to Bayesian computation, one of the variables, x or y, is used as the parameter of the model, the other as missing data so that they often have different and complex structures. In traditional applications of the Gibbs sampler, the variables are simpler and more homogeneous.

2.1 Correlation structure

Let $t(\cdot)$ be a measurable function on the space $\mathcal{X}$ such that $E(t^2(x)) < \infty$, $s(\cdot)$ a measurable function on the space $\mathcal{Y}$ and also square integrable.

Lemma 2.1.1 *The marginal chains $\{x_k\}$ and $\{y_k\}$ constructed by data augmentation are reversible Markov chains.*

PROOF: We can check the detailed balance condition directly:

$$\begin{aligned} K_x(x_2|x_1)\pi(x_1) &= \int \pi(x_2|y_2)\pi(y_2|x_1)dy_2\ \pi(x_1) \\ &= \int \pi(x_1|y_2)\pi(y_2|x_2)dy_2\ \pi(x_2) = K_x(x_1|x_2)\pi(x_2). \end{aligned}$$

The same is true for $\{y_k\}$. The Markov property is obvious. □

Lemma 2.1.2 *The one and two lag autocorrelations of $t(x)$ are nonnegative. The same is true for $s(y)$. The one lag correlations can be expressed as*

$$\begin{aligned} cov(t(x_0), t(x_1)) &= var(E(t(x)|y)), \\ cov(s(y_0), s(y_1)) &= var(E(s(y)|x)). \end{aligned}$$

PROOF: Without loss of generality, we assume t has mean zero. Therefore

$$\begin{aligned} \mathrm{cov}(t(x_0), t(x_1)) &= E(t(x_0)t(x_1)) = E(E(t(x_0)t(x_1)|y_1)) \\ &= E(E(t(x_0)|y_1) \cdot E(t(x_1)|y_1)) = E(E^2(t(x)|y)). \end{aligned}$$

The positivity of the two lag autocorrelation follows from the reversibility of the marginal chains and corollary 1.2.1. □

Though it is simple, the above lemma used two important properties of the marginal chains:

(i) Reversibility of the marginal chains $\{x_k, k = 1, 2, \ldots, \}$ and $\{y_k, k = 1, 2, \ldots, \}$.

(ii) Interleaving Markov property of the marginal chains defined as follows.

Definition 5 *A stationary Markov chain $\{x_k, k = 1, 2, \ldots, \}$ is said to have the interleaving Markov property if there exists a conjugate Markov chain $\{y_k, k = 1, 2, \ldots, \}$ such that*

(a) x_k and x_{k+1} are conditionally independent given y_k, $\forall k$.

(b) y_k and y_{k+1} are conditionally independent given x_{k+1}, $\forall k$.

(c) (y_{k-1}, x_k), (x_k, y_k) and (y_k, x_{k+1}) are identically distributed.

Lemma 2.1.3 *The marginal chains $\{x_k\}$ and $\{y_k\}$ constructed in data augmentation are reversible and have the interleaving Markov property.*

A fact that needs to be noted here is that we can actually derive the reversibility of the chain from the interleaving property. In other words, the interleaving property defined here implies reversibility. The reason we list them as two separate properties is to emphasize the concept of "interleaving".

Lemma 2.1.4 *Let x_0 and x_n be n lags apart from each other in the stationary marginal Markov chain described above, and $t(x) \in L_0^2(\mathcal{X})$. Then*

$$E(t(x_0)t(x_n)) = E[E(t(x)|y_n) \cdot E(t(x)|y_1)] \tag{2.1}$$

and the same is true for $E(s(y_0)s(y_n))$ with $s(\cdot)$ a square integrable centered function of y.

PROOF: A direct consequence of the interleaving Markov property. □

Theorem 2.1.1 *The n-lag covariances between $t(x_0)$ and $t(x_n)$, $s(y_0)$ and $s(y_n)$ are non-negative monotone decreasing with n, and have the following expressions :*

$$cov(t(x_0), t(x_n)) = var(E(E(\cdots E(t(x)|y)|x)|\cdots)), \tag{2.2}$$

$$cov(s(y_0), s(y_n)) = var(E(E(\cdots E(s(y)|x)|y)|\cdots)), \tag{2.3}$$

in which each has n expectation signs condition alternately on x and y.

PROOF: We use induction on 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 2.1.2. Assume the result is true for $n = m - 1$. For $n = m$, by applying lemma 2.1.4 and noting the Markov property, we get

$$\begin{aligned} E(t(x_0)t(x_m)) &= E(E(t(x)|y_1) \cdot E(t(x)|y_m)) \\ &= E(E(t(x)|y_0) \cdot E(t(x)|y_{m-1})). \end{aligned}$$

If we write $s(y) = E(t(x)|y)$ and use the induction assumptions for $n = m-1$, it is immediate that

$$\begin{aligned} E(t(x_0)t(x_m)) &= E(s(y_0)s(y_{m-1})) && (2.4)\\ &= E(E^2(E(\cdots E(s(y)|x)|y)|\cdots)) && (2.5)\\ &= E(E^2(E(\cdots E(E(t(x)|y)|x)|\cdots)). && (2.6) \end{aligned}$$

Expression (2.6) has $m-1$ expectation signs while expression (2.6) has m. Proceeding exactly the same as above, we can get the formula for $E(s(y_0)s(y_m))$. The monotone decreasing property of the covariances as lag increases follows easily from the inequality:

$$\text{var}(w(x)) \geq \text{var}(E(w(x)|y))$$

for any $w(\cdot)$. □

A more general way to obtain the positivity and monotonicity of the correlations between $t(x_0)$ and $t(x_n)$ is from operator theory. An operator F is called nonnegative if for any t in the space where F is defined on,

$$\langle Ft, t\rangle \geq 0 \quad , \quad \forall t.$$

Theorem 2.1.2 *Suppose F is the forward operator of a general Markov chain $\{X_i\}$. A necessary and sufficient condition for $cov(t(X_0), t(X_n))$ to be nonnegative and monotone decreasing with n for all $t(X) \in L_0^2(X)$ is that F is nonnegative (if and only if all the one lag autocorrelations are nonnegative), and self-adjoint (if and only if the chain is reversible).*

PROOF: Necessity is straightforward. For sufficiency, from the spectral theory of self-adjoint nonnegative operators (see, for example, Rudin [48]), there exists a self-adjoint operator A

such that $F = A^2$. Therefore, for any n

$$\mathrm{cov}(t(X_n)t(X_0)) = \langle F^n t, t\rangle = \langle A^n t, A^n t\rangle = \| A^n t \|^2$$

Since A is also a contracting operator, it follows that $\| A^n t \|$ is monotone decreasing. □

In the special case of data augmentation, the interleaving Markov property of the marginal chains leads to nonnegative marginal forward operators. Whether the interleaving Markov property is a necessary condition for the forward operator to be nonnegative is not clear.

2.2 Upper and lower bounds

Using the simple forms we derived in theorem 2.1.1 for the correlation structure of samples generated through data augmentation, it is possible for us to bound the autocorrelations

$$r_n(t) = \mathrm{corr}(t(x_0), t(x_n))$$

by some geometric sequences. From the geometric point of view, the iterative conditional expectation behaves like an iterative projection between two spaces. For a square integrable function $t(x)$, it is true that

$$r_1(t) = \sup_{s(y)} \mathrm{corr}(t(x), s(y)) = \sqrt{\frac{\mathrm{var}(E(t(x)|y)}{\mathrm{var}(t(x))}}.$$

Therefore this one-lag autocorrelation equals the length of the projection of $t(x)$ onto the space $L_0^2(y)$ of square integrable functions of y.

Definition 6 *A sequence of functions of x and y is defined recursively as*

$$G_0 = t(x), \;\; G_1 = E(t(x)|y), \;\; \ldots$$

$$G_{2k} = E(G_{2k-1}|x),\ G_{2k+1} = E(G_{2k}|y),\ \ldots\ .$$

We call this series of functions the conditional expectation sequence (CES).

Lemma 2.2.1 *The sequence of functions defined above have the relationship*

$$E(G_{2k-j} \cdot G_j) = E(G_k^2),$$

for any nonnegative integer k and $j < 2k$.

PROOF: It is similar to the proof of lemma 1.2.1 and theorem 2.1.1. From the definition, G_{2k} is a function of y while G_{2k+1} is a function of x for any k. Without loss of generality, we may assume $j < k$. When j is even, G_j and G_{2k-j} are functions of x, thus

$$\begin{aligned} E(G_{2k-j} \cdot G_j) &= E(E(G_{2k-j-1}|x) \cdot G_j) = E(E(G_{2k-j-1} \cdot G_j|x)) \\ &= E(G_{2k-j-1} \cdot G_j) = E(E(G_{2k-j-1} \cdot G_j)|y) \\ &= E(G_{2k-j-1} \cdot E(G_j|y)) = E(G_{2k-j-1} \cdot G_{j+1}) \end{aligned}$$

When j is odd, we can proceed exactly the same as above after interchanging the position of x and y. Therefore it is always true that

$$E(G_{2k-j} \cdot G_j) = E(G_{2k-j-1} \cdot G_{j+1})$$

for any integer j between 0 and k. The conclusion follows if we proceed the above operation until the two terms are equal. □

Lemma 2.2.2 *The autocorrelations are log-convex in lags: $r_n(t)^2 \le r_j \cdot r_{2n-j}$.*

PROOF: A simple application of the above lemma and Hölder's inequality. □

Since the marginal chain $\{x_k\}$ is reversible, the forward operator F_x and the backward operator B_x corresponding to it are identical and self-adjoint. The self-adjoint operator satisfies

$$\| F_x^n \| = \| F_x \|^n \ .$$

Therefore the spectral radius of the operator $R = \lim \| F_x^n \|^{\frac{1}{n}}$ is the same as the norm $\| F_x \|$. By lemma 1.2.4, $\| F_x \| = \gamma$ where γ is the maximal correlation between x_0 and x_1. However, in the special setting of data augmentation where the interleaving Markov property is present, γ can be further related to the maximal correlation γ_0 between x and y whose joint distribution is $\pi(x, y)$.

Lemma 2.2.3 *Let γ be the maximal correlation between x_0 and x_1 in the marginal chain, γ_0 be the maximal correlation between x and y whose joint distribution is $\pi(x, y)$. Then $\gamma = \gamma_0^2$.*

PROOF: Without loss of generality, we assume all functions used here have mean zero and variance one. Since $\gamma_0^2 = \sup_{\|t\|=1} E(E^2(t(x)|y))$,

$$\gamma = \sup_{t,s} E(t(x_1)s(x_0)) = \sup_{t,s} E(E(t(x_1)|y_1) \cdot E(s(x_0)|y_1)) \leq \gamma_0^2.$$

The last inequality follows from Cauchy-Schwartz. On the other hand

$$\gamma \geq \sup_{t} E(t(x_1)t(x_0)) = \sup_{t} E(E^2(t(x)|y)) = \gamma_0^2,$$

and thus we obtain the equality. □

Theorem 2.2.1 *For any $t(x) \in L_0^2(\mathcal{X})$, we have the bounds for the autocorrelations:*

$$r_1^n(t) \leq corr(t(x_0), t(x_n)) \leq \| F_x^n \| = \gamma_0^{2n},$$

which are sharp in the sense that they can be attained.

PROOF: By lemma 2.2.2 and an inductive argument, we can easily set up the first inequality. The second inequality follows from the definition of forward operator F, lemma 1.2.4, and the above lemma:

$$r_n(t) = \mathrm{corr}(t(x_0), t(x_n)) = \langle F_x^n t, t \rangle \leq \| F_x^n \| = \gamma_0^{2n}$$

The sharpness of the bounds is given in the following example. □

Example 1. The sharpness of the upper and lower bounds.

Assume that (x, y) is jointly distributed as bivariate normal with mean vector $(0, 0)$ and covariance matrix

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

The conditional distributions are

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

We proceed with the iteration between x and y using data augmentation. It follows that $E(x|y) = \rho y$, and $E(y|x) = \rho x$. Suppose the function of interest is $t(x) = x$, $r_1(t)$ defined in this section is just the square of the correlation coefficient ρ between x and y. Direct calculation shows

$$r_n(t) = E(x_0 x_n) = E(E^2(\cdots E(x|y)|x)\cdots)) = \rho^{2n} = r_1^n.$$

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

On the other hand, Lancaster [33] demonstrated that for bivariate normal variables x and y, the maximal correlation between x and y is just the absolute value of their usual

correlation coefficient ρ. Therefore $\gamma_0 = |\rho|$ in this example. Hence the upper bound is also attained. □

Example 2. Maximal correlation for multivariate normal variables.

Suppose x and y jointly has a nondegenerate multivariate normal distribution, but x and y are vectors instead of scalars, finding their maximal correlation is not so straightforward. We cite a result of Csàki and Fischer [9] and Sarmanov [49]: the pair of functions correspond to the maximal correlations are linear functions in x and y respectively. Hence finding the maximal correlation is equivalent to finding the maximum of the quadratic form

$$a\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}a'$$

with the constraint that $a\Sigma_{xx}a' = 1$, where Σ's are the corresponding covariance matrices, a is the coefficient vector for linear function of x. The value of this maximum is γ_0^2. Using Lagrange multipliers, the above maximization problem is equivalent to finding the maximal eigenvalue of

$$\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx},$$

and the corresponding eigenvector. If we write

$$\Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{yy} \end{pmatrix},$$

a result of Eaton [14] shows that the maximal correlation γ is bounded by

$$\gamma_0^2 \leq \frac{\mu_1 - \mu_n}{\mu_1 + \mu_n} < 1,$$

where $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$ are eigenvalues of Σ. □

2.3 Convergence rate problem

In light of lemma 1.2.3, to obtain geometric convergence for the marginal chain, we only require that the spectral radius of the marginal forward operator is strictly less than one. This is also equivalent to finding conditions so that the maximal correlation between x and y is strictly less than one. The earliest result of this kind was given by Gebelein [18]. Breiman and Friedman [8] also proved similar results in developing the ACE algorithm. Two more conditions are needed in addition to condition (A). We will generalize these conditions to higher dimensional cases to obtain convergence results for the general Gibbs sampler with various scans. The two conditions are termed as (B') and (C') temporarily.

Condition (B')

$$\int\int \frac{\pi^2(x,y)}{\pi_x(x)\pi_y(y)}dxdy < \infty,$$

where π is the equilibrium density of the joint chain, π_x and π_y are the corresponding marginals.

Condition (C') There exists no non-trivial functions $t(x)$ and $s(y)$ such that they are equal almost everywhere.

The proofs of the following lemma and theorem on convergence rate are postponed to section 4.1 where we will present a more general approach. Some relevant results are also found in Breiman and Friedman [8], Lancaster [33] and Renyi [44].

Lemma 2.3.1 *Under condition (B'), the marginal forward operator F_x is compact. There exists a pair of functions $t(x)$ and $s(y)$ such that $corr(t(x), s(y))$ attains its maximum γ_0.*

Theorem 2.3.1 *The chain with starting density $p_0(x,y)$ is functional geometric convergent provided that the conditions (A), (B') and (C') are satisfied. The maximal correlation γ_0*

between x and y can serves as the constant α in the definition 3.

In the case that the chain is started from a fix point (x_0, y_0) instead of a starting distribution, the result will also be the same if we think of the one-step evolved distribution $\pi((x,y)|(x_0,y_0))$ as the starting density. In that case, condition (A) is changed to

$$\int \frac{\pi^2(X|X_0)}{\pi(X)} dX < \infty$$

for the starting point $X_0 = (x_0, y_0)$. Conditions (B') and (C') are similar to those used in Breiman and Friedman [8]. (B') implies the compactness of the forward operator, while (C') guarantees that the spectral radius of the operator is strictly less than one if (B') holds.

If we investigate further, we find an additional interesting phenomenon. In the above discussions, we were dealing with the marginal forward operator of the marginal chain $\{x_k\}$. Hence our convergence result applies to this chain. What about the joint chain?

Corollary 2.3.1 *Under condition (B'), the spectral radius of the forward operator F corresponding to the joint chain $\{X_k = (x_k, y_k)\}$ equals γ_0^2 where γ_0 is the maximal correlation between x and y. However, the norm of F equals γ_0 instead.*

PROOF: We can write down the transition function for the joint chain:

$$K(X_1|X_0) = \pi(x_1|y_1)\pi(y_1|x_0).$$

For any $t(x,y) \in L_0^2(\mathcal{X} \times \mathcal{Y})$, we have

$$Ft(x_0,y_0) = \int t(x_1,y_1)\pi(x_1|y_1)\pi(y_1|x_0)dx_1dy_1 = E(E(t(x,y)|y)|x_0).$$

Under condition (B'), both F and F_x are compact (from lemma 2.3.1). The spectral radii of them are the magnitude of their largest eigenvalues. For $t(x,y)$ to be an eigenfunction of F,

it has to be a function of x alone. Therefore the eigenfunction of F is also an eigenfunction of F_x, and vise versa. This shows that the spectral radii of F and F_x are the same and equal to γ_0^2.

However, the norms of F and F_x are not the same. Suppose $t_0(x)$ and $s_0(y)$ is the pair of functions with norm one which attains the maximal correlation, then they must satisfy the following equations:

$$\begin{aligned} \gamma_0 \cdot t_0(x) &= E(s_0(y)|x), \\ \gamma_0 \cdot s_0(y) &= E(t_0(x)|y). \end{aligned}$$

If we choose $t(x,y) = s_0(y)$,

$$\| Fs_0 \| = \| E(s_o(y)|x) \| = \| \gamma_0 t_0 \| = \gamma_0.$$

Thus $\| F \| \geq \gamma_0$. On the other hand, since for general $t(x,y)$, $E(t(x,y)|y)$ is a function of y alone, by property of maximal correlation,

$$\| Ft \|^2 \leq \gamma_0^2 \cdot \mathrm{var}(E(t(x)|y)) \leq \gamma_0^2 \| t \|^2,$$

which implies $\| F \| \leq \gamma_0$. Therefore we obtain the equality. □

We already know that the norm of F_x is the same as its spectral radius γ_0^2. The above corollary implies that the joint chain is one step behind the marginal chains, although they have the same "speed" (spectral radii) of convergence.

CHAPTER 3

COMPARISONS OF ESTIMATORS AND SCHEMES

3.1 Mixture representation and histogram

As discussed in the introduction, we have to make a choice among the different estimators: histogram, mixture representation or a combination of both. The mixture representation described in section 1.1 utilizes the conditional expectation of the original histogram-type approximation. If the samples are drawn independently, the superiority of the mixture representation is obvious. However, in dependent drawings, the comparison of the two is not yet clear. In this section, our theorems demonstrate that in certain situations, the mixture representation is always better than the histogram, and a combination of both cannot bring us any benefit. The two properties we need for deriving such results have been mentioned in section 2.1 where we obtain the nice form of the autocorrelations:

(i) Reversibility of the chain.

(ii) Interleaving Markov property (definition 4).

These two properties bring us monotonicity and nonnegativity for the autocorrelations of both the x-marginal chain and the y-marginal chain, as functions of lag. The results in this section depend only on these two properties. Hence they can be applied not only to the marginal chains in data augmentation, but also to other interesting cases where the reversibility and interleaving property are satisfied. The following lemma was implied, though not explicitly stated, in section 2.1 where we obtained the structural theorem for autocorrelations in data augmentation.

Lemma 3.1.1 *If the chain $\{x_k, k = 1, 2, \ldots\}$ is reversible and has the interleaving Markov property with conjugate chain $\{y_k, k = 1, 2, \ldots\}$, then for any function $t(\cdot)$ of x and $s(\cdot)$ of y, we have the expressions*

$$cov(t(x_0), t(x_n)) = var(E(E(\cdots E(t(x)|y)|x)|\cdots)), \tag{3.1}$$

$$cov(s(y_0), s(y_n)) = var(E(E(\cdots E(s(y)|x)|y)|\cdots)), \tag{3.2}$$

in which each has n expectation signs condition alternately on x and y.

The proof of this lemma follows from the same argument given in section 2.1. In the following general discussion, the chain $\{x_k\}$ is assumed to be reversible and have the interleaving Markov property with the conjugate chain $\{y_k\}$.

Theorem 3.1.1 *If the chain $\{x_k, k = 1, 2, \ldots\}$ is reversible and has the interleaving Markov property with the conjugate chain $\{y_k, k = 1, 2, \ldots\}$, then for any function $t(\cdot)$ of x we have*

$$var\{t(x_1) + \cdots + t(x_n)\} \geq var\{E(t(x)|y_1) + \cdots + E(t(x)|y_n)\}. \tag{3.3}$$

Therefore the estimator (1.2) derived from the mixture representation is always better than the estimator (1.2) derived from the histogram approximation.

PROOF: Using the above lemma and the Markov properties of both $\{x_k\}$ and $\{y_k\}$, we have

$$\begin{aligned}\mathrm{cov}(t(x_m), t(x_{m+k})) &= \mathrm{cov}(t(x_1), t(x_{k+1})) \\ &= \mathrm{var}(E^2(E(\cdots E(t(x)|y)|x)|y\cdots)),\end{aligned}$$

where there are k expectation signs. Correspondingly, if we write $E(t(x)|y)$ as $s(y)$, then

$$\begin{aligned}\mathrm{cov}(s(y_m), s(y_{m+k}))) &= \mathrm{cov}(s(y_1), s(y_{k+1})) \\ &= \mathrm{var}(E^2(E(\cdots E(t(x)|y)|x)|y\cdots)) \\ &= \mathrm{cov}(t(x_1), t(x_{k+2})).\end{aligned}$$

Since $\mathrm{cov}(t(x_1), t(x_{k+1}))$ is a monotone decreasing function of k, it is obvious that

$$\mathrm{cov}(t(x_1), t(x_{k+1})) \geq \mathrm{cov}(t(x_1), t(x_{k+2})).$$

This implies that any term in the quadratic expansion of $\mathrm{var}\{E(t(x)|y_1) + \cdots + E(t(x)|y_n)\}$ is uniformly smaller than or equal to the corresponding term in the expansion of $\mathrm{var}\{t(x_1) + t(x_2) + \cdots + t(x_n)\}$. Thus the conclusion follows. □

To be more convenient, we write

$$\sigma_m^2(t) = \mathrm{cov}(t(x_0), t(x_m)).$$

From the above theorem, $\{\sigma_k\}$ is a monotone decreasing sequence.

$$\begin{aligned}n^2\mathrm{var}(\hat{t}_1) &= n\sigma_0^2 + 2(n-1)\sigma_1^2 + \cdots + 2\sigma_{n-1}^2, \\ n^2\mathrm{var}(\hat{t}_2) &= n\sigma_1^2 + 2(n-1)\sigma_2^2 + \cdots + 2\sigma_n^2.\end{aligned}$$

The difference between the variances of the two estimates are

$$\mathrm{var}(\hat{t}_1) - \mathrm{var}(\hat{t}_2) = \frac{1}{n^2}[n(\sigma_0^2 - \sigma_1^2) + 2(n-1)(\sigma_1^2 - \sigma_2^2) + \cdots + 2(\sigma_{n-1}^2 - \sigma_n^2)].$$

Often we can reduce the variance substantially by using the mixture estimate. A more striking result is that any linear combination with the histogram approximation will inflate the variance of the estimate. To see the result, we need the following lemma.

Lemma 3.1.2 *For any $n \geq 1$ we have*

$$\begin{aligned} cov(t(x_0), E(t(x)|y_n)) &= cov(E(t(x)|y_0), E(t(x)|y_n)) = \sigma_{n+1}^2, \\ cov(t(x_n), E(t(x)|y_0)) &= cov(E(t(x)|y_0), E(t(x)|y_{n-1})) = \sigma_n^2. \end{aligned}$$

PROOF: If we denote $g(y)$ by $E(t(x)|y)$,

$$\begin{aligned} E(t(x_0)E(t(x)|y_n)) &= E(E(t(x_0)|y_0) \cdot E(E(t(x)|y_n)|y_0)) \\ &= E(g(y_0)E(g(y_n)|y_0)) = E(g(y_0)g(y_n)). \end{aligned}$$

For $\mathrm{cov}(t(x_n), E(t(x)|y_0))$ the conclusion is true because of the equation

$$E(t(x_n)E(t(x)|y_0)) = E((E(t(x_n)|y_{n-1}) \cdot E(E(t(x)|y_0)|y_{n-1})).$$

□

Theorem 3.1.2 *The notations are the same as above. Suppose w_1 and w_2 are nonnegative and $w_1 + w_2 = 1$, then*

$$var(w_1\hat{t}_1 + w_2\hat{t}_2) \geq var(\hat{t}_2).$$

PROOF: We only need to look at the sign and the magnitude of $\mathrm{cov}(\hat{t}_1, \hat{t}_2)$. This can be done by using the above lemma:

$$\begin{aligned} E(t(x_k) \cdot \hat{t}_2) &= \frac{1}{n} \sum_{m=1}^{n} \mathrm{cov}(t(x_k), E(t(x)|y_m)) \\ &= \sigma_{k-1}^2 + \sigma_{k-2}^2 + \cdots + \sigma_1^2 + \sigma_1^2 + \cdots + \sigma_{n-k+1}^2 . \end{aligned}$$

Using the above equation for all k, we can explicitly write out the covariance between the two estimates:

$$\begin{aligned}\mathrm{cov}(\hat{t}_1, \hat{t}_2) &= \frac{1}{n^2}(n\sigma_1^2 + (n-1)(\sigma_1^2 + \sigma_2^2) + (n-2)(\sigma_2^2 + \sigma_3^2) + \cdots + (\sigma_{n-1}^2 + \sigma_n^2)) \\ &\geq \frac{1}{n^2}(n\sigma_1^2 + 2(n-1)\sigma_2^2 + \cdots + 2\sigma_n^2) = \mathrm{var}(\hat{t}_2).\end{aligned}$$

Therefore

$$\mathrm{var}(w_1\hat{t}_1 + w_2\hat{t}_2) = w_1^2\mathrm{var}(\hat{t}_1) + 2w_1w_2\mathrm{cov}(\hat{t}_1, \hat{t}_2) + w_2^2\mathrm{var}(\hat{t}_2) \geq \mathrm{var}(\hat{t}_2).$$

□

From the above, we see that any involvement of $t(x_k)$ will increase the variance of the estimates. In practice, we suggest using the mixture representation alone when the two conditions, reversibility and the interleaving property, are satisfied.

However, it is interesting that, when t is not restricted to functions of a single variable, there exists counter-intuitive examples where the mixture representation has larger variance. It is thus incorrect to take it for granted that the histogram approximation is always inferior to the mixture representation.

Example 3. Here we consider the bivariate normal example which has been introduced previously. The procedure is ordinary data augmentation.

$$x_1 \to y_1 \to x_2 \to y_2 \to \cdots$$

$$x|y \sim N(\rho y, 1-\rho^2) \quad \text{and} \quad y|x \sim N(\rho x, 1-\rho^2)$$

Suppose we are interested in a function of both variables, $t(x, y) = x - by$, where b is some constant. Then

$$E(t(x, y)|y) = (\rho - b)y.$$

Let $X_1 = (x_1, y_1)$ and $X_2 = (x_2, y_2)$ be consecutive samples from the chain, $s(y) = E(t(X)|y)$. Then

$$\begin{aligned} \mathrm{var}(t(X_1) + t(X_2)) &= 2 + 2b^2 - 6\rho b + 2(1 + b^2)\rho^2 - 2b\rho^3, \\ \mathrm{var}(s(y_1) + s(y_2)) &= (\rho - b)^2(2 + 2\rho^2). \end{aligned}$$

From the two equations, we can calculate the difference between the two terms:

$$\mathrm{var}(t(X_1) + t(X_2)) - \mathrm{var}(E(t(X_1)|y_1) + E(t(X_2)|y_2)) = 2(1 - \rho^2)(1 + \rho^2 - b\rho).$$

If we choose $b \geq 2/\rho$, this difference is less than zero, which implies that the variance of the histogram-type approximation to $E(t(X))$ using two samples is less than the mixture-type approximation in this special case. □

Later on, other chains generated by the general Gibbs sampler with various scans will be introduced. Many of these satisfy both reversibility and the interleaving Markov property. For them, similar conclusions concerning variances of different estimators hold.

3.2 Comparison of schemes corresponding to partitioning

The problem of comparing schemes has been introduced in section 1.1. Here we intend to detail our comparisons by making use of the forward operators defined in section 1.2. The three schemes for comparison are

[i] $x|y \ , \ y|x,$

[ii] $x|\{y, z\} \ , \ \{y, z\}|x,$

[iii] $x|\{y, z\} \ , \ y|\{x, z\} \ , \ z|\{x, y\},$

The implications have been given in section 1.1. A general discussion of the ordinary

systematic scan like scheme [iii] will be given in the next section. Relative to [iii], we will refer to scheme [i] as "*collapsing*", and scheme [ii] as "*grouping*".

We note that, relative to scheme [i], an additional variable is imputed in scheme [ii]. Iterating with an extra variable is often done to simplify the mathematics and computations involved. However, this may also severely impair the rate of convergence, increase the autocorrelations between samples and, as a consequence, inflate the variance of the resulting estimator. A better understanding of the trade-offs is desirable.

Having to make a choice between schemes [ii] and [iii] also occurs in applications. Intuitively, we expect scheme [ii] will converge faster and has smaller autocorrelations since y and z are drawn jointly conditional on x. On the other hand, grouping may complicate the computations and is not always feasible. Again there needs to be a compromise between the ease of implementation and the convergence rate. Our results in this section may help one to gain some insight into such comparisons and to decide what kind of compromise should be made. Our first theorem shows that the three schemes do have certain ordering in terms of the norm of respective forward operators. The transition functions corresponding to "*collapsing*", "*grouping*", and the ordinary systematic scan are:

$$\begin{aligned} K_1((x_2,y_2)|(x_1,y_1)) &= \pi(x_2|y_2)\pi(y_2|x_1), \\ K_2((x_2,y_2,z_2)|(x_1,y_1,z_1)) &= \pi(x_2|y_2,z_2)\pi(y_2,z_2|x_1), \\ K_3((x_2,y_2,z_2)|(x_1,y_1,z_1)) &= \pi(x_2|y_2,z_2)\pi(y_2|x_1,z_2)\pi(z_2|x_1,y_1). \end{aligned}$$

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

Theorem 3.2.1 *The norms of the above three forward operators have the ordering:*

$$\| F_1 \| \leq \| F_2 \| \leq \| F_3 \| .$$

Furthermore, the spectral radius of scheme [i] is less than or equal to that of scheme [ii].

PROOF: For any function $t(x)$ of a single variable, we notice that

$$E(t(x)|y) = E(E(t(x)|(y,z))|y).$$

Therefore

$$\mathrm{var}(E(t(x)|y)) \leq \mathrm{var}(E(t(x)|(y,z))),$$

which implies that the maximal correlation between x and y is always smaller than that between x and (y,z). Making use of corollary 2.3.1, we conclude that the claims on schemes [i] and [ii] are true.

For the second inequality, we need to prove, for any $s \in L_0^2(x,y,z)$,

$$\mathrm{var}(F_2 s(x,y,z)) \leq \mathrm{var}(F_3 s(x,y,z)).$$

This follows from

$$\begin{aligned} K_2((x_2,y_2,z_2)|(x_1,y_1,z_1)) &= \pi(x_2|y_2,z_2)\pi(y_2|x_1,z_2)\pi(z_2|x_1) \\ &= \int \pi(x_2|y_2,z_2)\pi(y_2|x_1,z_2)\pi(z_2|x_1,y_1)\pi(y_1|x_1)dy_1 \\ &= E(E(K_3((x_2,y_2,z_2)|(x_1,y_1,z_1))|x_1). \end{aligned}$$

Using Rao-Blackwell type inequality, we obtain

$$\begin{aligned} \| F_2 s \|^2 &= E(E_2^2(s(x_2,y_2,z_2)|x_1,y_1,z_1)) \\ &= E(E^2(E_3(s(x_2,y_2,z_2)|x_1,y_1,z_1))|x_1)) \\ &\leq E(E_3^2(s(x_2,y_2,z_2)|x_1,y_1,z_1)) = \| F_3 s \|^2, \end{aligned}$$

where E_2 indicates the conditional expectation under the transition function K_2, and E_3 indicates the conditional expectation under transition K_3. □

An intuitive idea behind this theorem and the proof is the concept of maximal correlation between random variables. By lemma 1.2.4 and corollary 2.3.1, the norm of the operator corresponding to scheme [i] equals the maximal correlation between x and y. Similarly for scheme [ii], the norm of F_2 is exactly the maximal correlation between x and (y, z). From the regression point of view that adding one more regressor will definitely increase R^2, we can intuitively see that with one more variable in second scheme, the corresponding maximal correlation is definitely larger. Therefore, if z is not very dependent on x, it will be fine to introduce it if that simplifies computations. If, on the other hand, z is highly dependent on x, we should probably avoid such a procedure and try some other ways. The comparison of [ii] and [iii] can also be explained in terms of maximal correlations. Further analysis can actually show that the norm of F_3 is equal to the maximal correlation between $\{y_2, z_2\}$ and $\{x_1, y_1\}$ where they have a joint distribution

$$\pi(y_2|z_2, x_1)\pi(x_1, y_1, z_2) \quad \text{or} \quad \pi(x_1, y_2, z_2)\pi(y_1|x_1, z_2).$$

For the same reason, it should be greater than or equal to the maximal correlation between $\{y_2, z_2\}$ and x_1 which is just the norm of F_2. Accordingly, we can also give some suggestion for practitioners in this situation: if any pair of the three variables is highly dependent, grouping these two together can substantially increase the convergence rate and reduce the autocorrelations. Later on, we will show that even when scheme [iii] is not ergodic, scheme [ii] can still be fine.

The second inequality in the theorem can be generalized to higher dimensions where we may group more than two variables together. One may refer to next section for the definition of the ordinary systematic scan for the general Gibbs sampler. We will omit the proof since it is similar to the two variable case.

Corollary 3.2.1 *Let $A = \{1, \ldots, l\}$ with $l < d$. Suppose $\{x(i) : i \in A\}$ can be grouped together. Then the forward operator F_1 of the scheme*

$$\{x(1), \ldots, x(l)\} \to x(l+1) \to \cdots \to x(d)$$

has a smaller norm than that of the forward operator F_2 correspond to the ordinary systematic scan

$$x(1) \to x(2) \to \cdots \to x(d).$$

Comparison of norms of operators can suggest the superiority of certain schemes and give us some insight into the structure of the Gibbs sampler. However, this comparison is far from "exact". The convergence rate will depend on more detailed interaction between variables, the function of interest, and also the starting density. The first inequality in the theorem only suggests that for "almost all" function $t(\cdot)$, $t(x_k)$ converges faster to the equilibrium in scheme [i] than in scheme [ii]. Using some spectral analysis as in section 5.1, we may write

$$t(x) = t_0(x) + \sum_{i=1}^{\infty} \sum_{j=1}^{n_i} \alpha_{ij} \phi_{ij}$$

where $\phi_{11}, \phi_{12}, \ldots, \phi_{1n_1}$ are basis for Λ_1, the eigenspace corresponding to the largest eigenvalue of the marginal forward operator F_x. If we use Λ_1^1 and Λ_1^2 to denote such spaces for the marginal forward operators F_{1x}, F_{2x} corresponding to schemes [i] and [ii] respectively, it is of measure zero that all the $\alpha_{1,j}$'s, which correspond to the largest eigenvalue of scheme [ii], are zero. Asymptotically, the largest eigenvalue term will dominate as k goes to infinity. Therefore, for large k we will almost surely have

$$\| F_{1x}^k t \| \leq \| F_{2x}^k t \|,$$

though it is still possible to find some function orthogonal to Λ_1^2, but not orthogonal to Λ_1^1.

Example 4. Let $x = (x(1), x(2))$ be bivariate normal with mean zero and identity covariance matrix, and y and z are independent standard one dimensional normal variates. Jointly, $(x(1), x(2), y, z)$ has covariance matrix ($\rho > 0$)

$$\begin{pmatrix} 1 & 0 & \rho & 0 \\ 0 & 1 & \rho & \rho \\ \rho & \rho & 1 & 0 \\ 0 & \rho & 0 & 1 \end{pmatrix}.$$

The maximal correlation between x and y is $\sqrt{2}\rho$ which can be attained by the pair of functions $(x(1) + x(2))/\sqrt{2}$ and y. However, the maximal correlation of x and (y, z) is the square root of the largest eigenvalue of

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \rho & 0 \\ \rho & \rho \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \rho & \rho \\ 0 & \rho \end{pmatrix} = \rho^2 \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}.$$

The two eigenvalues are: $\lambda_1 = \frac{3+\sqrt{5}}{2}\rho^2$ and $\lambda_2 = \frac{3-\sqrt{5}}{2}\rho^2$, with two orthogonal eigenvectors $(1, \frac{\sqrt{5}-1}{2})$, $(1, -\frac{\sqrt{5}+1}{2})$ respectively. If we take

$$t(x) = \frac{\sqrt{5}+1}{2}x(2) - x(1),$$

we can get the following from direct algebraic calculations:

$$\begin{aligned} F_{1x}^n t(x) &= \frac{\sqrt{5}-1}{2}\rho^2(2\rho^2)^{n-1}(x(1) + x(2)), \\ F_{2x}^n t(x) &= (\frac{3-\sqrt{5}}{2}\rho^2)^n t(x). \end{aligned}$$

Therefore, for this special $t(x)$, scheme [i] is worse than scheme [ii] in the sense that the convergence rate of $t(x_k)$ is slower and the long term correlations is larger in scheme [i]. As a consequence, the usual estimator $\hat{t}$ of $t(x)$ where

$$\hat{t} = \frac{t(x_1) + t(x_2) + \cdots + t(x_n)}{n}$$

has larger variance by using scheme [i] for this special t. □

The above example elaborates the point that in some extreme cases, scheme [ii] can be better than scheme [i], although the probability of the set of such cases is zero in asymptotic sense. The comparison between schemes [ii] and [iii] is even more unclear. It may be the case that F_3 has smaller spectral radius than F_2 even though the norm of F_3 is larger. The following example demonstrates this possibility.

Example 5. Let $(x, y, z) \sim N(0, \Sigma)$ with

$$\Sigma = \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & \rho \\ 0.5 & \rho & 1 \end{pmatrix}.$$

Then spectral radius of F_2 is therefore the square of the maximal correlation between x and y, z. Its value can be derived analytically:

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

However, if we denote Q by Σ^{-1}, where $Q = (q_{ij})$, the spectral radius of F_3 is the length of the eigenvalue with largest norm 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}. Especially when $\rho = 0$, the spectral radius of F_3 is $\frac{1}{3}$, while the spectral radius of F_2 is 0.5. However the relationship for the norms, $\| F_2 \| = 0.707 < 0.72 = \| F_3 \|$, still holds. Moreover, $\text{spec}(F_2) > \text{spec}(F_3)$ when $\rho < 0.25$.

This example can be generalized a little bit. Consider the situation where the covariance matrix is

$$\Sigma = \begin{pmatrix} 1 & a & a \\ a & 1 & b \\ a & b & 1 \end{pmatrix}.$$

Let a be fixed and b change, the behaviors of the largest eigenvalues corresponding to the two operators F_2 and F_3 are displayed in Figure 1, which shows that when b is small, the ordinary systematic Gibbs sampler corresponding to the operator F_3 will dominate the grouping scheme corresponding to F_2, in the sense that F_3 has a smaller spectral radius. The four plots correspond to four different values of a, while b is continuously changed from -0.4 to 0.4. □

However, even with above two "unpleasant" examples, we can still prove a "pleasant" result that the ergodicity of scheme [iii] always implies the ergodicity of the previous two, and [ii] implies [i].

Corollary 3.2.2 *If the forward operators corresponding to the three schemes are all compact, the ergodicity of [iii] implies the ergodicity of [ii], and [ii] implies [i].*

PROOF: The ergodicity of [ii] means that the maximal correlation between x and $\{y, z\}$ is strictly less than one. So it is obvious that the maximal correlation between x and y is less than one. Therefore scheme [i] is ergodic too. However, if scheme [ii] is not ergodic, there exists a non-trivial $t(x)$ such that

$$E(E(t(x)|y, z)|x) = t(x).$$

Let $s(y, z) = E(t(x)|y, z)$. Without loss of generality we can assume that $\mathrm{var}(t(x)) = 1$. It follows that $\mathrm{var}(s(y, z)) = 1$. Now we can proceed to prove that $t(x)$ is also an eigenfunction of F_3 so that the spectral radius of F_3 is one. We only need

$$E(s(y, z)|x, z) = t(x) \qquad a.e. \tag{3.4}$$

so that

$$E(E(E(t(x)|y, z)|x, z)|x, y) = E(E(s(y, z)|x, z)|x, y) = t(x).$$

But (3.4) follows from

$$t(x) = E(s(y,z)|x) = E(E(s(y,z)|x,z)|x).$$

So

$$\mathrm{var}(t(x)) \leq \mathrm{var}(E(s(y,z)|x,z)) \leq \mathrm{var}(s(y,z)) = \mathrm{var}(t(x))$$

leads to the conclusion. □

The example that [ii] is ergodic when [iii] is not can occur in genetic pedigree analysis where the classical positivity condition is generally not satisfied with the naive application of the Gibbs sampler without grouping. For some numerical results in this area of applications, see Kong et al [31].

In the following chapters, we will apply our method to the general Gibbs sampler for many variables by going through various scans in detail. The correlation structures of these scans are in general different. The convergence rate results, however, share the same conditions for all the scans we discuss. Some convergence rate results for the systematic scan may have been obtained partially by others in one way or another, for example Schervish and Carlin [50]. Our main contribution on this subject is to use a simple and unified approach to obtain more general results.

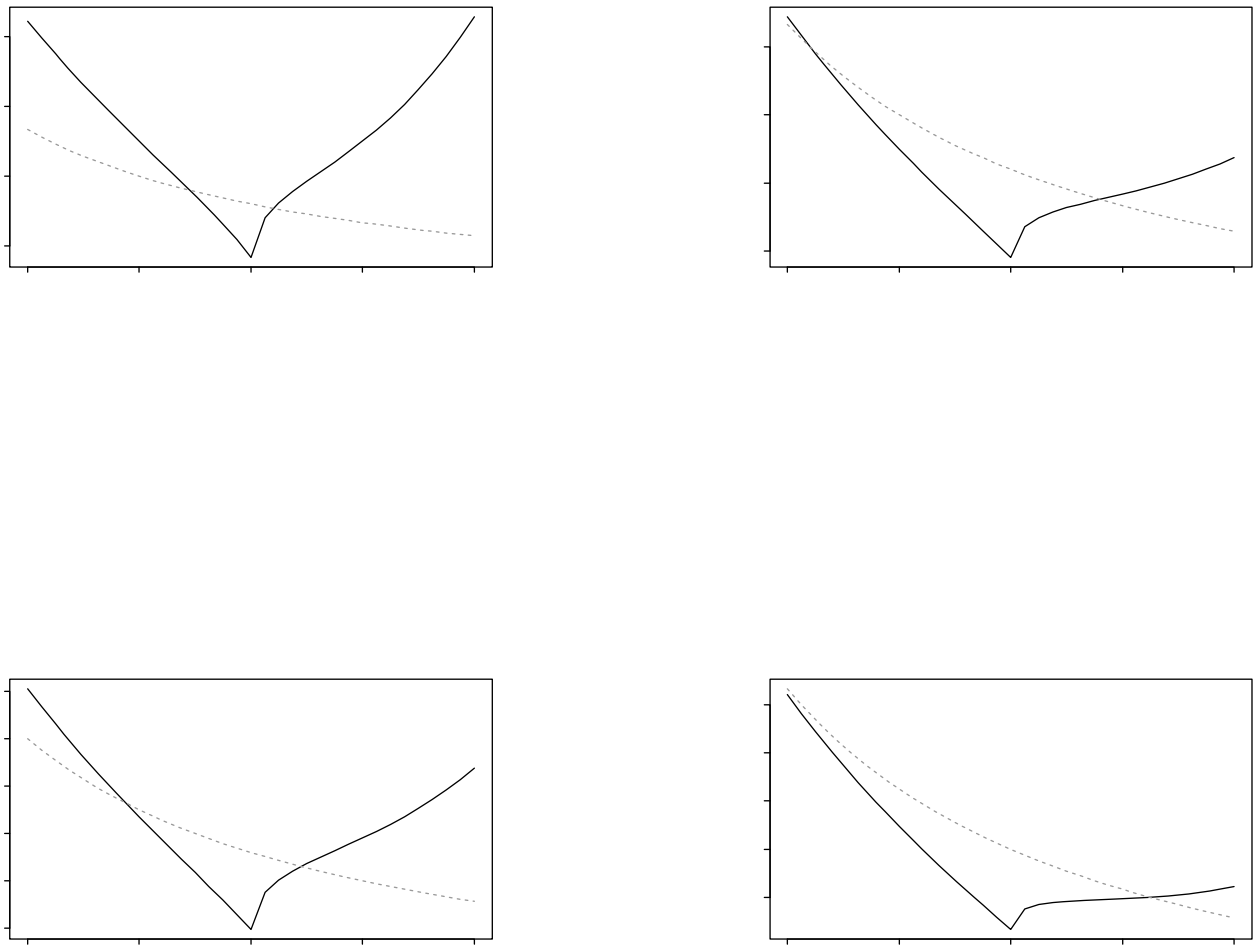


Figure 1: Figure 1: The comparison of the largest eigenvalues of F_2 and F_3, the forward operators corresponding to the grouping and the ordinary schemes respectively. F_2 — dash, F_3 — solid.

CHAPTER 4

RESULTS FOR DIFFERENT SCANS

4.1 Systematic scan Gibbs sampler

The first significant modern development of the Gibbs sampler is due to Geman and Geman [21] when they applied the method to image restoration, where $X = \{x(1), \ldots, x(d)\}$ is a collection of d random variables representing grey levels at the d sites (or pixels) of an image. In a binary image (Ising Model), $x(i)$ can either be -1 or 1. But, in general, X can be continuous. In this section we will focus on the convergence rate of the ordinary systematic scan (OSS). This scan can be described by the updating scheme

$$x(i_1) \to x(i_2) \to \cdots \to x(i_d),$$

where $(i_1, \ldots, i_d)$ is a permutation of the sites $(1, \ldots, d)$ which has been determined in advance. We are interested in getting samples from the joint density of $(x(1), x(2), \ldots x(d))$, i.e., X, or more generally any square integrable function $t(X)$ of the joint space. The Gibbs sampler prescribes a way to reach this goal by evolving a Markov chain. The corresponding

transition function is

$$K(Y|X) = \pi(y(i_1)|X^{-i_1})\pi(y(i_2)|X^{-\{i_1,i_2\}}, y(i_1)) \cdots \pi(y(i_d)|Y^{-i_d}),$$

where X^{-A} denotes components of X excluding those sites in A, i.e., $X^{-A} = \{x(i) : i \in A^c\}$, for A a subset of $I = \{1, 2, \ldots, d\}$, π is the equilibrium measure. The intuitive implication of this scan is that we first draw $x(i_1)$ conditioned on the current states of the rest components, then draw $x(i_2)$ the same way, then $x(i_3)$, etc, until $x(i_d)$. After such d updatings, we say that our Markov chain moves one step. We keep moving our Markov chain until the equilibrium is attained. In practice, people may use more sophisticated scans, for example, updating by alternating coding sets. When the sites is assumed to have a nearest neighbor Markov structure, the coding method turns a many component Gibbs sampler into a two component Gibbs simpler, in which we can still apply all the results obtained in this section.

To obtain geometric convergence, we need to bound the norm of the operators F and B. Such bounds can be derived under the following two conditions which are just the generalizations of conditions (B') and (C').

Condition (B). *There exists at least one permutation* $i_1, \ldots, i_d$ *of* $1, \ldots, d$ *such that*

$$\int \left(\frac{\pi(y(i_1)|X^{-i_1})\pi(y(i_2)|X^{-i_1-i_2}, y(i_1)) \cdots \pi(y(i_d)|Y^{-i_d})}{\pi(Y)}\right)^2 \pi(X)\pi(Y)dXdY < \infty. \quad (4.1)$$

The above condition can also be written in a simpler form

$$\int \frac{p^2(X,Y)}{\pi(X)\pi(Y)} dXdY < \infty,$$

where $p(X, Y)$ is the joint distribution of the two consecutive samples X and Y from the Markov chain. Hence the maximal correlation argument can be naturally applied here.

Condition (C). *There is no non-constant function $t(X)$ to satisfy*

$$E(t(X)|X^{-i}) = t(X) \ \ a.e. \ \ \forall i. \tag{4.2}$$

In almost all existing results on the convergence rate of Gibbs sampling, it is assumed that the support space for the stationary distribution $\pi(X)$ is the entire joint state space, which is also called the *positivity condition* (see Besag [5]). It follows from the following lemma that the positivity condition is stronger than condition (C) which is used in this paper.

Lemma 4.1.1 *$\pi_i(x)$ represents the marginal density of $x(i)$. Then the traditional condition that on the support of $\prod \pi_i(x(i))$*

$$\frac{\pi_1(x(1))\pi_2(x(2))\cdots\pi_d(x(d))}{\pi(X)} < \infty$$

implies condition (C).

PROOF: (Discrete case)

Suppose the equilibrium distribution $\pi(X)$ satisfies the positivity condition. In other words, the support of $\pi(X)$ can be expressed as

$$I_1 \times I_2 \times \cdots \times I_d,$$

where $I_i = \{x(i) : \pi_i(x(i)) > 0\}$.

We argue by contradiction. Assume such non-trivial function $t(X)$ exists to violate condition (C), so that $\exists X_1, X_2$ such that $t(X_1) < t(X_2)$ and $\pi(X_1) > 0$, $\pi(X_2) > 0$. Then we can find a "path" from X_1 to X_2 in which we construct

$$Y_1 \quad = \quad (x_2(1), x_1(2), \cdots, x_1(d)),$$

$$\begin{aligned} Y_2 &= (x_2(1), x_2(2), x_1(3), \cdots, x_1(d)), \\ &\cdots \\ Y_{d-1} &= (x_2(1), \cdots x_2(d-1), x_1(d)), \\ Y_d &= X_2 = (x_2(1), \cdots, x_2(d)), \end{aligned}$$

so that

$$X_1 \to Y_1 \to \cdots \to Y_d = X_2$$

by changing one coordinate each step. Since $\pi(X_1) > 0$,

$$\pi_i(x_1(i)) > 0, \text{ for all } i.$$

For the same reason,

$$\pi_i(x_2(i)) > 0, \text{ for all } i.$$

Then by the positivity condition it is straightforward that $\pi(Y_i) > 0$ for all i. However our condition (C) says that

$$E(t(X)|X_1^{-1}) = t(x, X_1^{-1}), \text{where } x \text{ is arbitrary,}$$

which implies that $t(X_1) = t(Y_1)$. By similar arguments we can conclude also that

$$t(Y_1) = t(Y_2) = \cdots = t(Y_d)$$

which contradicts our previous assumption that $t(X_1) < t(X_2)$.

Since the continuous case can be dealt with in the same spirit, but needs much more lengthy technical work, it is omitted. □

Since all the three conditions will be fully used in the later context, illustration of their implications in detail is needed.

Condition (A) can be visualized in the discrete case where it implies that $p_0(X)$ can not assign probability to those points having π zero probability. This condition is quite natural since otherwise we may have some nonzero probability of getting out of the system.

Condition (B) is a regularity condition which has been used by many authors, for example Breiman and Friedman [8], Schervish and Carlin [50]. It is a standard condition, but is also very hard to check and understand. Intuitively this condition reflects certain dependency between the two consecutive states X and Y, and also provides some restrictions on the shape of the support region of the probability distribution. For example, a "thorn" shape region has to be eliminated. Some kinds of degeneracy of the distribution are prohibited.

Condition (C) is quite awkward from the practitioner's point of view. Some illustration will be helpful. Roughly speaking, it prevents the random variables to be totally dependent. Let us look at the 2-dimensional case first: condition (C) means that there are no non-constant functions $f(x)$ and $g(y)$ so that they are equal almost everywhere. In the discrete case, the equilibrium distribution can be written in the matrix form:

$$\Pi = \begin{pmatrix} p_{11} & p_{12} & p_{13} & \cdots \\ p_{21} & p_{22} & p_{23} & \cdots \\ p_{31} & p_{32} & p_{33} & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix},$$

and the function defined on $\mathcal{X} \times \mathcal{Y}$ can also be expressed in the matrix form:

$$t = \begin{pmatrix} t_{11} & t_{12} & t_{13} & \cdots \\ t_{21} & t_{22} & t_{23} & \cdots \\ t_{31} & t_{32} & t_{33} & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix}.$$

We can see that condition (C) implies that the matrix Π is "connected" in the sense that there is no nontrivial decomposition $\mathcal{X}_1$ of $\mathcal{X}$ and $\mathcal{Y}_1$ of $\mathcal{Y}$ such that

$$p_{kl} = 0 \quad \forall\, (x_k, y_l) \in \mathcal{X}_1 \times \mathcal{Y}_1^c \quad or \quad \mathcal{X}_1^c \times \mathcal{Y}_1,$$

which means that the matrix Π cannot be rearranged to have the diagonal form:

$$\begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix},$$

where $A \neq 0$, $B \neq 0$. This argument can be extended from discrete distributions to the general bivariate case which is that the support for the density $\pi(x, y)$ cannot be decomposed into two nontrivial measurable subsets A and B so that there exist subsets $S_x \in \mathcal{X}$ and $S_y \in \mathcal{Y}$ such that $\pi(A \cap S_x \times S_y) = \pi(A)$ and $\pi(B \cap S_x^c \times S_y^c) = \pi(B)$. ($\pi(\cdot)$ is also used to denote the equilibrium probability measure of a set). We call a measure with this property a decomposable probability measure. Therefore condition (C) prohibits the equilibrium to be a decomposable probability measure. The implications of condition (C) in the multidimensional case can be similarly generalized using the above arguments. It implies that any 2-dimensional marginal probability measure $\pi_{ij}(x(i), x(j))$ of the joint equilibrium measure $\pi(x(1), \ldots, x(d))$ cannot be decomposable.

Lemma 4.1.2 *Condition (B) implies that the forward operator F with the updating order $i_1, \ldots, i_d$ is (sequentially) compact.*

PROOF: For simplicity, we may assume $i_1, \ldots, i_d$ to be $1, \ldots, d$. Hence for any $t \in L_0^2(X)$

$$\begin{aligned} Ft(X) &= \int t(Y)K(Y|X)dY \\ &= \int \frac{\pi(y(i_1)|X^{-i_1})\pi(y(i_2)|X^{-i_1-i_2}, y(i_1))\cdots\pi(y(i_d)|Y^{-i_d})}{\pi(Y)}\pi(Y)dY. \end{aligned}$$

It is an integral operator. According to Example 2, page 277 of Yosida [59], we have sequential compactness under condition (B). □

Lemma 4.1.3 *Conditions (B) and (C) imply that the spectral radius of the forward operator F corresponding to OSS with updating order $i_1, \ldots, i_d$ is strictly less than one.*

PROOF: Because the norm of F is less than or equal to one, it is automatic that all the eigenvalues are in the interval $[-1, 1]$. By compactness of the operator derived from lemma (4.1.2), we know that the spectrum of it is countable with zero as the only possible accumulating point. Hence there is an eigenfunction corresponding to the largest absolute eigenvalue λ_1. If $|\lambda_1| = 1$,

$$Ft(X) = \lambda_1 t(X) \quad a.e.$$

Therefore

$$\text{var}(Ft(X)) = \text{var}(t(X)). \tag{4.3}$$

Since $Ft(X) = E(t(Y)|X)$, by the Rao-Blackwell theorem, the equality (4.3) holds only if

$$t(X) = E(t(Y)|X) = Ft(X) \quad a.e.$$

Hence $\lambda_1 = 1$, the above equation automatically implies that

$$t(X) = E(t(X)|X^{-i}) \ \ a.e. \text{ for all } i$$

because

$$Ft(X) = E(E(\cdots E(E(t(X)|X^{-i_1})|X^{-i_2})\cdots)|X^{-i_d})$$

leads to the identities

$$t(X) = E(t|X^{-i_1}) = E(E(t|X^{-i_1}|X^{-i_2})) = \cdots = Ft(X) \quad a.e.$$

and then we can use induction to get the equality for all $E(t|X^{-i})$. This contradicts condition (C), thus the maximum eigenvalue must be smaller than one. □

Theorem 4.1.1 *The chain with starting density $p_0(X)$ is functional geometric convergent provided that the conditions (A), (B) and (C) are satisfied. The order of updating is assumed to be the same as the one in condition (B).*

PROOF: From lemma 4.1.2 and 4.1.3 we know that the spectral radius of the forward operator F is strictly less than one. Using lemma 1.2.3 we easily get the result. □

Theorem 4.1.2 *Under the setting of this OSS and the three conditions, the absolute value of the autocorrelations between $t(X_0)$ and $t(X_n)$ converges to zero in a geometric rate.*

PROOF: Easy. □

Theorem 4.1.1 is quite similar to the one in a recent paper by Schervish and Carlin (1990) independently. But their methods are quite different from ours since, as Tanner and Wong [54] did, they concentrated on the operator that transforms densities. The results they obtained is almost the same except that we use condition (C) to replace the traditional positivity assumption of the stationary distribution on the entire state space. Theorem 4.1.2 can actually guarantee the asymptotic normality of the estimators

$$\hat{t}_1 = \frac{t(X_1) + \cdots + t(X_n)}{n}.$$

With ordinary systematic scans, if one considers the function $t(X) \in L_0^2(X)$, it will be noted that the autocorrelations are no longer all-nonnegative. Actually, it is possible to

have all-negative autocorrelations in the absence of reversibility and the interleaving Markov property.

Example 6. The distribution is still taken to be the bivariate normal which appeared in section 2.2, i.e, (x, y) is jointly normal with mean vector $(0, 0)$ and variances 1, covariance (also correlation coefficient)ρ. The scanning scheme is the same as in section 2.2:

$$y_1 \to x_1 \longrightarrow y_2 \to x_2 \longrightarrow \cdots$$

We consider a function of the joint variable (x,y)

$$t(x, y) = x - y.$$

Then

$$\begin{aligned} E(t(x_1, y_1)t(x_2, y_2)) &= E(x_1x_2 - x_1y_2 - y_1x_2 + y_1y_2) \\ &= \rho^2 - \rho - \rho^3 + \rho^2 = -\rho(1 - \rho)^2. \end{aligned}$$

If we take $\rho > 0$, then the above value is negative. By induction, it can be found that

$$E(t(x_1, y_1)t(x_{n+1}, y_{n+1})) = -\rho^{2n-1}(1 - \rho)^2.$$

Therefore all the autocorrelations can be negative for this kind of systematic scan.

The example illustrates that the autocorrelations for such non-reversible chain are more complicated. The special functional form we deal with will play a more important role in this case. However, in the following section, it can be shown that the random scan Gibbs sampler possesses exactly the same property as data augmentation, so that the autocorrelations in that case is nonnegative and monotone decreasing.

If one is interested in the function $s(x(i))$ of only one variable, a preliminary result on the one lag correlation can be derived.

Proposition 1 *Suppose X and Y are two consecutive samples from the stationary Markov chain constructed by OSS, then for any square integrable centered function s of the i-th covariate, we have the expression*

$$E(s(x(i))s(y(i))) = E(E^2(s(x(i))|X^{-i})). \tag{4.4}$$

It is greater than or equal to the 2-lag autocorrelation.

PROOF: Without loss of generality, we assume the updating order being $(1, \ldots, d)$. The joint density function of X and Y can therefore be written down:

$$p(X,Y) = K(Y|X)\pi(X) = \pi(y(1)|Y^{-1})\pi(y(2)|Y^{-1-2}, x(1)) \cdots \pi(y(d)|X^{-d})\pi(X).$$

From the transition we can easily see that $x(i)$ and $y(i)$ are conditionally independent given $x(i+1), \ldots, x(d), y(1), \ldots, y(i-1)$ which will be denoted by U. Furthermore, $(x(i), U)$ has the same distribution as $(U, y(i))$ which is the equilibrium π. Therefore

$$\begin{aligned} E(s(x(i))s(y(i))) &= E(E(s(x(i))|U) \cdot E(s(y(i))|U)) \\ &= E(E^2(s(x(i))|X^{-i})). \end{aligned}$$

For the 2-lag autocorrelation, we assume X and Z are two lags apart with Y in between, then it is clear that

$$E(s(x_1)s(z_1)) = E(E(s(x_1)|X^{-1}) \cdot E(s(z_1)|Y^{-1})) \leq E(E^2(s(x(i))|X^{-i})).$$

It is the same for general i. □

If the autocorrelations for higher lags go to zero fast, this one lag autocorrelation will play the major role in determining the variance of the unbiased estimator $\hat{s} = (s(x_1(i)) + s(x_2(i)) + \cdots + s(x_n(i)))/n$.

4.2 Random scan

The random scan (RS) was used in Geman and Geman [21] when they applied the Gibbs sampler to image restoration. The convergence property of the Gibbs sampler when the state space is finite discrete was also derived under RS in that paper. However, the results for general state space is not available in Geman and Geman [21]. RS can be described as follows:

With probability α_i where $\sum_{i=1}^{d} \alpha_i = 1$, we randomly choose a site i from I, then replace the value of the random variable $x(i)$ corresponding to that site by a new sample drawn from the conditional distribution $\pi(x(i)|X^{-i})$; proceed until the equilibrium. Here $X^{-i} = X - \{x(i)\}$ has the same meaning as before. In Geman and Geman [21] they assume that all α_i are equal. We will consider the more general case of random scan with selection probability distribution $V = \{\alpha_i\}$. The distribution need not be uniform, but we do require that $\alpha_i > 0$ for all i.

4.2.1 *Correlation structure*

The first interesting property is the reversibility of the chain constructed by RS.

Lemma 4.2.1 *The Gibbs sampler with random scanning satisfies the detailed balance relation, and therefore generates a reversible Markov chain.*

PROOF: Suppose X_1 and X_2 are two consecutive realizations of a Markov chain constructed by the Gibbs sampler. According to the description of the random scan, X_1 and X_2 differ in at most one variable, say i. Then $X_1^{-i} = X_2^{-i}$, and

$$K(X_2|X_1)\pi(X_1) \quad = \quad \alpha_i \pi(X_2|X_2^{-i})\pi(X_1)$$

$$= \alpha_i \pi(X_2|X_2^{-i})\pi(X_1|X_1^{-i})\pi(X_1^{-i})$$

$$= \alpha_i \pi(X_2)\pi(X_1|X_2^{-i}) = K(X_1|X_2)\pi(X_2).$$

Hence the detailed balance relation is satisfied and the chain is reversible. □

Besides the nonnegative even-lag correlations guaranteed by reversibility and corollary 1.2.1, we also have the nonnegative odd-lag correlations in this special scan. The autocorrelations can also be expressed as the variances of certain iterative conditional expectations. To establish these properties, we first look at the 1-lag correlation.

Lemma 4.2.2 *Suppose $X_0, X_1, \ldots, X_n$ are n consecutive realizations of the Gibbs sampler using random scan. For any $t(X) \in L_0^2(X)$ and $E(t^2(X)) = 1$,*

$$E(t(X_0) \cdot t(X_1)) = E\{\sum_{i=1}^{d} \alpha_i E^2(t(X)|X^{-i})\} \geq 0.$$

PROOF: First we note that the following is always true for random scan Gibbs sampler:

$$E(t(X_1)|X_0) = \sum_{i=1}^{d} \alpha_i E(t(X)|X_0^{-i}). \tag{4.5}$$

This expression together with a conditional expectation argument gives us

$$\begin{aligned} E(t(X_0) \cdot t(X_1)) &= E(t(X_0)E(t(X_1)|X_0)) \\ &= E\{\sum_{i=1}^{d} \alpha_i E(t(X)|X_0^{-i})t(X_0)\} \\ &= E\{\sum_{i=1}^{d} \alpha_i E^2(t(X)|X_0^{-i}))\} \geq 0. \end{aligned}$$

The lemma is therefore established. □

By theorem 2.1.2 and above lemma, it is clear that the correlations $\mathrm{corr}(t(X_0), t(X_n))$ are nonnegative and monotone decreasing. We can further establish an analog of theorem 2.1.1 by proving that the chain $\{X_n\}$ constructed by random scanning has the interleaving Markov property.

Lemma 4.2.3 *The chain $X_0, X_1, \ldots$ generated by the random scan Gibbs sampler has the interleaving Markov property.*

PROOF: If we think of X^{-i} as a joint function of X and i, say $X^{-i} = f(X, i)$, with i independently identically distributed according to $V = \{\alpha_i\}$, for the chain $\{X_k\}$ we will have a conjugate chain $\{f(X_k, i_k)\}$ in which $\{i_k\}$ are i.i.d. This conjugate chain plays the role of $\{y_k\}$ in case of data augmentation. Using this notation,

$$Ft(X_0) = E(t(X_1)|X_0) = E(E(t(X_1)|f(X_0, i))|X_0).$$

It is clear that given $f(X_0, i) = X_0^{-i}$, X_1 and X_0 are independent, and given X_1, $f(X_0, i_0)$ and $f(X_1, i_1)$ are independent because i_0 and i_1 are independent. All other conditions for the interleaving Markov property can also be easily checked. □

With reversibility and the interleaving Markov property, we can establish similar results as in data augmentation by using exactly the same arguments. Therefore we have the following results.

Theorem 4.2.1 *Let $X_0, X_1, \ldots,$ be consecutive samples taken from the Markov chain generated by random scanning. For $t(X) \in L_0^2(X)$, the correlation of $t(X_k)$ and $t(X_{k+n})$ is a nonnegative monotone decreasing function of n. It has the expression that*

$$E(t(X_k)t(X_{k+n})) = E(E^2(E(\cdots E(t(X)|f(X, i))|X)|\cdots)), \tag{4.6}$$

where there are $n + 1$ expectation signs, the conditional expectations are taken alternately on $f(X, i) = X^{-i}$ and X.

Bounds similar to those given in section 2.2 can be established for the autocorrelations.

Theorem 4.2.2 *If $t(X) \in L_0^2(X)$, then $r_1^n(t) \leq r_n(t) \leq \| F \|^n$.*

Here $\|F\| = \sup_{\|s\|=1} E(\sum_{i=1}^d \alpha_i E^2(s(X)|X^{-i}))$ is the norm of the forward operator, $r_1(t) = E(\sum_{i=1}^d \alpha_i E^2(t(X)|X^{-i}))$ is the one-lag correlation for $t(X)$.

Furthermore, the results in section 3.1 for comparing different estimators can also be applied here. We only need to substitute x_k there by X_k, and y_k by $X_k^{-i_k}$ in theorem 3.1.1.

4.2.2 *Geometric convergence*

Recall the forward operator for the random scan Gibbs sampler:

$$F \ : \quad t(X) \longrightarrow \sum_{i=1}^d \alpha_i E(t(X)|X^{-i}).$$

We consider a general state space for X. The forward operator F corresponding to RS is generally not a compact one (except the finite discrete state space case). There will be difficulties trying to apply the method in section 4.1 directly. However, by making use of the results there, we can still prove that under the same conditions (B) and (C), the norm of F is strictly less than one.

Theorem 4.2.3 *If conditions (B) and (C) hold, we have $\| F \| < 1$.*

PROOF: We may write the operators

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

then $F = (\alpha_1 A_1 + \alpha_2 A_2 + \cdots + \alpha_d A_d)$. Thus

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

The term corresponding to the permutation $(i_1, \ldots, i_d)$ in condition (B) is an operator

$$A = A_{i_d} A_{i_{d-1}} \cdots A_{i_1}.$$

According to lemma 4.1.2, A is a compact operator. From lemma 4.1.3, there exists n_0 such that $\| A^{n_0} \| < 1$. Since all A_i's satisfy $\| A_i \| \leq 1$, $\forall i$,

$$\begin{aligned} \| (F^d)^{n_0} \| &\leq 1 - (\prod_{k=1}^{d} i_k)^{n_0} + (\prod_{k=1}^{d} i_k)^{n_0} \| A^{n_0} \| \\ &< 1. \end{aligned}$$

We further note that, in random scanning, the forward operator F is self-adjoint. Hence

$$\|F^{n_0 d}\| = \|F\|^{n_0 d},$$

the norm of F is also less than one. □

Corollary 4.2.1 *Under the conditions (A), (B) and (C), the random scan Gibbs sampler is functional geometric convergent, and the autocorrelations* $corr(t(X_0), t(X_n))$ *for any square integrable function* $t(\cdot)$ *is geometrically decreasing.*

PROOF: Follow from lemma 1.2.3 and theorem 4.2.3. □

For discrete and finite state spaces considered in [21], it is a trivial fact that the operator F is compact and self-adjoint. If condition (C) is also satisfied, it has been shown that the norm of F is less than one. Its maximal eigenvalue equals this norm. Theoretically, the maximal eigenvalue can be obtained by finding the largest λ such that nontrivial solution $t(X)$ of the following linear equation exists:

$$Ft(X) = \lambda t(X).$$

It is usually formidable in practice because the system is huge. If we use $I_i = \{1, \ldots, l_i\}$ to denote the possible states of each $x(i)$, the classical proof used by [21] requires that $\pi(X) > 0$ for all $X \in I_1 \times \cdots \times I_d$, where π denotes the stationary distribution. In our proof we replace it by a weaker condition that there exists no nontrivial $t(X)$ such that

$$E(t(X)|X^{-i}) = t(X) \ \ a.e. \ \text{for all } i.$$

This new condition is actually the weakest possible in the sense that if it is violated, there will be a nontrivial function $t_0(X)$ such that $Ft_0(X) = t_0(X)$. Hence the correlations $\text{corr}(t_0(X_n), t_0(X_0))$ will stay the same. The chain $\{t_0(X_n)\}$ is not even ergodic. Therefore, the Gibbs sampler will usually not converge if we cannot start from a "good" starting density. Whereas some functional chains of $\{X_i\}$ may still be geometric ergodic. It is a consequence of the fact that if we choose some function orthogonal to the eigenspace corresponding to the largest eigenvalue, the covariances between $t(X_0)$ and $t(X_n)$ will still converge to zero in geometric rate. Furthermore, if we can choose a starting density which satisfies a certain property, the geometric convergence rate of the density can also be obtained. This idea will be summarized in the section of spectral analysis of the reversible chains.

4.3 Other scans

4.3.1 *Symmetric random permutation scan (SRPS)*

In the case of pure random scanning, it is possible that we choose the same site in consecutive iterations. This is obviously inefficient because repeated drawings from $\pi(x(i)|X^{-i})$ is the same as drawing once. The symmetric random permutation scan (SRPS) is introduced here

to avoid such immediate repetition. The SRPS is an updating scheme which at each step chooses an ordering $(i_1, \ldots, i_d)$ of the d sites according to a certain probability distribution Ψ on the set of the permutations of $1, \ldots, d$, and update $X = (x(1), \ldots, x(d))$ in the order we choose. The distribution Ψ is required to be symmetric in the sense that

$$\Psi(i_1, \ldots, i_d) = \Psi(i_d, \ldots, i_1).$$

It can be visualized heuristically that the Markov chain constructed by the above procedure is reversible because we may think that the reverse of the updating order $(1, \ldots, d)$ is equivalent to updating X using ordering $(d, \ldots, 1)$. A rigorous proof is contained in the following lemma. To be consistent in the thesis, we keep the notations in previous chapters. Further, we use $\mathcal{O}$ to denote the set of permutations of $\{1, \ldots, d\}$. For each permutation $o \in \mathcal{O}$ we use

$$A_o = A_{i_1} A_{i_2} \cdots A_{i_d}.$$

The reverse of the permutation o is denoted by o'. Therefore the forward operator F corresponding to SRPS can be written as

$$F = \sum_{o \in \mathcal{O}} \phi_o A_o,$$

where $\sum_o \phi_o = 1$, and $\phi_o = \phi_{o'}$.

Lemma 4.3.1 *The chain constructed by SRPS is a reversible Markov chain.*

PROOF: For any fixed ordering, say $o = (1, \ldots, d)$, we can write down the one step transition function $K_o(Y|X)$ from X to Y :

$$K_o(Y|X) = \pi(y(1)|Y^{-1})\pi(y(2)|Y^{-1-2}, x_1) \cdots \pi(y(d)|X^{-d}).$$

To obtain reversibility, we only need to check that

$$(K_o(Y|X) + K_{o'}(Y|X))\pi(X) = (K_o(X|Y) + K_{o'}(X|Y))\pi(Y),$$

since the two orderings have the same probability assignment. But it is easy to see that

$$\begin{aligned} K_o(Y|X)\pi(X) &= \pi(y(1)|Y^{-1})\pi(y(2)|Y^{-1-2}, x(1)) \cdots \pi(y(d)|X^{-d})\pi(X) \\ &= \pi(Y)\pi(x(1)|Y^{-1}) \cdots \pi(x(d-1)|X^{-d-(d-1)}, y(d))p(x(d)|X^{-d}) \\ &= \pi(Y)K_{o'}(X|Y). \end{aligned}$$

Therefore the conclusion follows. □

We can easily see that if condition (B) is satisfied, the norm of F is less than one since F is a weighted average of all the differently ordered systematic scan. F is furthermore a self-adjoint operator because of reversibility. Hence the upper bound for the autocorrelations is the power of $\|F\|$. The following theorem is a summary.

Theorem 4.3.1 *If conditions (B) and (C) are satisfied, the norm of the forward operator F is strictly less than one. If condition (A) is also satisfied for the starting density $p_0(X)$, then the chain generated by SRPS is functional geometric convergent. The autocorrelations are also geometrically decreasing.*

By corollary 1.2.1, the reversibility of the chain also provides us with nonnegative monotone decreasing even-lag autocorrelations between $t(X_n)$ and $t(X_0)$. The following example shows that the odd-lag autocorrelations can indeed be negative.

Example 7. Here we again consider the bivariate normal example previously studied. Our objective here is to show that a reversible Markov chain may have negative odd-lag

correlations. Let us consider a new transition function from (x_1, y_1) to (x_2, y_2):

$$K((x_2, y_2)|(x_1, y_1)) = \frac{1}{2}\{p(x_2|y_2)p(y_2|x_1) + p(y_2|x_2)p(x_2|y_1)\},$$

which corresponds to a SRPS in two dimensions. Here we demonstrate that the 1-lag correlation, however, is negative if we choose a positive ρ in the joint distribution of x, y and the same function t as in Example 1.

$$E(t(x_1, y_1)t(x_2, y_2)) = E(x_1x_2) + E(y_1y_2) - E(x_1y_2) - E(x_2y_1).$$

Direct calculations show $E(x_1x_2) = E(y_1y_2) = \rho^2$, which are the same as in example 1; and

$$E(x_1y_2) = E(x_2y_1) = \frac{1}{2}(\rho + \rho^3).$$

Hence $E(t(x_1, y_1)t(x_2, y_2)) = -\rho(1 - \rho^2)$ is negative for any $0 < \rho < 1$. □

Similar to the systematic scan, if we are interested in the function $t(x(i))$ of a single component, it can be shown that the one lag autocorrelation is nonnegative and greater than the two lag autocorrelation.

Proposition 2 *Suppose X and Y are two consecutive samples taken from the stationary SRPS chain. Then for any square integrable centered function t of their i-th covariate, we have the expression*

$$E(t(x(i))t(y(i))) = E(E^2(t(x(i))|X^{-i})). \tag{4.7}$$

It is greater than the 2-lag covariance which equals $E(E^2(t(x(i))|Y))$.

PROOF: From proposition 4.1, for any order o of updating, we always have

$$E_o(t(x(i))t(y(i))) = E(E^2(t(x(i))|X^{-i})).$$

SRPS incorporates a probability distribution on such orders. Hence the expectation taken under SRPS chain is a weighted sum of $E_o(t(x(i))t(y(i)))$. Therefore the expression for the one lag correlation is still the same as that in OSS.

For the second part, we notice that

$$E(t(x(i))|Y) = \sum_{o\in\mathcal{O}} \phi_o A_o t(Y)$$

and we also have $A_o t(Y) = E(E(t(x(i))|x(i+1),\ldots,x(d),y(1),\ldots,y(d-1))|Y)$. Hence the discrete version of Cauchy-Schwartz inequality shows that

$$E(E^2(s(x(i))|Y)) = E((\sum_{o\in\mathcal{O}} \phi_o A_o t(Y))^2) \le E(\sum_{o\in\mathcal{O}} \phi_o (A_o t(Y))^2) \le E(E^2(s(x(i))|X^{-i})).$$

□

4.3.2 *Symmetric systematic scan (SSS)*

This is a kind of systematic scan which also possesses reversibility and the interleaving Markov property. It has already been used in practice, see Johnson, et al [26], but has not been analyzed theoretically.

Description of the scan: suppose $X = \{x(1),\ldots,x(d)\}$, one round of scanning consists of updating $x(1), x(2), \ldots$ one by one until reaching $x(d)$ and then reversing the order and updates $x(d-1), x(d-2), \ldots$, sequentially until returning to $x(1)$. It is illustrated by the following diagram:

$$x(1) \to x(2) \to \cdots x(d) \to x(d-1) \to \cdots \to x(1).$$

The one-step transition function from X to Y can be written as

$$K(Y|X) \quad = \quad \int \pi(z(1)|X^{-1})\,\pi(z(2)|X^{-\{1,2\}}, z(1)) \cdots \pi(z(d-1)|x(d), Z^{-\{d,d-1\}})$$

$$\pi(y(d)|Z^{-d}) \cdots \pi(y(2)|z(1), Y^{-\{1,2\}})\, \pi(y(1)|Y^{-1})\, dZ.$$

In this scan, one step of the Markov chain involves $2d-1$ individual updates. The convergence property of this scan is exactly the same as the systematic scan.

Lemma 4.3.2 *The Markov chain constructed by SSS is reversible.*

PROOF: We need to check the *detailed balance* condition for this chain, i.e., we require the following equality:

$$K(Y|X)\pi(X) = K(X|Y)\pi(Y).$$

Look at the first part of the transition $K(Y|X)$,

$$\begin{aligned} & \pi(X)\pi(z(1)|X^{-1})\pi(z(2)|X^{-\{1,2\}}, z(1)) \cdots \pi(y(d)|Z^{-d}) \\ = \quad & \pi(x(1)|X^{-1})\pi(x(2)|X^{-1-2}, z(1)) \cdots \pi(x(d)|Z^{-d})\pi(Z^{-d}, y(d)). \end{aligned}$$

Continuing with the other half of the transition we have

$$\begin{aligned} & \pi(Z^{-d}, y(d))\pi(y(d-1)|Z^{-\{d,d-1\}}, y(d)) \cdots \pi(y(1)|Y^{-1}) \\ = \quad & \pi(z(d-1)|Z^{-\{d,d-1\}}, y(d)) \cdots \pi(z(1)|Y^{-1})\pi(Y). \end{aligned}$$

Putting the two parts together gives us the detailed balance condition. □

As we noted, Z is of $(d-1)$-dimensional. This Z plays the role of y in data augmentation. It can be easily proved that the chain $\{X_k\}$ constructed by SSS has interleaving property with the conjugate chain $\{Z_k\}$. Therefore we have the correlation structure theorem for SSS.

Theorem 4.3.2 *The correlation between $t(X_k)$ and $t(X_{k+n})$ is a monotone decreasing non-negative function of n. It has the expressions*

$$E(t(X_k)t(X_{k+n})) \quad = \quad E(E^2(E(\cdots E(E(t(X)|Z)|X)\cdots|Z))\,, \quad for \;\; n = odd,$$

$$E(t(X_k)t(X_{k+n})) \quad = \quad E(E^2(E(\cdots E(E(t(X)|Z)|X)\cdots|X)), \quad for \;\; n = even,$$

in which X and Z are half-lag apart as described in the expression of the transition function. The expectations are taken by conditioning alternately on Z and X.

The convergence property is obvious. We will not repeat the proof here. The conditions needed for geometric convergence are still (A), (B) and (C).

CHAPTER 5

FURTHER ANALYSIS AND EXAMPLES

5.1 Spectral analysis on the forward operators

We discuss in this section some finer properties of the self-adjoint compact operators studied. These include the forward operators in data augmentation, the SRPS with strengthened condition (B) and the SSS. The operators with respect to OSS and RS will typically not be included in this group except the case when the state space is discrete and finite. The literature on the theory of self-adjoint compact operators are overwhelming. Here we only make use of some elementary results taken from the theory. Readers who are interested in details of this theory are referred to Dunford and Schwartz [13], Riesz and Nagy [45], and Yosida [59].

From here on we assume F to be a self-adjoint compact operator on the Hilbert space $L_0^2(X)$. According to Riesz-Schauder's theorem (see page 283 of Yosida [59]) on the spectrum of compact operators, there exist countable eigenvalues $\lambda_0, \lambda_1, \lambda_2, \ldots$ and the only

possible accumulating point of them is zero. Without loss of generality, we may assume

$$\lambda_0 = 0, \ \ |\lambda_1| > |\lambda_2| > \cdots$$

Also the eigenspace Λ_i corresponding to λ_i is finite dimensional. If Λ_0 is the null space of the operator F, which means for any $t \in \Lambda_0$ $Ft = 0$, we have the direct sum decomposition of the space

$$\Lambda_0 \oplus \Lambda_1 \oplus \Lambda_2 \oplus \cdots = L_0^2(X),$$

where Λ_i and Λ_j are orthogonal for $i \neq j$ in the sense that $E(t_i(X)t_j(X)) = 0$ for $t_i \in \Lambda_i$, $t_j \in \Lambda_j$.

Since $\Lambda_i, i \geq 1$ are all of finite dimensions, we may find the orthonormal basis for each of such subspaces, say, for example, $\{\phi_{ij}, j = 1, \ldots, n_i\}$ is a basis of Λ_i. All such basis together give a set of orthonormal basis of $\Lambda_0^{\perp}$. Now any $t(X) \in L_0^2(X)$ has an expansion:

$$t = t_0 + \sum_{i=1}^{\infty} \sum_{j=1}^{n_i} \alpha_{ij} \phi_{ij}. \tag{5.1}$$

Hence

$$F^p t = \sum_{i=1}^{\infty} \sum_{j=1}^{n_i} \alpha_{ij} \lambda_i^p \phi_{ij} \tag{5.2}$$

$$\| F^p t \|^2 = \sum \sum \alpha_{ij}^2 \lambda_i^{2p}. \tag{5.3}$$

The last equality is also called Parseval's identity. If the largest eigenvalue λ_1 is less than one, the norm of F^p converges to zero geometrically. But the point we want to make here is that even if the maximal eigenvalue is one, which means the condition (C) is violated, we can still find some starting density which leads to a geometric rate of convergence. It is clear that if we choose some function $s(X) \in \Lambda_1^{\perp}$, from the Parseval identity and the fact

that λ_1 term will be zero after one iteration:

$$\| F^{p+1} s \|^2 \leq \lambda_2^{2p} \| Fs \| .$$

Therefore the iterative use of the operator on such s will make its norm converge to zero geometrically fast.

If the starting density $p_0(X)$ is chosen so as to make $p_0(X)/\pi(X)$ orthogonal to Λ_1, which means that for any $f_1(X) \in \Lambda_1$

$$E_\pi[f_1(X)(\frac{p_0(X)}{\pi(X)} - 1)] = E_{p_0}(f_1(X)) - E_\pi(f_1(X)) = 0,$$

the convergence rate will always be geometric. But unfortunately this is practically useless. Whether this information can be utilized will depend on more detailed knowledge about the problem we deal with. For example, in the bivariate normal case, from the discussion of example 4 we know that if we can choose the starting density to be symmetric about the same mean as the true one, it will double the convergence rater.

Example 8. This example illustrates the spirit of the above discussion. We consider the case $d = 2$, $X \in I_1 \times I_2$ where $I_1 = I_2 = \{1, 2, 3, 4\}$. The stationary distribution is assumed to be

$$P = \begin{pmatrix} 0.15 & 0.15 & 0 & 0 \\ 0.15 & 0.15 & 0 & 0 \\ 0 & 0 & 0.10 & 0.10 \\ 0 & 0 & 0.10 & 0.10 \end{pmatrix},$$

which is decomposable. Therefore we can find a t_0 which violates condition (C)

$$t_0(i,j) = \begin{cases} \sqrt{2/3} & i,j = 1,2, \\ -\sqrt{3/2} & i,j = 3,4, \\ 0 & \text{Otherwise.} \end{cases}$$

If our starting density is chosen so that $p_0(X)/\pi(X)$ is orthogonal to the eigenspace corresponding to the eigenvalue 1, which is a linear space spanned by t_0, the random scan procedure will still converge to the equilibrium in a geometric rate. The possible choices of such P_0 can be characterized by the condition that

$$\sum_{i,j=1}^{2} P(i,j) = \frac{3}{2} \sum_{i,j=3}^{4} P(i,j).$$

On the other hand, if the chosen prior distribution does not satisfy this, the chain will not converge to the required distribution. To be extreme, if we choose starting distribution to be uniform, the Gibbs sampler will have no effects at all, it will remain uniform no matter how many iterations one may do. □

The spectral analysis not only provides us some theoretical understanding of the schemes, but also implies some practical methods to derive useful stopping rules for the Gibbs sampler. In the following, we propose a tentative method.

For a function $t(X) \in L^2(X)$ which is not required to have zero expectation, by the expansion (5.1) we have

$$t(X) = E(t) + t_0(X) + \sum_{i=1}^{\infty} \sum_{j=1}^{n_i} \alpha_{ij} \phi_{ij}(X).$$

If the function $t(X)$ is chosen in random, then with probability one, it will not be orthogonal to the subspace Λ_1. Hence, after n iterations, using the differences between two different

starting values gives

$$F^n t(x) - F^n t(y) = \sum_{i=1}^{\infty} \sum_{j=1}^{n_i} \alpha_{ij} \lambda_i^n (\phi_{ij}(x) - \phi_{ij}(y)).$$

As $n \to \infty$, the largest eigenvalue term will dominate. So,

$$\lim_{n \to \infty} \frac{\log |F^n t(x) - F^n t(y)|}{n} = \log \lambda_1, \tag{5.4}$$

provided that $\sum_j (\phi_{1j}(x) - \phi_{1j}(y)) \neq 0$.

For example, the multivariate normal case has been extensively discussed in the literature, e.g. Amit [2]. In that case, the eigenspace Λ_1 consists of certain linear functions. If we choose random coefficients a_i such that

$$t(X) = a_1 x_1 + a_2 x_2 + \cdots + a_d x_d,$$

there is probability zero that $t(X)$ and Λ_1 are orthogonal. If one can get hold of the analytical expressions of the powers of the operator F, he will likely be able to determine the largest eigenvalue and hence the convergence rate. A straightforward algorithm is to iteratively apply the operator F on such a $t(X)$ and estimate the limiting value of $\frac{1}{n} \log |F^n t(x) - F^n t(y)|$. Here we suggest using the linear test functions for other cases too because of its simplicity and the fact that it is unusual for the forward operator to have the eigenspace corresponding to the largest eigenvalue to be orthogonal to all the linear functions. However, the problem arises when we cannot get the analytic forms of $F^n t$, where we have to use Monte Carlo method. When using simulations to estimate $F^n t(X)$ for $X = x$, estimates in the form of a mean have the errors that will swamp the information about the eigenvalues. Writing out explicitly we have, after m independent runs of the

chain starting from x by n steps,

$$\hat{F^n}t(x) = \frac{t(x_{n1}) + t(x_{n2}) + \cdots + t(x_{nm})}{m}.$$

The error for this estimate is of order $m^{-1/2}$. Hence m has to be exponentially greater than n to make the estimated eigenvalue meaningful. On the other hand, we do not really need this eigenvalue. Our aim is to make sure that the chain has converged. If the chain really converges, the value

$$\frac{1}{n}\log|\hat{F^n}t(x) - \hat{F^n}t(y)| \tag{5.5}$$

will go to zero and maybe fluctuate around zero. Therefore this value can serve as an indicator for checking convergence.

5.2 Examples

1. Gaussian distribution

In this example we consider the application of the Gibbs sampler to Gaussian models. X is assumed to have a nondegenerate multivariate normal distribution with mean zero and the covariance matrix $\Sigma_{d\times d}$ which is unknown to us in practice. Using the notations in Amit [2], we can write out the density function

$$\phi(x) = \frac{det(Q)}{(2\pi)^{\frac{d}{2}}}\exp\{-\frac{1}{2}x^tQx\},$$

where $Q = \Sigma^{-1}$. All the conditional distributions, $\phi(x(i) \mid X^{-i})$, are known to us. By lemma 1 of Amit [2], it can be written as

$$x(i) \mid X^{-i} \sim N\left(-\sum_{j\neq i}\frac{q_{ij}}{q_{ii}}x(j), q_{ii}^{-1}\right),$$

where $Q = (q_{ij})_{d\times d}$. In image processing, these conditional distributions depend only on the "neighborhood structure" of the model. The simplest such case is the so-called "nearest neighbor structure" in which the conditional distribution of the site i is determined by the four nearest neighbors in the lattice structure of the model.

Now consider the ordinary systematic scan in which the transition function from X to Y can be written as

$$K(Y|X) = \phi(y(1)|Y^{-1})\phi(y(2)|Y^{-1-2}, x(1))\cdots\phi(y(d)|X^{-d})$$

where Y and X are identically distributed as $N(0,\Sigma)$. Therefore

$$(y(1),\ldots,y(d),x(1),\ldots,x(d))$$

is also normally distributed with covariance matrix

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

Here Σ_{yx} is determined by the special structure in the real problem where the Gibbs sampler is applied. We start by proving that Ψ is nonsingular. Suppose it is not, then there exist nontrivial vectors $a = (a_1,\ldots,a_d)$ and $b = (b_1,\ldots,b_d)$ such that

$$\sum a_i y(i) + \sum b_i x(i) \equiv 0.$$

Because Y and X are not degenerate, there exists at least one k such that

$$y(k) = \sum_{i>k} a_i y(i) + \sum_{i<k} b_i x(i) + \sum_{i\le k} b_i x(i). \tag{5.6}$$

However, from the sampling scheme, we know that given $y(k+1),\ldots,y(d),x(1),\ldots,x(k-1)$, we have conditional independence between z_k and $x(k),x(k+1),\ldots,x(d)$. Hence

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

which is a linear function of $y(k+1), \ldots, y(d), x(1), \ldots, x(k-1)$. On the other hand, from equation (5.6), we know that

$$E(y(k)|y(k+1), \ldots, y(d), x(1), \ldots, x(d)) = \sum_{i>k} a_i y(i) + \sum_{i<k} b_i x(i) + \sum_{i \leq k} b_i x(i).$$

All b_i , $i \geq k$ must be zero. On the other hand, $(y(k), y(k+1), \ldots, y(d), x(1), \ldots, x(k-1))$ is normally distributed with a nonsingular covariance matrix Σ. This leads to a contradiction.

Now with the nonsingularity result we may apply the method of example 2 in section 2.2. There we pointed out that the maximal correlation between X and Y is the maximal eigenvalue of

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

By lemma (1.2.4), this value equals the norm of the forward operator. An explicit bound can also be expressed as the function of the maximal and minimal eigenvalues of Ψ. To summarize, we have the following theorem.

Theorem 5.2.1 *Suppose X is nondegenerate Gaussian distributed with covariance matrix Σ and mean vector zero, Ψ is the $2d \times 2d$ matrix defined above, then the ordinary systematic scan Gibbs sampler is functional geometric convergent with rate γ where*

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

λ_1 and λ_{2d} are the largest and smallest eigenvalues of the matrix Ψ, and λ_{2d} is strictly greater than zero.

This theorem provides us with some theoretical understanding of the scheme, but it is not very useful in practice because of the lack of knowledge of the covariance matrix Σ. By using the idea of spectral analysis, we propose a practical method to estimate the

convergence rate. Since the the maximal correlation is attained by a pair of linear functions of X and Y, which also corresponds to the eigenspace Λ_1 of the largest eigenvalue of the forward operator, we may start with a linear function t of X by choosing random coefficients $a_1, a_2, \ldots, a_d$ such that

$$t(X) = a_1 x(1) + a_2 x(2) + \cdots + a_d x(d).$$

It is with probability one that this function is not orthogonal to the eigenspace Λ_1. It can be seen that the forward operator of the OSS on t is

$$Ft(X_0) = E(a \cdot Y | X_0) = a \cdot (I - D_1 Q)(I - D_2 Q) \cdots (I - D_d Q) X_0,$$

where $a = (a_1, \ldots, a_d)$, the matrix D_i has all entries zero except for the i-th entry on the diagonal which is q_{ii}^{-1}, $Q = (q_{ij})$ is the inverse of the covariance matrix Σ. At this stage we basically have two ways to proceed:

1. If the conditional expectations are analytically manageable, we can directly assess n-round iterations of the operator F, and use the method in (5.1) to estimate an approximate value of the maximal eigenvalue λ_1.

2. However, analytical calculations are usually formidable in practice. We may employ a Monte Carlo method to estimate. Since the error may soon swamp the information on λ_1 as pointed out at the end of section 5.1, we propose here a modified way of estimation. Because of the linearity of the conditional expectation in the Gaussian case, we may amplify the value by multiplying a constant (3 or 4) during each iteration, and at the end we subtract this amount from the estimate. This can be summarized as follows:

ALGORITHM:

a) From $X = X_0$, start m independent OSS chains. (m need not be too large.) X_0 is chosen randomly.

b) After getting the m independent first round OSS samples $X_{11}, \ldots, X_{1m}$ we multiply each one by c (c can be 3 or 4, in spirit it should be chosen a little bit greater than $\approx 1/\lambda_1$.) Then do the second round OSS starting from the m different values $cX_{11}, \ldots, cX_{1m}$. So we still have m independent OSS chains.

c) Repeat step (b) till a moderate nth round, then start to calculate the number

$$\delta_n = \frac{1}{n}\log|\frac{t(X_{n1}) + \cdots + t(X_{nm})}{m}| - \frac{n-1}{n}\log(c).$$

d) Stop until $|\delta_n - \delta_{n-1}|$ is small. This value is an estimate of the convergence rate.

NOTE: If the mean is not known, we can similarly initiate another set of chains from a different starting value y, and use formula (5.5).

2. Murray's data

The data in the following table are supposed to be drawn from a normal distribution with mean $\mu_1 = \mu_2 = 0$. This data was originally given by Murray [41], and used later by Tanner and Wong [54] to illustrate the data augmentation method.

TABLE 1. Twelve Observations From a Bivariate Normal Distribution With Known Mean

1	1	-1	-1	2	2	-2	-2	*	*	*	*
1	-1	1	-1	*	*	*	*	2	2	-2	-2

* indicates that the data is missing.

Jeffery's non-informative prior is given to the covariance matrix, see Box and Tiao [6]

$$p(\Sigma) \propto |\Sigma|^{-\frac{m+1}{2}},$$

where $m = 2$ is the dimension of the distribution. Our purpose here is not to investigate the posterior distribution of the correlation ρ. We are interested here in whether the conditions (B) and (C) can be verified.

Condition (C) is fine with this example because the densities required are all positive. The main difficulty is condition (B). We can rewrite condition (B) as

$$\int p(X|Y)p(Y|X)dXdY < \infty.$$

Suppose we denote the missing data by x_i and y_i as the follows, corresponding to the original pattern:

$$\begin{array}{cccccccccccc} 1 & 1 & -1 & -1 & 2 & 2 & -2 & -2 & x_1 & x_2 & -x_3 & -x_4 \\ 1 & -1 & 1 & -1 & y_1 & y_2 & -y_3 & -y_4 & 2 & 2 & -2 & -2. \end{array}$$

Using the notations in Tanner and Wong [54], $Z = (x_1, \ldots, x_4, y_1, \ldots, y_4)$ is the latent data and

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

The two conditional distributions are

$$\begin{aligned} p(Z|\Sigma) &= \frac{1}{(2\pi\sigma_1\sigma_2(1-\rho^2))^4}\exp[-\frac{1}{2(1-\rho^2)}\{\frac{\sum x_i^2}{\sigma_1^2}+\frac{\sum y_i^2}{\sigma_2^2}-\frac{4\rho\sum(x_i+y_i)}{\sigma_1\sigma_2}+(\frac{16\rho^2}{\sigma_2^2}+\frac{16\rho^2}{\sigma_1^2})\}], \\ p(\Sigma|Z) &= \frac{|\Sigma|^{-15/2}|B|^6}{(2\pi)^12}\exp[-\frac{1}{2(1-\rho^2)}\{\frac{\sum x_i^2}{\sigma_1^2}+\frac{\sum y_i}{\sigma_2^2}-\frac{4\rho\sum(x_i+y_i)}{\sigma_1\sigma_2}+\frac{20}{\sigma_2^2}+\frac{20}{\sigma_1^2}\}], \end{aligned}$$

where

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

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

$$\frac{1}{2(1-\rho^2)}\{2\sum_i(\frac{x_i}{\sigma_1}-\frac{2\rho}{\sigma_2})^2+2\sum_i(\frac{y_i}{\sigma_2}-\frac{2\rho}{\sigma_1})^2+\frac{20-16\rho^2}{\sigma_1^2}+\frac{20-16\rho^2}{\sigma_2^2}\}.$$

Since $|B|^6$ is a polynomial of x_i and y_i with the maximal degree 12 on each x_i and y_i, when we integrate out x_i and y_i we get moments up to order 12. The result is a polynomial on $1/\sigma_1$, $1/\sigma_2$ and is also a rational function of ρ. Together with all the other coefficients left, we get an integral

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

where f is a polynomial on first two arguments, a rational function on the third one. The integral is less than infinity because of the exponential term. Thus we have verified condition (B). The convergence rate of data augmentation is therefore geometric in this case .

CHAPTER 6

A REVIEW OF STOCHASTIC RELAXATION TECHNIQUES

6.1 The Metropolis Algorithm

The root of the Gibbs sampler can be traced back to the work of Metropolis et al. [39], a group of physicists and chemists, who invented a method for investigating equations of state for substances consist of interacting individual molecules. In order to solve the problem they had to carry out an integration over the $2N$-dimensional configuration space, where N, the total number of particles, was usually very large. A Monte Carlo method is therefore introduced.

Since their integration has the form

$$\frac{\int t(X)\exp(-E/kT)dX}{\int \exp\{-E/kT\}dX},$$

where E, the energy of the system, is a function of the $2N$-dimensional variable X, the naive Monte Carlo method, which is to simulate N random positions uniformly in the square and

to calculate the energy according to the equation

$$E = \frac{1}{2}\sum V(x(i)).$$

then give $t(X)$ a weight $\exp\{-E/kT\}$, is quite inefficient. This led them to think about simulating X directly from the probability distribution

$$\pi(X) \propto \exp\{-E/kT\},$$

and use the law of large number, or more precisely, the ergodic theorem for the Markov chain.

The method, which is later named the "Metropolis algorithm" in statistical physics, is related to rejection techniques since it involves explicitly proposing a tentative value which may be rejected. Because the rejection rule only needs the ratio between the two density values of two different configurations, the renormalization constant of the sampling distribution $\pi(X)$ is irrelevant, so that we never need to know it. This property allows us to sample from any distribution known up to a constant.

The Metropolis algorithm is of great simplicity and power; it can be used to sample essentially any density function regardless of analytic complexity in any number of dimensions. Of course, higher dimensionalities may make the algorithm slower. The disadvantages are, just like the Gibbs sampler, that sampling is correct only asymptotically and that successive samples produced are correlated.

Since its invention, the method had mostly been used in theoretical chemistry and statistical mechanics. Later and also quite recently, it was recognized by researchers in image science. In the past two or three years, statisticians began to find its potentials in statistical computations, and extensively apply the method to various seemingly unrelated problems.

The general description of the Metropolis algorithm can be found in Valleau and Whittington [56], Kalos and Whitloc [28]. The method was motivated by an analogy to the behavior of systems in statistical mechanics that approach an equilibrium whose statistical properties are independent of the kinetics of the system. In probability terms, it implies that the system will converge to a stationary process eventually.

Let $X \in \Omega$, where Ω (typically in R^n) be the sample space. $K(Y|X)$ is the transition function that governs the evolution of the system (Markov chain) as before. The Metropolis algorithm prescribes a way of finding a tractable transition function $K(Y|X)$ so that the system will have the desired probability distribution $\pi(X)$, the one we want to simulate samples from, as its equilibrium distribution. The function $K(Y|X)$ is also required to satisfy the "detailed balance" condition

$$K(X|Y)\pi(Y) = K(Y|X)\pi(X)$$

so that the induced Markov chain is reversible. The condition is needed because it automatically implies that $\pi(X)$ is invariant under the transition $K(Y|X)$, and is therefore a possible equilibrium distribution of the chain. This point is easily seen because

$$\pi(Y) = \pi(Y)\int K(X|Y)dX = \int \pi(Y)K(X|Y)dX = \int \pi(X)K(Y|X)dX$$

To find such a proper transition function $K(Y|X)$, a tentative initial one is proposed from, say, X to Y', using essentially any distribution $T(Y'|X)$. (Therefore we can make it easy enough to meet our capacity.) Then on comparing $\pi(Y')$ with $\pi(X)$, and taking into account T as well, the system is either moved to Y' (accepted). or returned to X (rejected).

Acceptance of the move occurs with probability $A(Y'|X)$. Consequently we have

$$K(Y|X) = \begin{cases} A(Y|X)T(Y|X) & \text{if } Y \neq X, \\ 1 - \int_{Y' \neq X} A(Y'|X)T(Y'|X)dY' & \text{if } Y = X, \end{cases}$$

$A(Y|X)$ has to be chosen so that $K(Y|X)$ satisfies the detailed balance condition, i.e.,

$$A(X|Y)T(X|Y)\pi(Y) = A(Y|X)T(Y|X)\pi(X)$$

If we define

$$q(Y|X) = \frac{T(X|Y)\pi(Y)}{T(Y|X)\pi(X)},$$

then $q(Y|X) = \frac{A(Y|X)}{A(X|Y)}$.

There will be many ways of choosing $A(Y|X)$. Metropolis [39] suggested using

$$A(Y|X) = \min(1, q(Y|X)).$$

Barker [4] later suggested a more "continuous" acceptance function as follows,

$$A(Y|X) = \frac{q(Y|X)}{1 + q(Y|X)}.$$

Peskun (1973) [42] shows that Baker's prescription is inferior to Metropolis' one in most situations. The heuristic argument is that Metropolis' method allows more transitions, so when using

$$\frac{t(X_1) + t(X_2) + \cdots + t(X_n)}{n}$$

to estimate $Et(X)$, the average is over more independent terms.

In either cases, it can be easily checked that the corresponding Markov chain under the transition $K(Y|X)$ is reversible, so $\pi(X)$ is the invariant function of the transition which, together with other conditions, assures that the asymptotic probability distribution exists and is unique. Therefore $\pi(X)$ will be its asymptotic distribution.

The original proposals of the initial transition function $T(Y|X)$ suggested by both Metropolis et al. [39] and Barker [4] are those "symmetric" ones in the sense that

$$T(Y|X) = T(X|Y), \text{ for all } X \text{ and } Y.$$

In that case, the acceptance probability of each step only depend on the ratio $\frac{\pi(Y)}{\pi(X)}$. Hastings [24] extended such arguments and made the choice of T arbitrary. So the algorithm described here should be more precisely named as the "Metropolis-Hastings algorithm". For convenience, we use the term "Metropolis algorithm" to denote such a class of similar Monte Carlo schemes shown above, including the individual algorithms proposed by Metropolis et al., Barker, and Hastings.

A probabilistic argument to solve the problems of uniqueness of the equilibrium measure and the convergence to it, when the state space is discrete and finite, is quite illustrative. Here we assume $A(Y|X)$ to be the one given by Metropolis et al. Then it is apparent that

$$K(Y|X) > 0 \quad \text{if only if} \quad T(Y|X) > 0.$$

Hence the K-chain is irreducible if and only if the T-chain is irreducible which can be guaranteed by purposely choosing the initial transition $T(Y|X)$ to be so. By the theorems in chapter 15 of Feller [16], the equilibrium distribution of the K-chain is unique because of the irreducibility. Furthermore, if the chain is aperiodic too, the distribution $p_n(X)$ of the n-th iterated sample will converge to $\pi(X)$. Note that the chain is aperiodic if we can find one X_0 such that $K(X_0|X_0) > 0$. This is true for the K-chain except for the trivial case with $\pi(X)$ being uniform on Ω. If it is not true, since $K(X|X) = 0$ implies $T(X|X) = 0$ and also

$$\pi(Y) \geq \pi(X) \quad \text{for all } Y \text{ with } T(Y|X) > 0,$$

we can construct the set of minimum

$$A = \{X : \pi(X) = \min_{Y \in \Omega} \pi(Y)\}.$$

Then T can never move out of A once the chain gets into it. Therefore it contradicts to the assumption that the T-chain is irreducible. Hence the K-chain is aperiodic. (Note that we don't even require that T is aperiodic.) Thus a sufficient condition if π is not constant is to check that if it is possible to move from any state to any other under T.

If we take a view from a new angle, it is possible to analysis the forward operator

$$Ft(X) = E(t(Y)|X) = \int t(Y)A(Y|X)T(Y|X)dY + t(X)(1 - \int A(Y|X)T(Y|X)dY)$$

as well. If we further denote $c(X) = \int A(Y|X)T(Y|X)dY$, which represents the probability of not staying at X in the next step, and the associate operator

$$F_1 t(X) = \frac{1}{c(X)} \int t(Y)A(Y|X)T(Y|X).$$

then F can be further decomposed as

$$Ft(X) = (1 - c(X))t(X) + c(X)F_1 t(X).$$

Therefore, if a function $t(X)$ is an eigenfunction of F_1 corresponding to the eigenvalue 1, it is also an eigenfunction of F with eigenvalue one, and vise versa. We can easily choose the initial transition function $T(Y|X)$ so that F_1 is a compact operator. Furthermore, if we consider the discrete state space of Ω, both F and F_1 are compact operators. If the maximal eigenvalue of F_1 is not unity, then it is also true for F. So the geometric convergence of the associated chain corresponding to F_1 will imply the geometric convergence of the chain corresponding to F. However a general result does not exist because of the complex nature

of the initial choice of transition T, the rejection rule, and the possible non-compact sample space.

The algorithm can be described more verbally: at step n of the evolution (random walk of the Markov chain). the value of X is X_n; a possible next value for X, X'_{n+1}, is sampled from $T(X'_{n+1}|X_n)$, and the probability of accepting X'_{n+1} is computed according to a rejecting rule $A(Y|X)$. With probability $A(X'_{n+1}|X_n)$ we set $X_{n+1} = X'_{n+1}$; otherwise we set $X_{n+1} = X_n$.

Since the procedure can only guarantee to sample from $\pi(X)$ asymptotically, we need to throw away first L steps of the iteration. This value L is usually hard to determine. Heuristically, one may want to choose the starting density $p_0(X)$ and the transition $T(Y|X)$ as close to the true density $\pi(X)$ as possible so as to obtain rapid convergence and small sample correlations.

Some ad hoc methods of deciding when to stop have been suggested. For example, one way is to extract some summary quantities such as mean, variance, histogram, etc., which can characterize the distributions to certain degree, by averaging over (say) every 100 steps and observe the behavior of such quantities. A decision is then made about whether in the random walk the observed values have converged. Careless observation of the attainment of the asymptotic distribution in the Metropolis algorithm has led to some bad Monte Carlo calculation in the past. It is also the case for applications of the Gibbs sampler.

Another problem arises from choosing of the original transition function $T(Y|X)$. If it is badly chosen, most of the initial steps will be rejected, which will result in inefficiency of the algorithm. Some initial rough approximation of the true density $\pi(X)$ may be helpful to resolve the problem. However, in practice a simple uniform distribution of $T(Y|X)$ is

often employed for the sake of simplicity.

Example. Suppose we are going to sample from

$$\pi(X) = \begin{cases} 2X, & X \in (0,1). \\ 0, & \text{otherwise.} \end{cases}$$

We can choose the initial transition function to be uniform:

$$T(Y|X) = \begin{cases} 1, & Y \in (0,1). \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$q(Y|X) = \begin{cases} 1, & Y/X \in (0,1). \\ 0, & \text{otherwise.} \end{cases}$$

Hence we can choose $A(Y|X) = \min\{1, Y/X\}$, so that

$$K(Y|X) = \begin{cases} \min\{1, Y/X\}, & Y \neq X, \\ 1 - X/2, & Y = X. \end{cases}$$

If we use $\phi_n(X)$ to denote the density function after n-th iteration, the recursive equation can be derived,

$$\phi_{n+1}(X) = \int_X^1 \frac{X}{Y}\phi_n(Y)dY + \int_0^X \phi_n(Y)dY + \frac{X\phi_n(X)}{2}.$$

□

Recently, there are several articles on estimating the rate of convergence for the Markov chain with finite discrete state space, by bounding the second largest eigenvalue of the corresponding transition operator with a geometric method. The basic tools used are Poincaré inequalities and Cheeger's inequality. The so-called Poincaré inequality is a discrete analog of the classical method of Poincaré for estimating the spectral gap of the Laplacian on a

domain (see, e.g., Bandle [3]). A quite thorough overview of such methodology can be found in Diaconis [11], Diaconis and Stroock [12], and Fill [17]. Some stimulating results for solving a problem in computer science are derived by Sinclair and Jerrum[52] from using such methods. Sinclair [51] also used these inequalities to get bounds in the approach to equilibrium for using Metropolis algorithm to simulate Ising Models. Ingrassia [25] has used the techniques to get bounds on the rate of convergence in simulated annealing.

6.2 The Gibbs Sampler

In 1984, Geman and Geman [21] introduced an algorithm combining both stochastic relaxation and an annealing procedure for computing the maximum a posteriori (MAP) estimate of a large stochastic system used to model a true image. The relaxation step, which can be viewed as a variation of the old Metropolis algorithm, the so-called "the Gibbs sampler", is named after the great physicist and mathematician of the last century, Josiah Williard Gibbs, who is apparently not quite responsible for the method. This is another verification of Stigler's law of eponymy. The reason why Geman and Geman named it so is, perhaps, that they first applied such technique to a system with the Gibbs distribution structure. By a theorem in Besag [5], the Hammersley-Clifford theorem, the system with a Gibbs distribution is equivalent to a Markov random field.

Consider a process that takes one of L values, $v_1, \ldots, v_L$, at each of a set of sites, $S = \{1, \cdots, d\}$. These sites will be related to each other by a graphical structure, for example, sites on a square lattice, so each site has a number of neighbors. The process is specified by giving the joint distributions of all the random variables $(x(s)|x \in S)$. A

Markov random field has the special property that

$$P(x(r) = v_i | x(s).s \neq r) = P(x(r) = v_i | x(s).s \text{ a neighbor of } r).$$

In other words, the conditional distribution at a site given the rest depends only on the values at the neighboring sites. Such processes are thus specified by giving all these conditional distributions with the result that the joint distribution is known only up to a renormalizing constant.

The Gibbs sampler proposed by Geman and Geman makes full use of the special Markov random field structure to model a true image. The transition function $T(Y|X)$ is such created so that each transition can be accepted. Actually the function T proposed by Geman and Geman [21] satisfies

$$T(Y|X)\pi(X) = T(X|Y)\pi(Y).$$

There will be no need for rejection. Later extensions of such technique make it possible that T is not reversible, but still has $\pi(X)$ as its equilibrium distribution, i.e.,

$$\int T(Y|X)\pi(X)dX = \pi(X).$$

Therefore these extensions, for example, Tanner and Wong [54] which is equivalent to a two-dimensional Gibbs sampler using systematic scan, are not entirely equivalent to the original Metropolis algorithm where the detailed balance for a reversible chain is required.

An obvious advantage of such choice of the transition function is that the inefficiency caused by rejections is diminished. However the samples from such a procedure may be more correlated than the ones from the original Metropolis algorithm because in each step of the Gibbs sampler we only allow one variable, $x(i)$ for some $i \in S$, to change. The

"improvement" in each step is therefore restricted to only one direction. An explicit formula for such correlations in the case of the random scan has been derived in Chapter 4, which is always nonnegative for all lags.

Geman and Geman's transition can be specially written as

$$K(Y|X) = T(Y|X) = \begin{cases} \frac{1}{d}\pi(y(i)|X^{-i}). & \text{if } Y^{-i} = X^{-i}, \\ 0 & \text{otherwise,} \end{cases}$$

which was used in their examples. However, their theoretical result of convergence is for more general cases where $K(Y|X)$ is even allowed to be inhomogeneous in the sense that $K(Y|X)$ may change with the evolution of the system, as long as each site is visited infinitely often. The following theorem about the convergence and also the rate for the case of finite discrete state space was provided by the original Geman and Geman [21]. The positivity condition was assumed everywhere. In this special case of the random field, it implies that for any possible one of the L^d configurations of the process, X, its probability assignment is not zero, i.e., $\pi(X) > 0$ for all possible X. The reason why we can't use the result about the Metropolis algorithm directly is that the initial transition T there is the one we choose and needs not be related directly to the equilibrium distribution, but here the transition is determined entirely by the structure of the equilibrium distribution. Therefore the convergence results need to be renewed.

Theorem 6.2.1 (Geman and Geman) *Assume that for each $i \in S = \{1, \ldots, d\}$, the sequence $\{i_t, t \geq 1\}$, in which t represents the number of iteration, contains i infinitely often. Then under the positivity condition, for every starting configuration $\eta \in \Omega$ and any $\omega \in \Omega$,*

$$\lim_{t \to \infty} P(X_t = \omega | X_0 = \eta) = \pi(\omega).$$

To prove the theorem, some lemmas are needed. Some settings are useful too. Let T_1 be the shortest time that all sites are visited, i.e.,

$$T_1 = \min\{t : S \subset \{i_1, \ldots, i_t\}\}$$

Recursively, we can define

$$T_k = \min\{t > T_{k-1} : S \subset \{i_{T_{k-1}+1}, \ldots, i_t\}\}.$$

Also we define $k(t) = \sup\{k : T_k < t\}$, which goes to infinity as t goes to infinity.

Lemma 6.2.1 *For any ω and η'', we will have*

$$P(X_{T_k} = \omega | X_{T_{k-1}} = \eta'') \geq \delta^d$$

for all k, where $\delta = min(P(x(i)|X^{-i}))$.

PROOF: Let $m_i = \sup\{t : t \leq T_k, i_t = i\}$, for $i = 1, \cdots, d$, the latest time for the site i to be drawn in the period (T_{k-1}, T_k). Without loss of generality, we may assume that $m_1 > \cdots > m_d$. So

$$\begin{aligned}
& P(X_{T_k} = \omega |\ X_{T_{k-1}} = \eta'') \\
= \ & P(x_{m_1}(1) = \omega(1). \ldots, x_{m_d} = \omega(d) | X_{T_{k-1}} = \eta'') \\
= \ & \prod_{j=1}^{d} P(x_{m_j}(j) = \omega(j) | x_{m_{j+1}}(j+1) = \omega(j+1). \ldots, x_{m_d}(d) = \omega(d). X_{T_{k-1}} = \eta'') \\
\geq \ & \delta^d
\end{aligned}$$

The last inequality follows directly from the Markov property of the chain X_t. □

Lemma 6.2.2 *There exists a constant r, $0 \leq r < 1$, such that for every $k = 1, 2, \ldots,$*

$$\sup_{\omega, \eta', \eta''} |P(X_{T_k} = \omega | X_{T_{k-1}} = \eta') - P(X_{T_k} = \omega | X_{T_{k-1}} = \eta'')| \leq r$$

holds for any ω, η' and η".

PROOF:

$$\begin{aligned} & \sup_{\omega,\eta',\eta''} |P(X_{T_k} = \omega | X_{T_{k-1}} = \eta') - P(X_{T_k} = \omega | X_{T_{k-1}} = \eta'')| \\ = & \sup_{\omega} \{ \sup_{\eta} P(X_t = \omega | X_0 = \eta) - \inf_{\eta} P(X_t = \omega | X_0 = \eta) \} \\ = & \sup_{\omega} \{ I - II \} \end{aligned}$$

where

$$\begin{aligned} I &= \sup_{\eta} \sum_{\omega'} P(X_t = \omega | X_{T_1} = \omega') P(X_{T_1} = \omega' | X_0 = \eta) \\ &\leq \sup_{\mu(\omega') \geq \delta^d} \sum_{\omega'} P(X_t = \omega | X_{T_1} = \omega') \mu(\omega') \end{aligned}$$

in which the last supremum is taken over all possible probability measure with the restriction that $\mu(\omega') \geq \delta^d$ due to lemma 6.2.1. Now suppose $P(X_t = \omega | X_{T_1} = \omega')$, for a fixed ω, attains its maximum at $\omega' = \omega^o$, i.e.,

$$P(X_t = \omega | X_{T_1} = \omega^o) = \sup_{\omega'} P(X_t = \omega | X_{T_1} = \omega').$$

Then the best possible probability measure μ for the quantity I is to assign $1 - (L^d - 1)\delta^d$ on ω_0, the rest for δ^d each. (Recall that $|\Omega|$ is equal to L^d, which is the total number of possible configurations.) Hence

$$I \leq (1 - (L^d - 1)\delta^d) P(X_t = \omega | X_{T_1} = \omega^o) + \delta^d \sum_{\omega' \neq \omega^0} P(X_t = \omega | X_{T_1} = \omega').$$

Similarly, we can find the minimum possible value of II as

$$II \geq (1 - (L^d - 1)\delta^d) P(X_t = \omega | X_{T_1} = \omega_o) + \delta^d \sum_{\omega' \neq \omega_o} P(X_t = \omega | X_{T_1} = \omega').$$

where $P(X_t = \omega | X_{T_1} = \omega_o) = \inf_{\omega'} P(X_t = \omega | X_{T_1} = \omega')$. Therefore

$$I - II \leq (1 - L^d \delta^d)[P(X_t = \omega | X_{T_1} = \omega^o) - P(X_t = \omega | X_{T_1} = \omega_o)]$$

The result follows inductively, with the bounding constant $r = 1 - L^d \delta^d$. □

Proof of the theorem:

$$\begin{aligned} & \overline{\lim_{t\to\infty}} \sup_{\omega,\eta} |P(X_t=\omega|X_0=\eta)-\pi(\omega)| \\ = & \overline{\lim_{t\to\infty}} \sup_{\omega,\eta} |\sum_{\eta'}\{\pi(\eta')P(X_t=\omega|X_0=\eta)-P(X_t=\omega|X_0=\eta')\}| \\ = & \overline{\lim_{t\to\infty}} \sup_{\omega,\eta,\eta'} |P(X_t=\omega|X_0=\eta)-P(X_t=\omega|X_0=\eta')|. \end{aligned}$$

We can apply lemma 6.2.2 to get the geometric convergence rate of the scheme with the bound $r^{k(t)}$, where $r=1-L^d\delta^d$, $k(t)=\sup\{k:T_k\le t\}$. □

The above proof illustrates the difficulties of the problem and the limitations of Geman and Geman's approach, in which the discreteness and the positivity condition are relied on too heavily.

When considering the case of the systematic scan where $T_k = kd$, a much simpler argument based on lemma 6.2.1 can be given. Let us consider the embedded Markov chain consists of

$$X_{T_1}, X_{T_2}, \ldots, X_{T_k}, \ldots,$$

which, by lemma 6.2.1, is aperiodic and irreducible. Hence the chain is ergodic. When the scan used is purely random as described in section 4.2, the argument is the same, except that in this case $\{T_k\}$ are identically and independently distributed random variables.

The stochastic relaxation and simulating annealing techniques proposed by Geman and Geman [21] is used for image restoration. The essence of their approach to restoration is a relaxation algorithm which generates a sequence of images that converges in a appropriate sense to the MAP estimate. The stochastic relaxation permits changes that decrease the posterior probability values, and is made on a random basis. The effect of which is to avoid converging to a local maxima. It can also be viewed as an optimization algorithm for a

large system.

Example. Ising Model

Ising model is a special Markov random field in which an $m \times m$ square lattice structure is assumed. Each component of X, i.e., $x(s)$, $s = 1, \ldots, m^2$, takes only two values, -1 or $+1$. The conditional distribution is given by

$$\pi(x(i) = 1 | X^{-i}) = \frac{e^{\eta}}{1 + e^{\eta}}$$

where $\eta = \beta \sum_{j \neq i} x(j)x(i)$, the sum is over all the neighbors of site i, β is a global constant to control the flatness of the distribution. The form of the joint density can be easily written down, up to a constant ($d = m^2$). as:

$$\pi(x(1)\ldots., x(d)) \propto \exp(\beta \sum_{i \neq j} x(i)x(j))$$

For implementing the Metropolis algorithm, an initial transition function $T(Y|X)$ must be selected. For simplicity we may choose T to be a transition which picks a site at random and changes its value to -1 and $+1$ randomly with probability $1/2$ each. This transition is clearly irreducible, and aperiodic. Therefore the Metropolis algorithm applies. If the acceptance function $A(Y|X)$ is chosen according to Metropolis' method, we end up with a function

$$A(Y|X) = \min(1, \pi(Y)\pi(X)).$$

Such a choice of the acceptance function has been shown to dominate the function $A(Y|X)$ suggested by Barker [4], which in this special case is

$$A(Y|X) = \frac{\pi(Y)}{\pi(X) + \pi(Y)} = \pi(Y|Y^{-i})$$

where Y possibly differs from X only at site i. This is exactly the same as the Gibbs sampler used with a random scan. Therefore by choosing different $A(Y|X)$, we can derive a

Gibbs sampler scheme directly from a Metropolis algorithm. Also, in this case, the original Metropolis algorithm dominates the Gibbs sampler. □

6.3 Applications in statistics

The applicability of the Metropolis algorithm, the Gibbs sampler, and their variations is due to the rapid growth of the computer facilities. It is also the case for many other modern developments of statistics, for example, bootstrap, ACE algorithm, EM algorithm, to name a few. The first formal application of the Gibbs sampler technique to the computation of posterior densities should be ascribed to Tanner and Wong [54], though concurrently Li [34] has applied the method to the imputation of missing data.

Tanner and Wong [54]'s work focuses on the missing data problem setting from Bayesian point of view, and is also applicable to many others which can be treated as missing data problems. The basic assumption for their method to work, which is similar to the basic requirement for the EM algorithm, is that the complete data posterior is easy to deal with, i.e., to be able to draw sample from. Suppose we use θ or ϕ to denote the parameter of interests; $x = (y, z)$ to denote the complete data in which y represents the observed part, while z the missing part; and π to stand for various kinds of distributions derived from the true joint distribution. The basic algorithm, which is later referred as "data augmentation," involves two steps in general:

(a) Generate samples $z^{(1)}, z^{(2)}, \ldots, z^{(m)}$ from the current approximation to the predictive density $\pi(z|y)$.

(b) Update the current approximation to $\pi(\theta|y)$ to be the mixture of conditional densities of θ given the m sets of the augmented complete data with the form $(y, z^{(k)})$, $k = 1, \ldots, m$, where $z^{(k)}$ are generated from step (a). that is

$$g_{i+1}(\theta) = \frac{1}{m} \sum_{k=1}^{m} \pi(\theta|(y, z^{(k)})).$$

To implement step (a). there will be two more steps:

(a1) generate θ from $g_i(\theta)$,

(a2) generate z from $\pi(z|\theta, y)$, where the θ is from step (a1).

If the value of m is taken to be one, which implies that there is no repeated sampling at each iteration, we boil down to a simpler version of the algorithm by combining steps (a1). (a2). and (b):

(1) Draw z_i from $\pi(z|\theta_i, y)$, where θ_i is from previous iteration,

(2) Draw θ_{i+1} from $\pi(\theta|z_i, y)$, where z_i is from step (1).

This is exactly the same as a 2-dimensional Gibbs sampler used with the systematic scan. When Tanner and Wong [54] first invented such an algorithm they didn't realize the connection between the two. It only became clear later on. Until quite recently, the paper by Gelfand and Smith [19] formally explores the relations between the Gibbs sampler, data augmentation, importance sampling etc., and uses the name "sampling-based techniques" as a summary. The later version of the data augmentation scheme (i.e., steps (1). (2).) may works more efficiently in lots of cases than the original one because, as was also pointed out by Tanner and Wong [54] (section 7). the first few iterations are far from the truth anyway,

it is not worth being accurate, i.e., simulating a large amount of samples, for the beginning iterations.

To theoretically prove the convergence of their scheme, Tanner and Wong adopted an elegant argument using L^1 theory on the transition operator. The condition they require is called condition (C) in their paper. To distinguish it from our condition (C). we will call that the C-Condition. Let

$$K(\theta|\phi) = \int \pi(\theta|y,z)\pi(z|\phi,y)dz,$$

which is the same as what we have defined as the transition function of the marginal chain in section 3.1.

C-Condition: $K(\theta|\phi)$ is uniformly bounded and is equicontinuous in θ. For any $\theta_0 \in \Theta$, there is an open neighborhood U of θ_0, so that $K(\theta,\phi) > 0$ for all θ, $\phi \in U$.

This condition is quite strong and is usually hard to check in practice. As for the method of proof, they are using L^1 theory on the operator defined as

$$Tf(\theta) = \int K(\theta|\phi)f(\phi)d\phi$$

which transit a density function to another one. Obviously, the true posterior $\pi(\theta|y)$ is a fixed point of such transition, or in other words, the eigenvalue of the operator T because, if we use $g_*(\theta) = \pi(\theta|y)$,

$$\begin{aligned} Tg_*(\theta) &= \int K(\theta|\phi)g_*(\phi)d\phi \\ &= \int\int \pi(\theta|y,z)\pi(z,|\phi,y)\pi(\phi|y)dx\,d\phi \\ &= \int\{\int \pi(z,|\phi,y)\pi(\phi|y)d\phi\}\pi(\theta|y,z)dz \\ &= \pi(\theta|y) \end{aligned}$$

L^1-theory is relatively more awkward than L^2 theory, which creates obstacles to the theoretical result of Tanner and Wong [54]. In fact, the proof of theorem 3 in Tanner and Wong [54], which claims a geometric convergence rate for the scheme, i.e.,

$$\| g_i - g_* \| \leq \alpha^i \| g_0 - g_* \|$$

contains a mistake at step (e). The question remains an unsettled issue for quite long, until our latest work (Liu et al. [35] [36],) and independently the work of Schervish and Carlin [50]. The new result in Schervish and Carlin [50] is more general than required by Tanner and Wong and is applicable to the general Gibbs sampler used with a systematic scan. The method can be viewed as a substantial extension of the original Tanner and Wong [54]'s argument. Instead of applying L^1 theory on the operator T as was done by Tanner and Wong, Schervish and Carlin adopted an analysis on the operator $U = T^*T$, where T^* is the adjoint operator of T, using L^2-theory. Such an operator U is self-adjoint, and compact under certain condition. Although the largest eigenvalue of U is still unity with the eigenfunction $\pi(\theta|y)$, they proved that the second largest eigenvalue of U is strictly less than one when the positivity condition holds. This, together with an assumption on the starting density which is the same as our condition (A) in chapter 1, gives the geometric convergence of the scheme. Our approach, as has been shown in previous chapters, is more direct and elegant by dealing with the forward operator which is slightly different from the traditional transition operator T. Other convergence results concerning Gaussian models can be found in the works of Knoerr [30], Amit and Grenander [2], Goodman and Sokal [23].

A huge amount of literature on applying the Gibbs sampler and data augmentation techniques have emerged recently since the papers of Tanner and Wong [54] and Gelfand and

Smith [19]. Its conceptual simplicity and practical ease of implement have been recognized and appreciated by more and more statisticians. The stochastic relaxation techniques open a new page for solving hard statistical computation problems. To illustrate, I name a few areas where such techniques have been actively used. For example, one can read about the inference for normal models by using the Gibbs sampler from Gelfand et al. [20]; genetic linkage analysis from Kong et al. [31], Geyer and Thompson [22], and Thompson et al. [55]; hierarchical models, variance components and errors-in-variables in Gelfand and Smith [19]; latent-class model in original Tanner and Wong [54]; outlier detection in Verdinelli and Wasserman [57]; level-changing problem in autoregressive time series model in McCulloch and Tsay [38]; and even non-parametric Bayesian analysis using Dirichlet Process prior in Escobar [15].

The boom of the area may lead people to think that the Gibbs sampler is omnipotent. However this is not true. In applications, it is possible to get a nonergodic chain for the iteration; it is also possible for individuals to misjudge the convergence of the algorithm. There are also many issues like computing time, choosing proper variables to iterate, setting good starting values etc., which are all very important. The technique is not fool-proof. It was invented to allow us to solve harder problems. To conclude, I would say that I warmly welcome the fully appreciation and application of the Gibbs sampler, but one should not apply it blindly.

BIBLIOGRAPHY

[1] Amit, Y. (1990). On rates of convergence of stochastic relaxation for Gaussian and non-Gaussian distributions, *Technical report*, Division of Applied Mathematics, Brown University.

[2] Amit, Y. and Grenander, U. (1989). Comparing sweep strategies for stochastic relaxation, *Technical report*, Division of Applied Mathematics, Brown University.

[3] Bandle, C. (1980). *Isoperimetric inequalities and Applications*, Pitman, Boston.

[4] Barker, A.A. (1965). Monte Carlo calculations of radial distribution functions for a proton-electron plasma. *Aust. J. Physics.* **18**, 119-133.

[5] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with Discussion). *J. R. Statistics Soc.* **B, 36**, 192-236.

[6] Box, G.E.P. and Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis*, Addison-Wesley Publishing Company, Reading, Massachusetts.

[7] Bradley, R. (1986). Basic properties of strong mixing conditions, *Dependence in Probability and Statistics*, Birkhäuser, Boston, 165-192.

[8] Breiman, L. and Friedman, J.H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *J. of Amer. Stat. Assoc.*, **80**, 580-619.

[9] Csàki, P. and Fischer, J.H.(1960). Contributions to the problem of maximal correlation, *Matematikao Kotato Intezet, Kozlemenyei*, **5**, 325-337.

[10] Dempster, A.P. and Laird, N.M. and Rubin, D.B. (1977). Maximum Likelihood From Incomplete Data Via the EM Algorithm, *Journal of the Royal Statistical Society*, Ser. B, **39**, 1-38.

[11] Diaconis, Persi (1988). *Group representations in probability and statistics.* IMS, Hayward, Calif.

[12] Diaconis, P. and Stroock, D. (1991). Geometric bounds for eigenvalues of Markov chains, *The Annals of Applied Probability*, **1**, 36-61.

[13] Dunford, N. and Schwartz, J.T. (1963). *Linear Operators, Part I*, Interscience, New York.

[14] Eaton, M.L. (1976). A maximization problem and its application to canonical correlation, *I. Multivariate Anal.*, **6**, 422-425.

[15] Escobar, M.D. (1991). Estimating normal means with a Dirichlet process prior, *Technical report No. 512*, Dept. of Statistics, Carnegie Mellon University.

[16] Feller, W. (1971). *An Introduction to Probability Theory and its Applications, Vol. 1*, second edition. John Wiley & Sons, New York.

[17] Fill, James Allen (1991). Eigenvalue bounds on convergence to stationarity for nonreversible Markov chains, with an application to the exclusion process, *The Annals of Applied Probability*, **1**, 62-87.

[18] Gebelein, H. (1947). Cited by Breiman, L. and Friedman, J.H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *J. of Amer. Stat. Assoc.*, **80**, 580-619.

[19] Gelfand, A.E. and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities, *J. of Amer. Stat. Assoc.*, **85**, 398-409.

[20] Gelfand,A.E. and Hills, S.E. and Racine-Poon, S. and Smith, A.F.M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling, *J. of Amer. Stat. Assoc.* **85**, 972-985.

[21] Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **6**,721-741.

[22] Geyer, G.J. and Thompson, E.A. (1990). Three papers on maximum likelihood in exponential families, *Technical report No. 188*, Dept. of Statistics, University of Washington, Seattle.

[23] Goodman, J. and Sokal, A.D. (1989). Multigrid Monte Carlo method, conceptual foundations, *Physical Review D*, **40** no. 6, 2035-2071.

[24] Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications, *Biometrika*, **57**, 97-109.

[25] Ingrassia, S. (1990). Ph.D. dissertation. Cited by Diaconis, P. and Stroock, D. (1991). Geometric bounds for eigenvalues of Markov chains, *The Annals of Applied Probability*, **1**, 36-61.

[26] Johnson, V. (1989) On statistical image reconstruction, *Ph.D. thesis*, Dept. of Statistics, University of Chicago.

[27] Kagan, A.M. and Linnik, Y.V. and Rao, C.R. (1973). *Characterization Problems in Mathematical Statistics*, John Wiley & Sons, New York.

[28] Kalos, H.M. and Whitloc, P.A. (1986). *Monte Carlo Methods, Vol.1*, John Wiley & Sons, New York.

[29] Kendall, M.A. and Stuart, A. (1967). *The Advanced Theory of Statistics, Vol.2*, Hafner Publishing, New York.

[30] Knoerr, A.P. (1988). Global models of natural boundaries theory and applications, Ph.D. thesis, Division of Applied Mathematics, Brown University.

[31] Kong, A. and Wong, W. and Frigge, M. and Cox, N. (1990). *A sampling-based method for linkage analysis with multiple parameters*. In preparation.

[32] Lancaster, H.O. (1957). Some properties of the bivariate normal distribution considered in the form of a contingency table, *Biometrika*, **44**, 289-292.

[33] — — (1958). The structure of Bivariate Distributions, *Annals of Mathematical Statistics*, **29**, 719-736.

[34] Li, Kim-Hung (1988). Imputation using Markov chains, *Journal of Statistical Computation and Simulation* **30** 57-79.

[35] Liu, Jun and Wong, W.H. and Kong, A. (1991). Correlation structure and convergence rate of the Gibbs sampler (I): applications to the comparisons of estimators and augmentation schemes, *Technical report No. 299*, Dept. of Statistics, University of Chicago.

[36] Liu, Jun and Wong, W.H. and Kong, A. (1991). Correlation structure and convergence rate of the Gibbs sampler (II): applications to various scans, *Technical report No. 304*, Dept. of Statistics, University of Chicago.

[37] Maung, K. (1941). Measurement of association in a contingency table with special reference to the pigmentation of hair and eye colours of Scottish school children. *Ann. Eugen., London,* **11**, 189.

[38] McCulloch, R.E. and Tsay, R.S. (1991). Bayesian analysis of autoregressive time series via the Gibbs sampler, *Technical report*, Graduate School of Business, University of Chicago.

[39] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953). Equations of state Calculations by fast computing machines, *J. Chem. Phys.*, **21**, 1087-1091.

[40] Muirhead, R.J. (1982). *Aspects of Multivariate Statistical Theory*, John Wiley & Sons, New York.

[41] Murray, G.D. (1977). Comment on "Maximum Likelihood From Incomplete Data Via the EM Algorithm" by A.P. Dempster, N.M. Laird, and D.B. Rubin, *Journal of the Royal Statistical Society*, Ser. B, **39**, 27-28.

[42] Peskun, P.H. (1973). Optimal Monte-Carlo sampling using Markov chains. *Biometrika* **60**, 607-612.

[43] Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*, John Wiley & Sons, New York.

[44] Renyi, A. (1959). On measure of dependence, *Acta Mathematica Academiae Scientiarum Hungaricae*, **10**, 441-451.

[45] Riesz, F. and Nagy, B. (1952). *Lecons d'Analyse Fonctionelle*, Akad. Kiado, Budapest.

[46] Ripley, B.D. (1987). *Stochastic simulation.* John Wiley & Sons, New York.

[47] Rosenblatt, M. (1971). *Markov process, structure and asymptotic behavior*, Springer-Verlag, New York.

[48] Rudin, W. (1973). *Functional Analysis*, McGraw-Hill, New York.

[49] Sarmanov, O.V. (1958). The maximal correlation coefficient, *Doklady Akademii Nauk USSR*, **120**, 715-718.

[50] Schervish, M.J. and Carlin, B.P. (1990). On the convergence of Successive Substitution Sampling, *Technical Report No. 492*, Department of Statistics, Carnegie-Mellon University.

[51] Sinclair, A. (1990). Improved bounds for mixing rates of Markov chains on combinatorial structures. Cited by Diaconis, P. and Stroock, D. (1991). Geometric bounds for eigenvalues of Markov chains, *The Annals of Applied Probability*, **1**, 36-61.

[52] Sinclair, A. and Jerrum, M. (1989). Approximate counting, uniform generation and rapidly mixing Markov chains, *Inform. and Comput.* **82**, 93-133.

[53] Stigler, S. (1980). Stigler's law of eponymy, *Trans. of the New York Academy of Sciences*, Series II, **39**, 147-158.

[54] Tanner, M.A. and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. of Amer. Stat. Assoc.*, **52**, 528-550.

[55] Thompson, E. (1991). Stochastic simulation for complex genetic analyses. Presented in the workshop of Bayesian computation via stochastic simulation, Feb. 15-17, 1991, Ohio State University.

[56] Valleau, J.P. and Whittington, S.G. (1976). A guide to Monte Carlo for statistical Mechanics: 1. Highways, 2. Byways, in *Statistical Mechanics*, Part A: Equilibrium Techniques, Modern Theoretical Chemistry Series, Vol. 5, B. Berne, Ed., Chap. 4, 5. Plenum, New York.

[57] Verdinelli, I. and Wasserman, L. (1991). Bayesian analysis of outlier problems using the Gibbs sampler, *Statistics and Computing (in press)*.

[58] Wood, W.W. and Parker, F.R. (1957). Monte Carlo equation of state of molecules interacting with the Lennard-Jones potential. I. Supercritical isotherm at about twice the critical temperature, *J. Chem. Phys.* **27**, 720.

[59] Yosida, K. (1978). *Functional Analysis*, Springer-Verlag, New York.