The Collapsed Gibbs Sampler in Bayesian Computations With Applications to a Gene Regulation Problem

Jun S. Liu*

This article describes a method of "grouping" and "collapsing" in using the Gibbs sampler and proves from an operator theory viewpoint that the method is in general beneficial. The norms of the *forward operators* associated with the corresponding nonreversible Markov chains are used to discriminate among different simulation schemes. When applied to Bayesian missing data problems, the idea of collapsing suggests skipping the steps of sampling parameter(s) values in standard data augmentation. By doing this, we obtain a predictive update version of the Gibbs sampler. A procedure of calculating the posterior odds ratio via the collapsed Gibbs sampler when incomplete observations are involved is presented. As an illustration of possible applications, three examples, along with a Bayesian treatment for identifying common protein binding sites in unaligned DNA sequences, are provided.

KEY WORDS: Data augmentation; Importance sampling; Markov chain; Metropolis algorithm; Missing data; Maximal correlation; Predictive distribution.

1. INTRODUCTION

Integrating out irrelevant (nuisance) parameters is a standard practice in conducting Bayesian inference. In this article we call this integration the "collapsing-down" procedure for reducing random components. The idea can also be carried through in Monte Carlo computations. In applying importance sampling, the collapsing-down idea is known as the dimension-reduction technique (Rubinstein 1981, sec. 4.3.7). In Markov chain Monte Carlo, especially the Gibbs sampler, similar treatments have also been applied explicitly or implicitly by many authors.

The Gibbs sampler is an iterative simulation scheme for generating samples that converge to draws from a target distribution $\pi(X)$ of a random variable X. To simplify notation, all marginal or conditional distributions derivable from the target density are denoted by $\pi(\ \ )$ or $\pi(\,|\,)$ throughout the article. The basic idea of the Gibbs sampler is to construct a Markov chain with $\pi(X)$ as its equilibrium distribution. For example, let $X = (x_1, x_2, x_3)$ be a random variable with three components. The Gibbs sampler is easy to implement when the set of three conditional distributions $\pi(x_i \mid X_{[-i]})$, $i = 1, 2, 3$, are easy to draw samples from, where $X_{[-i]}$ denotes $\{x_j,\ j \neq i\}$. The chain is initiated by a draw from some starting density $p_0(X)$ (or a fixed point); then each component x_i is visited and updated by a sample drawn from the conditional distribution $\pi(x_i \mid X_{[-i]})$. The most widely used visiting scheme, for example, is a systematic one that visits each variate in turn. Detailed descriptions and extensive discussions have been provided by Gelfand and Smith (1990), Smith and Roberts (1993), and others.

Now suppose that we are able to draw x_2 and x_3 together from the conditional distribution $\pi(x_2, x_3 \mid x_1)$. This would be true if we are able to draw x_2 from $\pi(x_2 \mid x_1)$ and then draw x_3 from $\pi(x_3 \mid x_1, x_2)$. Should we use the original Gibbs sampler or a modified version obtained by grouping x_2 and x_3 together? The latter procedure is referred as *grouping* (or *blocking*). Furthermore, suppose that we can draw x_1 directly from $\pi(x_1 \mid x_2)$ and x_2 directly from $\pi(x_2 \mid x_1)$ (i.e., with x_3 integrated out); then the Gibbs sampler can be applied directly to (x_1, x_2). After the chain converges, the third component x_3 can be drawn from $\pi(x_3 \mid x_1, x_2)$. This procedure is called *collapsing*. Is this new sampler even better? Liu, Wong, and Kong (1994) compared such schemes with no more than three components based on operator theory. In this article, we extend their arguments to the general Gibbs sampler and apply the ideas especially to Bayesian computations.

Treatment similar to collapsing was implicitly used by Escobar (1994) to skip the step of sampling an infinite dimensional parameter involved in a nonparametric Bayesian problem. Rubin and Schafer (1990) presented the proposal for effectively producing multiple imputations for multivariate normal data with missingness as a special usage of the grouping and collapsing methodology. Besag (1974)'s coding set method is a primitive version of grouping in spatial statistics. Although in some particular settings, carefully choosing an auxiliary variable to augment can help to speed up the convergence (see, for example, Besag and Green 1993 and Swendsen and Wang 1987), we believe that mere addition of extra variables with no other changes slows convergence and increases sample autocovariances. This article provides numerical as well as theoretical evidences to support our belief.

The article is organized as follows. Section 2 concentrates on Bayesian missing data problems. Its first part presents the general procedure for collapsing in data augmentation and discusses some advantages of doing it. The latter part addresses a problem of calculating the posterior odds ratio when null space Ω_0 is degenerate with respect to the alternative space Ω_1. Section 3 gives three examples to illustrate and support the ideas. Section 4 provides the theoretical justification for the general methodology, and Section 5 proposes a collapsed Gibbs-Metropolis algorithm for a DNA

* Jun S. Liu is Assistant Professor, Department of Statistics, Harvard University, Cambridge, MA 02138. The author thanks Ye Ding for suggesting the DNA problem, Charles Lawrence for insightful discussions, and Hal Stern, Ralph D'agostino Jr., and Andrew Neuwald for a critical reading of the paper and many helpful suggestions. Two referees and an associate editor made many valuable suggestions that helped improve this article greatly.

sequence alignment problem. The Appendix presents theoretical proofs.

2. BAYESIAN MISSING DATA PROBLEMS

A standard Bayesian problem is usually formulated as follows. Let θ be the parameter of interest, and let $X = \{x_1, \ldots, x_n\}$ be a set of complete iid observations from a density that depends on θ: $\pi(x|\theta)$. Here the iid assumption is not crucial for implementing our method but will be useful to keep the arguments simple. A prior distribution $\pi(\theta)$ is incorporated, and the inference is based on the posterior distribution of θ calculated from the Bayes theorem,

$$\pi(\theta|X) = \prod_{i=1}^{n} \pi(x_i|\theta)\pi(\theta)/\pi(X),$$

where $\pi(X) = \int \prod_{i=1}^{n} \pi(x_i|\theta)\pi(\theta)\,d\theta$ is the marginal density of X.

For a future observation x_{n+1}, its predictive distribution after observing $x_1, \ldots, x_n$ can be easily computed as

$$\pi(x_{n+1}|X) = \int \pi(x_{n+1}|\theta) \prod_{i=1}^{n} \pi(x_i|\theta)\pi(\theta)/\pi(X)\,d\theta$$

$$= \pi(\{X, x_{n+1}\})/\pi(X) \qquad (1)$$

(see, for example, Aitchison and Dunsmore 1975). When the model is in an exponential family with a conjugate prior, the posterior distribution typically has a nice form. Furthermore, if the family has a quadratic variance function, then the Bayesian predictive distribution can be written out explicitly (Morris 1983). For example, in multivariate normal problems, the predictive distribution is usually a multivariate t distribution. In multinomial problems with Dirichlet priors, the predictive distribution is again multinomial.

In many practical situations, however, x_i may not be completely observed. Let us assume that the unobserved values are *missing completely at random* (Little and Rubin 1987). Let $x_i = (y_i, z_i)$, $i = 1, \ldots, n$, where y_i is the observed part, and z_i is the missing part. We also write $X = (Y, Z)$, where $Y = (y_1, \ldots, y_n)$ and $Z = (z_1, \ldots, z_n)$. Based on a simple formula,

$$\pi(\theta|Y) = \int \pi(\theta|Y, Z)\pi(Z|Y)\,dZ, \qquad (2)$$

the idea of multiple imputations can be applied to deal with missingness. That is, multiple values, $Z^{(1)}, \ldots, Z^{(m)}$ are drawn from $\pi(Z|Y)$ so as to form m complete data sets. With these imputed data sets and the ergodicity theorem, we can approximate the posterior distribution of θ by a mixture of the complete data posterior distributions; that is,

$$\pi(\theta|Y) \approx \frac{1}{m}\{\pi(\theta|Y, Z^{(1)}) + \cdots + \pi(\theta|Y, Z^{(m)})\}.$$

Many related computations are then simplified.

A difficulty of doing this "exact imputation," however, is that in most applied problems it is impossible to draw Z from $\pi(\cdot|Y)$ directly. Tanner and Wong (1987)'s data augmentation (DA), which can regarded as a two-component Gibbs sampler applied to draw multiples of θ's and multiples of Z's jointly from $\pi(\theta, Z|Y)$, manages to cope with the difficulty by evolving a Markov chain. Here we point out that the standard DA method can be improved by collapsing down the parameter θ in many Bayesian missing data problems.

2.1 Collapsing in Data Augmentation

We assume that with complete observations $X = (Y, Z)$, the posterior distribution of θ is easy to compute or to draw samples from. By iterating between drawing θ from $\pi(\theta|Y, Z)$ and drawing Z from $\pi(Z|\theta, Y)$, DA constructs a Markov chain whose equilibrium distribution is $\pi(\theta, Z|Y)$ (see Gelfand and Smith 1990 and Tanner and Wong 1987 for details).

If we treat Z as a random variable with n components instead of one [i.e., $Z = (z_1, \ldots, z_n)$], we find that the standard DA procedure is also equivalent to the general Gibbs sampler applied to a random variable with $n + 1$ components: $\{\theta, z_1, \ldots, z_n\}$. This fact is made clear by noting that the step of drawing Z from $\pi(Z|\theta, Y)$ is the same as the n steps of drawing z_i from $\pi(z_i|\theta, Z_{[-i]}, Y) = \pi(z_i|\theta, Y)$, $i = 1, \ldots, n$, because of the conditional independence between any z_i and z_j for a given θ. Furthermore, we note that the parameter θ can be collapsed down in the foregoing procedure, and consequently each z_i can be drawn from its *Bayesian predictive distribution* conditioned on the current value of $Z_{[-i]}$. In doing this, we collapse a $(n+1)$-component Gibbs sampler to a n-component one. More precisely, because

$$\pi(Z, \theta|Y) \propto \pi(Y, Z|\theta)\pi(\theta),$$

we have

$$\pi(Z|Y) \propto \int \pi(Y, Z|\theta)\pi(\theta)\,d\theta = \pi(Y, Z),$$

where $\pi(Y, Z)$ is the marginal density for the augmented complete data. Without loss of generality, suppose that we want to draw z_n conditioned on $Z_{[-n]}$. Then, as displayed in (1), the complete data predictive distribution $\pi(x_n|x_1, \ldots, x_{n-1})$ usually has a nice explicit form, and consequently the conditional predictive distribution of z_n,

$$\pi(z_n|Z_{[-n]}, Y) \propto \pi(Y, Z_{[-n]}, z_n) \propto \pi(y_n, z_n|X_{[-n]}),$$

is often easy. This implies that conditioned on the imputed values of $z_1, \ldots, z_{n-1}$ and the observed value Y, the value of z_n can be easily updated by a draw from its conditional predictive distribution. Hence we can go through $z_1, \ldots, z_n$, updating the corresponding predictive distributions and then drawing from them to finish one iteration of this *predictive update* version of the Gibbs sampler.

There are two advantages of collapsing down θ. Within each iteration, the time-consuming step of drawing θ from $\pi(\theta|Y, Z)$ is skipped; between iterations, the sample autocorrelations are usually reduced. Examples in Section 3 show numerically how this method is an improvement over the standard DA procedure. Theoretical justifications are deferred to Section 4.

2.2 Computing Posterior Odds Ratios via Collapsed Gibbs Sampling

Let $X = (Y, Z)$ follow from $\pi(X|\theta)$, where Y is the observed part and Z is the missing part. Let Ω be a general space for θ. We want to do a Bayesian test on the hypothesis H_0: $\theta \in \Omega_0$, where Ω_0 is degenerate with respect to Ω, versus H_A: $\theta \notin \Omega_0$. Let $\pi_1(\theta)$ be a density function on the general space Ω, and let $\pi_0(\theta)$ be a density function on the degenerated space Ω_0. A Bayesian usually uses a prior distribution of the form

$$\alpha\pi_0(\theta)\delta_{\Omega_0}(\theta) + (1 - \alpha)\pi_1(\theta), \tag{3}$$

where $\alpha/(1 - \alpha)$ is the prior odds ratio of H_0 to H_A and δ_{Ω_0} is the Dirac delta function indicating that we put mass 1 on Ω_0. Then the *posterior odds ratio*, defined as

$$r = \frac{\alpha\pi_0(Y)}{(1 - \alpha)\pi_1(Y)} = \left(\frac{\alpha}{1 - \alpha}\right)\frac{\int_{\Omega_0}\pi(Y|\theta)\pi_0(\theta)\,d\theta}{\int_{\Omega}\pi(Y|\theta)\pi_1(\theta)\,d\theta},$$

is of interest for testing purposes. Without loss of generality, we take $\alpha = \frac{1}{2}$, in which case r is also called the *Bayes factor* by Berger and Sellke (1987).

With complete observations, exact computation of r is manageable. When missing data occur, however, there is generally no cheap way of carrying out the necessary computations. Noting that it is easy to compute both marginal densities $\pi_0(Y, Z)$ and $\pi_1(Y, Z)$ when Z is augmented, where

$$\pi_0(Y, Z) = \int_{\Omega_0}\prod_{i=1}^{n}\pi(y_i, z_i|\theta)\pi_0(\theta)\,d\theta,$$

and

$$\pi_1(Y, Z) = \int_{\Omega}\prod_{i=1}^{n}\pi(y_i, z_i|\theta)\pi_1(\theta)\,d\theta,$$

we seek our solution from Markov chain Monte Carlo. It is observed, however, that a brute force application of DA or the Gibbs sampler does not work, because the degeneracy of Ω_0 will result in a reducible Markov chain if a prior distribution of the form (3) is used for θ. That is, if one starts with $\theta \in \Omega$, the chain will have zero probability of moving into the subspace Ω_0, and vice versa. In other words, the posterior distributions $\pi_0(\theta, Z|Y)$ and $\pi_1(\theta, Z|Y)$ do not have the same support. In contrast, once the parameter θ is collapsed down, the resulting chain is irreducible and $\pi_0(Z|Y)$ and $\pi_1(Z|Y)$ have the same support. From these observations, we are able to derive two Gibbs samplers for approximating the posterior odds ratio.

Method 1. Draw $Z^{(k)}$, $k = 1, \ldots, m$ from the distribution $\pi_1(Z|Y)$ by running a collapsed Gibbs sampler described in Section 2.1, where $\pi_1(Z|Y) \propto \pi_1(Y, Z)$. Then for each sampled $Z^{(k)}$, we can compute

$$r_z^{(k)} = \frac{\pi_0(Y, Z^{(k)})}{\pi_1(Y, Z^{(k)})} = \frac{\int_{\Omega_0}\pi(Y, Z^{(k)}|\theta)\pi_0(\theta)\,d\theta}{\int_{\Omega}\pi(Y, Z^{(k)}|\theta)\pi_1(\theta)\,d\theta},$$

$$k = 1, \ldots, m.$$

Thus we obtain m such ratios, $r_z^{(1)}, \ldots, r_z^{(m)}$. Then $\hat{r} = (r_z^{(1)} + \cdots + r_z^{(m)})/m$ is a consistent estimate of the true posterior odds ratio r, because, by the ergodicity theorem of Markov chain, the foregoing quantity converges to

$$E_{\pi_1}\left\{\frac{\pi_0(Y, Z)}{\pi_1(Y, Z)}\middle|Y\right\} = \int\frac{\pi_0(Y, Z)}{\pi_1(Y, Z)}\pi_1(Z|Y)\,dZ = \frac{\pi_0(Y)}{\pi_1(Y)}.$$

In other words, we only need to run the Gibbs sampler under one hypothesis, say H_A, to compute the odds ratio. It is also valid to run the sampler to draw samples from $\pi_0(Z|Y)$ and to approximate $1/r$ using the same procedure. But some numerical experiences given by Chen and Liu (1993), where this idea was applied to testing a hidden Markov structure in a time series, show that it is generally better to run the chain under π_1. An example is presented in Section 3 to illustrate the method.

Method 2. Let the prior of θ be $[\pi_0(\theta)\delta_{\Omega_0}(\theta) + \pi_1(\theta)]/2$. Then, after integrating out θ, we have a distribution of Z as $\pi(Z/Y) \propto \pi_0(Y, Z) + \pi_1(Y, Z)$. Thus

$$\pi(z_i|Y, Z_{[-i]}) \propto \pi_0(z_i|Y, Z_{[-i]}) + \frac{\pi_1(Y, Z_{[-i]})}{\pi_0(Y, Z_{[-i]})}\pi_1(z_i|Y, Z_{[-i]}).$$

So each step of the collapsed Gibbs sampler is just a draw from a mixture of the two predictive distributions, provided that the ratio $\pi_1(Y, Z_{[-i]})/\pi_0(Y, Z_{[-i]})$ can be easily computed, where we note that

$$\frac{\pi_1(Y, Z_{[-i]})}{\pi_0(Y, Z_{[-i]})} = \frac{\pi_1(y_i|X_{[-i]})}{\pi_0(y_i|X_{[-i]})}.$$

After drawing $Z^{(k)}$, $k = 1, \ldots, m$, from this chain, we can compute $u^{(k)} = \pi_0(Y, Z^{(k)})/[\pi_0(Y, Z^{(k)}) + \pi_1(Y, Z^{(k)})]$ for each k. Then, by a similar argument as in Method 1, the average of the $u^{(k)}$'s is an unbiased estimate of $r/(1 + r)$.

3. SOME EXAMPLES

3.1 Data from Multivariate Normal Distribution with Missingness

For a Normal covariance inference problem with missing observations, a natural tool of computation is DA (Tanner and Wong 1987). To implement the standard DA, one needs to sample from an inverse Wishart distribution, which is computationally intensive. We propose here that this step can be collapsed down to iterate only among the missing parts using the corresponding predictive distributions.

Multivariate Normal Data of Murray (1977). Table 1 contains 12 observations that are assumed drawn from a bivariate normal distribution with known means $\mu_1 = \mu_2 = 0$ and unknown covariance structure. The data were created by Murray in a discussion of work by Dempster, Laird, and Rubin (1977) and were used later by Tanner and Wong

Table 1. 12 Bivariate Normal Observations

1	1	−1	−1	2	2	−2	−2	*	*	*	*
1	−1	1	−1	*	*	*	*	2	2	−2	−2

NOTE: * indicates that the value is missing.

(1987). For notation, let ρ denote the correlation coefficient and let σ_1^2 and σ_2^2 denote the marginal variances. The original interest is in the posterior distribution of ρ given the incomplete data. We are only interested in comparing different posterior sampling schemes here.

Using Jeffreys's prior $\pi(\Sigma) \propto |\Sigma|^{-3/2}$, the predictive distribution for the nth observation conditioned on $x_1, \ldots, x_{n-1}$, all of which are two-dimensional, is

$$\pi(x_n | x_1, \ldots, x_{n-1}) \sim t_2(0, S_{n-1}/(n-2), n-2),$$

where $t_2(\ )$ indicates a bivariate t distribution and $S_{n-1} = \sum_{i=1}^{n-1} x_i x_i^T$ is the 2×2 sample covariance matrix. Hence if part of x_n (i.e., y_n) is observed, then the conditional distribution $\pi(z_n | x_1, \ldots, x_{n-1}, y_n)$—whose accurate form has been provided by Kong, Liu, and Wong (1994, sec. 3.1)—is a noncentral t distribution, as was shown by Box and Tiao (1973). Therefore, conditioned on the currently imputed values of $z_1, \ldots, z_{n-1}$, we can easily update z_n by a draw from the noncentral t distribution. The collapsed Gibbs sampler can be implemented as indicated in Section 2.1.

To compare the collapsed and the standard schemes, we compute autocovariance curves for each of the imputed samples, z_i, $i = 1, \ldots, 8$. Figure 1 contains two groups of autocovariance curves, each group with eight curves for eight missing values. The curves are estimated from simulations of 100 independent chains and 100 iterations for each chain. Because both chains are geometrically mixing, we fit model $\text{auto}(n) = C\rho^n + \varepsilon$ to the autocovariances for the two bundles, where $\text{auto}(n)$ denotes the lag-n autocovariance. It is seen that the $\hat{\rho}$ estimated from the standard scheme is about 2.5 times larger than the $\hat{\rho}$ estimated from the collapsed scheme.

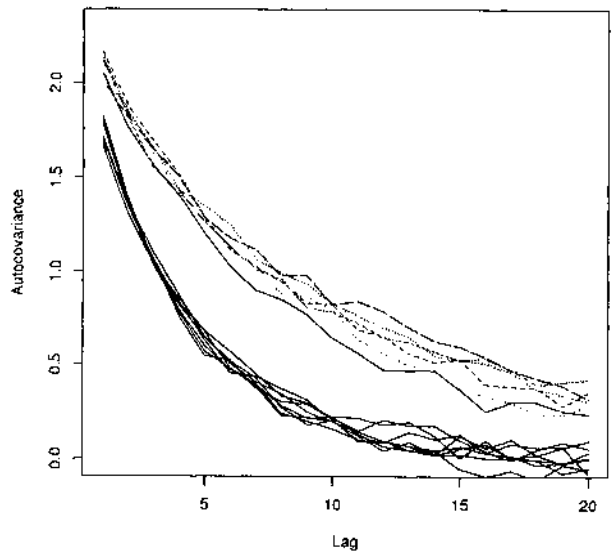

Figure 1. Autocovariance Plot for Both the Standard and the Collapsed Gibbs Sampling Schemes. The upper group represents the standard; the lower group, the collapsed.

3.2 Bayes Factor for a Discrete Data Example

Let us consider an example of simple graphical model with three variables, x_a, x_b, and x_c, each of which is binary (0 or 1). We assume that x_a and x_c are conditional independent given x_b, and that

$$\pi(x_b = 1 | x_a = 0) = \pi(x_b = 0 | x_a = 1) = \alpha,$$

$$\pi(x_c = 1 | x_b = 0) = \pi(x_c = 0 | x_b = 1) = \beta.$$

This structure can be intuitively expressed by the diagram $x_a \overset{\alpha}{\rightarrow} x_b \overset{\beta}{\rightarrow} x_c$. Observations involving incomplete configurations of $X = (x_a, x_b, x_c)$ are (1, 1, 0), (1, ?, 1) and (1, ?, 1), where "?" indicates missing. We are interested in computing the Bayes factor r for H_0: $\alpha = \beta$ versus H_A: $\alpha \neq \beta$. Because the likelihood function of the observations is $(1 - \alpha)\beta[\alpha\beta + (1 - \alpha)(1 - \beta)]^2$, the Bayes factor can be computed explicitly through tedious integrations, provided that the priors are conjugate. With flat priors, the Bayes factor is computed as $39/35 \approx 1.114$. But this computation will be infeasible if the number of incomplete observations is, say, 10 times larger.

Based on the approach described in Section 2.2, a Monte Carlo method is available to compute r and is applicable to much more complicated situations. Let Y denote the observed data and let z_1 and z_2 denote the missing parts. Then the complete data likelihood is

$$\pi(Y, z_1, z_2 | \alpha, \beta) \propto \alpha^{2-z_1-z_2}(1-\alpha)^{1+z_1+z_2}\beta^{3-z_1-z_2}(1-\beta)^{z_1+z_2}. \quad (4)$$

Therefore, if z_1, z_2 were known, then the Bayes factor with flat priors for H_0 versus H_A would be

$$r_z = \frac{\Gamma(6 - 2z_1 - 2z_2)\Gamma(2 + 2z_1 + 2z_2)\Gamma(5)^2}{\Gamma(3 - z_1 - z_2)\Gamma(2 + z_1 + z_2) \times \Gamma(4 - z_1 - z_2)\Gamma(1 + z_1 + z_2)\Gamma(8)}.$$

By integrating out α and β (in space $[0, 1] \times [0, 1]$) from (4), we obtain the predictive distribution

$$\frac{\pi(z_2 = 0 | z_1, Y)}{\pi(z_2 = 1 | z_1, Y)} = \frac{(2 - z_1)(3 - z_1)}{(2 + z_1)(1 + z_1)},$$

and the same for $\pi(z_1 | z_2, Y)$. A collapsed Gibbs sampler is run based on these conditional distributions; after stationarity, we compute $\hat{r}$ as indicated in Method 1 of Section 2.2. With $m = 400$ consecutive samples from the chain, we estimate $\hat{r} = 1.117$. It is also easy to figure out the exact distribution of r_z without actually running the sampler; that is, the r_z equals $8/7$ with probability .75 and $36/35$ with probability .25. So the ergodic average of the r_z's is equal to $r = 1.114$, which verifies the first method in Section 2.2.

The neat expression of the predictive distributions in the foregoing example is not an accident. With complete data and conjugate priors, Dawid and Lauritzen (1993) and Spiegelhalter and Lauritzen (1990) demonstrated the simplicity of the posterior and predictive distributions in both undirected and directed decomposable graphical models.

3.3 Nonparametric Bayesian Analysis

Many theoretical developments have been made on nonparametric Bayesian methods in the past two decades, by Ferguson (1974), Antoniak (1974), and many others. A major obstacle to widespread use of the method is computa-

tional. Recent developments of the Markov chain Monte Carlo seem to have rejuvenated the area, however.

To illustrate, we consider a binomial model for n unknown coins,

$$y_j \sim \text{Binomial}(l_j, \zeta_j) \qquad 1 \le j \le n,$$

where l_j is the total number of flips for coin j and ζ_j is its probability of landing heads up. The ζ_j's are assumed to be independently drawn from a common population with distribution F. Our interest is in drawing inference about F and the ζ_j based on the observed data y_j, $j = 1, \ldots, n$. Taking a Bayesian nonparametric approach, we assume that F, which is an infinite-dimensional parameter, is a priori drawn from a Dirichlet process $\mathcal{D}(\alpha)$, where α is a finite measure on interval [0, 1]. Note that $\mathcal{D}(\alpha)$ is a probability measure on $\mathcal{P}$, where $\mathcal{P}$ is the space of all probability measures on [0, 1]. (Readers not familiar with this area are referred to Ferguson 1974 for more details.)

If ζ_j's are actually observed, Ferguson (1974) explained that the posterior distribution of F is simply $\mathcal{D}(\alpha')$ with $\alpha' = \alpha + \sum_{j=1}^{n} \delta_{\zeta_j}$, where δ_{ζ_j} is a Dirac delta function. Thus this can be regarded as a missing data problem where $Z = (\zeta_1, \ldots, \zeta_n)$ plays the role of the missing values and F corresponds to the parameter θ in a standard setting. Because F is infinite-dimensional, it is difficult to directly apply the traditional DA procedure to iterate between the "missing data" and the "parameter," because the latter cannot be drawn correctly. Kong et al. (1994) found a noniterative method to overcome the difficulty, and Escobar (1994) described a Gibbs sampling approach. Both methods use the simple form of the Bayesian predictive distribution of ζ_i conditioned on the other ζ's. Precisely, because the likelihood $\pi(y_i \mid \zeta_i) \propto \zeta_i^{y_i}(1 - \zeta_i)^{l_i - y_i}$ and

$$[\zeta_i \mid Z_{[-i]}] \sim \frac{1}{n-1}\left(\alpha + \sum_{j \ne i} \delta_{\zeta_j}\right)$$

(Antoniak 1974), applying the Bayes theorem we obtain the predictive distribution of ζ_i as

$$\pi(\zeta_i \mid y_i, Z_{[-i]}) \propto \zeta_i^{y_i}(1 - \zeta_i)^{l_i - y_i}\alpha(\zeta_i) + \sum_{j \ne i} \zeta_j^{y_j}(1 - \zeta_j)^{l_j - y_j}\delta_{\zeta_j}(\zeta_i).$$

Hence a collapsed Gibbs sampler (with F collapsed down) can be applied to iteratively sample ζ_i using the foregoing simple predictive distribution.

4. COLLAPSING IN A GENERAL GIBBS SAMPLER

Let $X = (x_1, \ldots, x_d)$ be a random variable that can be partitioned into d components, with density $\pi(X)$. We consider a systematic scan Gibbs sampler applied to sample from this target distribution. That is, a Markov chain $\{X^{(k)} = (x_1^{(k)}, \ldots, x_d^{(k)}), k = 0, 1, \ldots\}$ is constructed with its transition function defined by the d-component Gibbs sampler,

$$K(X^{(k)}, X^{(k+1)}) = \prod_{l=1}^{d} \pi\{x_l^{(k+1)} \mid x_1^{(k+1)}, \ldots, x_{l-1}^{(k+1)}, x_{l+1}^{(k)}, \ldots, x_d^{(k)}\}. \quad (5)$$

It is easy to check that $\pi(X)$ is invariant under this transition.

Now suppose that the last two components, x_{d-1}, x_d, can be drawn together; then we have a reduced Gibbs sampler on a new partition of the random variable $X^* = \{x_1, \ldots, x'_{d-1}\}$, where $x'_{d-1} = \{x_{d-1}, x_d\}$, by *grouping*. Furthermore, suppose that the component x_d can be integrated out; then an even more reduced sampler on $X^- = \{x_1, \ldots, x_{d-1}\}$, with its marginal density $\pi(X^-) = \int \pi(X)\, dx_d$, results from *collapsing*. We will compare the original scheme and the two new ones.

To argue rigorously, we introduce some concepts concerning a Markov chain and its associated function spaces. Let $L^2(\pi)$ denote the set of all functions t that are square integrable with respect to π; that is, $\text{var}\{t(X)\} < \infty$. This set is a *Hilbert space* with an inner product defined by $\langle t, s\rangle = E\{t(X)s(X)\}$. Let $X^{(0)}, X^{(1)}, \ldots$, be a general state-space Markov chain with transition function $K(X, Y) = P(X^{(1)} = Y \mid X^{(0)} = X)$. We define a conditional expectation operator $\mathbf{F}$ on $L^2(\pi)$ for the Markov chain as

$$\mathbf{F}t(X) = \int K(X, Y)t(Y)\, dY = E\{t(X^{(1)}) \mid X^{(0)} = X\}.$$

We observe immediately that the *norm* of the operator is at most 1, where the norm is defined as $\|\mathbf{F}\| = \sup \|\mathbf{F}t(X)\|$ with the supremum taken over all functions with $E(t^2) = 1$. On the other hand, because the constant function c is an eigenfunction of the operator corresponding to eigenvalue 1, we know that the norm of $\mathbf{F}$ is exactly 1. When the chain is *reversible* (i.e., the *detailed balance* condition $\pi(X)K(X, Y) = \pi(Y)K(Y, X)$ is satisfied), $\mathbf{F}$ is a self-adjoint operator. When $\mathbf{F}$ is compact and self-adjoint (which is true when the state space is finite and the chain is reversible), the second largest eigenvalue (in absolute value) of $\mathbf{F}$ characterizes the mixing rate, or convergence rate, of the Markov chain. Many methods are available for bounding the second largest eigenvalue and finding the actual rate of convergence for this case (see Diaconis 1988 and Diaconis and Strook 1991 for details). Methods for dealing with nonreversible chain have been rare, however. (See Fill 1991 for the "reversibilization" method in treating nonreversible chains.)

Now we consider a subspace of $L^2(\pi)$. Let $L_0^2(\pi) = \{t(X) \in L^2(\pi) : E\{t(X)\} = 0\}$, which is a space of mean 0 functions with finite variance. Clearly, this is again a Hilbert space with the same inner product and is invariant under the operator $\mathbf{F}$. We use $\mathbf{F}_0$, called the *forward operator*, to denote the operator on $L_0^2(\pi)$ induced by $\mathbf{F}$. Then the largest eigenvalue of $\mathbf{F}_0$ is exactly the same as the second largest eigenvalue of $\mathbf{F}$. Typically, the spectral radius of $\mathbf{F}_0$ characterizes the rate of convergence of the Markov chain in both reversible and nonreversible cases. When the chain is reversible, the spectral radius of $\mathbf{F}_0$ is the same as its norm. A general relationship between the norm and the spectral radius of an operator is

$$\lim_{n \to \infty} \|\mathbf{F}_0^n\|^{1/n} = r,$$

where r is the spectral radius. This suggests that one can compare different Markov chains by comparing the norms

Table 2. DNA Sequences From 18 Loci, Each 105 Bases Long

taatgtttgtgctggtTTTTGTGGCATCGGGCGAGAATagcgcgtggtgtgaaagactgtttttttgatcgttttcacaaaaatggaagtccacagtcttgacag
gacaaaaacgcgtaacaaaagtgtctataatcacggcagaaaagtccacattgattatttgcacggcgtcacactttgctatgccatagcatttttatccataag
acaaatcccaataacttaattattgggatttgttatatataactttataaattcctaaaattacacaaagttaataactgtgagcatggtcatatttttatcaat
cacaaagcgaaagctatgctaaaacagtcaggatgctacagtaatacattgatgtactgcatgtatgcaaaggacgtcacattaccgtgcagtacagttgatagc
acggtgctacacttgtatgtagcgcatctttctttacggtcaatcagcatggtgttaaattgatcacgttttagaccatttttttcgtcgtgaaactaaaaaaacc
agtgaattatttgaaccagatcgcattacagtgatgcaaacttgtaagtagatttccttaattgtgatgtgtatcgaagtgtgttgcggagtagatgttagaata
gcgcataaaaaacggctaaattcttgtgtaaacgattccactaatttattccatgtcacactttcgcatctttgttatgctatggttatttcataccataagcc
gctccggcggggttttttgttatctgcaattcagtacaaaacgtgatcaacccctcaattttcccttgctgaaaaattttccattgtctcccctgtaaagctgt
aacgcaattaatgtgagttagctcactcattaggcaccccaggctttacactttatgcttccggctcgtatgttgtgtggaattgtgagcggataacaatttcac
acattaccgccaattctgtaacagagatcacacaaagcgacggtggggcgtaggggcaaggaggatggaaagaggttgccgtataaagaaactagagtccgtttta
ggaggaggcgggaggatgagaacacggcttctgtgaactaaaccgaggtcatgtaaggaatttcgtgatgttgcttgcaaaaatcgtggcgattttatgtgcgca
gatcagcgtcgttttaggtgagttgttaataaagatttggaattgtgacacagtgcaaattcagacacataaaaaaacgtcatcgcttgcattagaaaggtttct
gctgacaaaaaagattaaacatacctttatacaagacttttttttcatatgcctgacggagttcacacttgtaagttttcaactacgttgtagactttacatcgcc
ttttttaaacattaaaattcttacgtaatttataatctttaaaaaaagcatttaatattgctcccgaacgattgtgattcgattcacatttaaacaatttcaga
cccatgagagtgaaattgttgtgatgtggttaacccaattagaattcgggattgacatgtcttaccaaaaggtagaacttatacgccatctcatccgatgcaagc
ctggcttaactatgcggcatcagagcagattgtactgagagtgcaccatatgcggtgtgaaataccgcacagatgcgtaaggagaaaataccgcatcaggcgctc
ctgtgacggaagatcacatcgcagaataaataaatcctggtgtccctgttgataccgggaagccctgggccaacttttggcgaaaatgagacgttgatcggcacg
gatttttatactttaacttgttgatatttaaaggtatttaattgtaataacgatactctggaaagtattgaaagttaatttgtgagtggtcgcacatatcctgtt

of the corresponding forward operators. It is interesting to note here that $\|\mathbf{F}_0\|^2$ equals the second largest eigenvalue of the transition operator for the reversibilized chain, which ties in with the method of Fill (1991).

Let $\mathbf{F}_s$ denote the forward operator for the standard Gibbs sampler, corresponding to the transition function (5); let $\mathbf{F}_g$ denote the forward operator corresponding to the grouping procedure, and let $\mathbf{F}_c$ denote the collapsed Gibbs sampler with x_d integrated out. The three samplers can be illustrated by the diagrams of their respective visiting schemes:

$$\begin{aligned}
&\mathbf{F}_s\colon \quad x_1 \to x_2 \to \cdots \to x_d \\
&\mathbf{F}_g\colon \quad x_1 \to x_2 \to \cdots \to \{x_{d-1}, x_d\} \\
&\mathbf{F}_c\colon \quad x_1 \to x_2 \to \cdots \to x_{d-1}.
\end{aligned} \tag{6}$$

Theorem 1 (Three-schemes Theorem). The norms of the three forward operators are ordered as

$$\|\mathbf{F}_c\| \le \|\mathbf{F}_g\| \le \|\mathbf{F}_s\|.$$

Generally, the Gibbs sampler itself does not specify exactly how the random variable should be partitioned. This is a decision that users have to make, providing an opportunity to their ingenuity. A good Gibbs sampling algorithm must meet two conflicting criteria: (1) drawing one component conditioned on the others must be computationally simple, and (2) the Markov chain induced by the Gibbs sampler with such partitioning components must converge reasonably fast to its equilibrium distribution. For example, drawing the variables jointly with no partitioning at all is optimal for convergence but is formidable; this is the reason why the Gibbs sampler was invented. Theorem 1 provides a theoretical confirmation of such a confliction. It seems to be a reasonable strategy to "group" or "collapse" whenever it is computationally feasible, like those examples in Section 3. But as a whole, it is left to the reader to make compromises to balance all the aforementioned factors.

5. REGULATORY BINDING SITE PROBLEM

The collapsed Gibbs sampler was applied to a set of unaligned DNA fragments, previously analyzed by Lawrence and Reilly (1990) using an EM algorithm, that contain cyclic AMP receptor protein (CRP) binding sites. CRP is a positive control factor necessary for the expression of catabolite repressable genes. The location of the CRP binding sites in these sequences have been experimentally determined (see Lawrence and Reilly 1990 for references), so that using this set allows us to test the ability of our method (versus the EM algorithm) to locate the sites. Two obvious advantages of the Gibbs sampling approach over the EM approach are (1) the stochastic nature of the Gibbs sampler makes it more able to escape from a local mode, and (2) the idea of considering one random variable a time with the rest of the variables held fixed enables us to build more sophisticated and realistic models. Elsewhere we show (Lawrence et al. 1993) that the Bayesian models and the Gibbs sampling approach provide flexibility in describing biological sequences and power for finding subtle motifs and, unlike the EM algorithm, can identify multiple motifs. We first develop a simple collapsed Gibbs sampling scheme, and then build in a Metropolis step to cope with the special multimodality feature of the problem.

Eighteen DNA sequences from different loci are given, each $L = 105$ bases long (Table 2). It is known that there is at least one binding site (maybe more) of $J = 22$ bases long in each of the 18 sequences. For example, a binding site in the first sequence of Table 2 starts at the 17th position and stops at the 38th position (indicated by capital letters). We are interested in designing a computational method to aid in determining positions and base frequencies of the binding sites (thereby allowing the experimenter to more quickly characterize such sites).

Because regulatory DNA binding proteins typically recognize conserved sites, we adopt the strategy of looking for similar regions among the 18 DNA sequences. Following Lawrence and Reilly (1990), we treat the starting positions of all the binding sites as missing data Z and the actual residue frequencies in each binding site as a sample from a product multinomial model parameterized by Θ. Then a data augmentation scheme (Tanner and Wong 1987) can be applied to find the posterior distribution of Θ as well as the posterior mode of $\pi(Z \mid Y)$; that is, the most probable positions of the binding sites a posteriori.

More precisely, a 4×22 dimensional parameter Θ is needed to characterize the residue frequencies of the binding sites, where

$$\Theta = (\theta_1, \ldots, \theta_J),$$

in which $\theta_j = (\theta_{ja}, \theta_{jt}, \theta_{jg}, \theta_{jc})^T$ represents the frequencies of the four nucleotides A, T, G, and C at the jth position of a binding site. For example, the probability of observing the nucleotide T at the jth base of a binding site is θ_{jt}. For a given Θ, all bases in a binding site are assumed to be mutually independent. So the probability of seeing a binding site as the one indicated in the first sequence of Table 2 is $\theta_{1t}\theta_{2t} \ldots \theta_{21a}\theta_{22t}$. The bases in all the DNA sequences are also assumed independent. Therefore, for example, for a given Θ and an observed DNA sequence $B = (b_1, b_2, \ldots, b_L)$, the probability that the starting position for the binding site is 17, conditioned on the knowledge that there is a site in the sequence and that all starting positions are equally likely, is proportional to

$$\theta_{1b_{17}} \times \cdots \times \theta_{Jb_{17+J-1}}.$$

Here J is 22. Therefore, it is easy to sample a starting position of a binding site in a sequence for a given Θ. On the other hand, had we known the actual binding sites, we would have been able to update the posterior distribution of Θ, which would be a product of J independent Dirichlet distributions had a Dirichlet prior been used. A usual data augmentation scheme is then easy to implement.

For the sake of a simple argument, the prior distribution on Θ is assumed to be

$$\pi(\Theta) = \pi_1(\theta_1) \ldots \pi_J(\theta_J),$$

where $\pi_j(\theta_j)$ is a Dir(1, 1, 1, 1) distribution. Let $B_1, \ldots, B_{18}$ denote the 18 DNA sequences, each of length $L = 105$; that is, $B_i = (b_{i1}, \ldots, b_{iL})$. Let $Z = (z_1, \ldots, z_{18})$ denote the vector of starting positions of binding sites in 18 DNA sequences that cannot be observed and are treated as missing values. Imagine a data augmentation scheme applied to iteratively draw Θ conditioned on Z and then Z conditioned on Θ (all steps are conditioned on the observed sequences). For this problem, the step of drawing Θ involves sampling from 22 four-dimensional Dirichlet distributions and is time consuming. For a problem with protein sequences (Lawrence et al. 1993), the corresponding Dirichlet distributions will be 20-dimensional, which will further increase the computational burden. But we can apply collapsing, as discussed in Section 2.1, to design a more efficient algorithm.

With known $\mathbf{Z}_{[-i]}$, the posterior distribution of Θ is a product of independent Dirichlet distributions,

$$\prod_{j=1}^{J} \text{Dir}(c_{j,a} + 1, c_{j,t} + 1, c_{j,g} + 1, c_{j,c} + 1),$$

where $c_{j,a}$, for example, is the count of nucleotide type A among all the jth base of the 17 "known" sites. Hence the predictive distribution for the starting position z_i in sequence B_i is

$$P(z_i = s | \mathbf{Z}_{[-i]}, B_1, \ldots, B_{18}) \propto \prod_{j=1}^{J} (c_{j,b_{t(s+j-1)}} + 1)$$

$$s = 1, \ldots, L - J + 1. \quad (7)$$

This predictive distribution can be used to implement a collapsed Gibbs sampling algorithm.

Now we observe a special feature of the problem. Let $Z^0 = (z_1^0, \ldots, z_{18}^0)$ be the starting positions of the true sites, and suppose that it is the true mode of the distribution. Then those sites $Z = Z^0 + \delta = (z_1^0 + \delta, \ldots, z_{18}^0 + \delta)$, where δ is small, are also local modes of the distribution. We call them *shift modes* because they differ from the true mode by a common shift. The reason for this is that shifting δ bases for all the z's simultaneously still results in $22 - \delta$ correctly aligned bases. Thus when $z_1, \ldots, z_{17}$ are shifted from the true mode by δ, the 18th position predicted by formula (7) is also highly likely to have the same amount of shift. In this way, all the sites are still aligned. Thus the Gibbs sampler will not enable a global change for all random components simultaneously. This feature suggests a combination with another kind of transition to encourage a global "shifting."

For an integer δ, let $Z + \delta = (z_1 + \delta, \ldots, z_{18} + \delta)$. If one of the z_i equals 1, we assume that $Z - 1 \equiv Z$; if one of the z_i is $L - J + 1 = 84$, then $Z + 1 \equiv Z$. We construct a hybrid algorithm by inserting the following Metropolis step (Metropolis et al. 1953) after a few Gibbs sampling iterations:

Suppose that the current state is $Z^{(k)} = Z$; first choose $\delta = +1$ or -1 with probability $1/2$ each, and compute

$$p = \frac{\pi(Z + \delta)}{\pi(Z)}$$

using formula (8). If $p \geq 1$, then update the chain by $Z^{(k+1)} = Z + \delta$: if $p < 1$, then let $Z^{(k+1)} = Z + \delta$ with probability p and $Z^{(k+1)} = Z$ with probability $1 - p$.

Here $\pi(Z)$ can be easily computed up to a normalizing constant after collapsing down Θ; that is,

$$\pi\{Z = (z_1, \ldots, z_{18})\}$$

$$\propto \prod_{j=1}^{22} \{n_{ja}(\mathbf{Z})!n_{jt}(\mathbf{Z})!n_{jg}(\mathbf{Z})!n_{jc}(\mathbf{Z})!\}, \quad (8)$$

where, for example, $n_{ja}(\mathbf{Z})$ is the number of counts of nucleotide type A among the $(z_i + j - 1)$th base of sequence B_i, for all $i = 1, \ldots, 18$. An important fact to note is that this step involves little extra computation, because there are many cancellations between the common parts of $\pi(Z + \delta)$ and $\pi(Z)$. Simulation results show that whereas the ordinary Gibbs sampler may be easily stuck in a shift mode even with more than 3,000 iterations, the new Gibbs–Metropolis algorithm reaches the global mode within 400 steps for all of the different starting points (about 20) that we have tried.

Applying only a simple model similar to the mononucleotide model of Lawrence and Reilly (1990), we can identify 18 of the 24 footprinting sites. A comparison of the experimentally determined sites with sites located using Lawrence and Reilly's EM algorithm and our method is shown in Table 3. As pointed out by Lawrence and Reilly, more sophisticated models can be built to improve the ability of identifying the

Table 3. Starting Positions of the Sites

Sequence	Footprint sites	Two most likely sites: EM First	EM Second	GM First	GM Second
cole 1	17, 61	61	45	61	
eco arabop	17, 55	55	76	55	
eco bglrl	76	76	40	76	26
eco crp	63	63	73	63	45
eco cya	50	50	15	50	15
eco deop	7, 60	7	39	7	60
eco gale	42	24	76	42	24
eco ilvbpr	39	39	20	39	20
eco lac	9, 80	9	73	9	75
eco male	14	14	12	14	
eco malk	29, 61	61	29	61	35, 29
eco malt	41	41	11	41	51
eco ompa	48	48	12	48	12
eco tnaa	71	71	34	71	26
eco uxul	17	17	26	17	48
pbr-p4	53	53	84	53	27
trn9cat	1, 84	5	66	5	84, 66
(tdc)	78	78	76	78	76

sites. The proposed Gibbs–Metropolis method is also applicable in those cases. More details have been reported by Lawrence et al. (1993).

APPENDIX: MATHEMATICAL PROOFS

For any two random variables X and Y, both of which can be multidimensional, we define the *maximal correlation* between them as

$$\rho(X, Y) \overset{\text{def}}{=} \sup \operatorname{cov}\{t(X), s(Y)\} = \sup \sqrt{\operatorname{var}[E\{t(X)|Y\}]}, \quad \text{(A.1)}$$

where the supremum is taken over all functions $t(X)$ and $s(Y)$ with variances equal to 1. Clearly, it is always true that $0 \le \rho(X, Y) \le 1$. Under certain regularity conditions (e.g., compactness), one has $\rho(X, Y) < 1$. Lancaster (1958) and Csàki and Fischer (1960) gave such conditions.

Lemma 1. If $Y' = (Y, Z)$ where Z is another random variable, then $\rho(X, Y') \ge \rho(X, Y)$.

Proof: Because for any function $t(X)$ it is true that

$$E\{t(X)|Y\} = E[E\{t(X)|Y, Z\}|Y],$$

it is obvious that $\operatorname{var}[E\{t(X)|Y'\}] \ge \operatorname{var}[E\{t(X)|Y\}]$. Hence the statement is true.

Lemma 2. For a time-homogeneous and stationary Markov chain $X^{(0)}, X^{(1)}, \ldots$, the norm of its forward operator $\mathbf{F}_0$ is equal to the maximal correlation between the two consecutive states. That is, $\|\mathbf{F}_0\| = \rho(X^{(0)}, X^{(1)})$.

Proof: By definition,

$$\|\mathbf{F}_0\| = \sup_{\|t\|=1, E(t)=0} \operatorname{var}[E\{t(X^{(1)})|X^{(0)} = X\}],$$

which is equal to $\rho(X^{(0)}, X^{(1)})$ as indicated in (A.1).

Lemma 3. Suppose that $X = (x_1, \ldots, x_d)$ and that the transition function for the Gibbs sampling chain is the one defined in (5). Then, under the stationarity of the chain, the joint distribution between two consecutive states X and Y is

$$\pi(X, Y) = \pi(X)\pi(y_1|X_{[-1]})\pi(y_2|y_1, X_{[-\{1,2\}]}) \cdots \pi(y_d|y_1, \ldots, y_{d-1}).$$

Proof: By definition of the Gibbs sampler.

Lemma 4. Under the same setting as Lemma 3, the norm of the forward operator corresponding to the standard Gibbs sampler $\mathbf{F}_s$ is equal to the maximal correlation between $X_{[-1]}$ and $Y_{[-d]}$.

Proof: By Lemma 2, it is understood that $\|\mathbf{F}_s\| = \rho(X, Y)$. Because, given $X_{[-1]}$, x_1 is conditionally independent of Y, and given $Y_{[-d]}$, y_d is conditionally independent of X, it is intuitively true that X and Y are dependent only through $X_{[-1]}$ and $Y_{[-d]}$. Mathematically, because of the conditional independence, it is true that for any function t of X,

$$E\{t(X)|Y\} = E[E\{t(X)|X_{[-1]}\}|Y] = E\{t_1(X_{[-1]})|Y\},$$

where $t_1(X_{[-1]}) = E\{t(X)|X_{[-1]}\}$. Therefore, $\operatorname{var}[E\{t(X)|Y\}] = \operatorname{var}[E\{t_1(X_{[-1]})|Y\}] \le \operatorname{var}(t_1)\rho^2(X_{[-1]}, Y)$. Because $\operatorname{var}(t_1) \le \operatorname{var}(t)$, it shows that

$$\rho(X, Y) \le \rho(X_{[-1]}, Y).$$

On the other hand, for any function $t(X)$ not involving the first component, that is, $t_1(X_{[-1]}) = t(X)$,

$$\operatorname{var}[E\{t(X_{[-1]})|Y\}] = \operatorname{var}[E\{t(X)|Y\}] \le \rho^2(X, Y)\operatorname{var}\{t(X_{[-1]})\},$$

which shows that $\rho(X_{[-1]}, Y) \le \rho(X, Y)$ as well. As a consequence $\rho(X, Y) = \rho(X_{[-1]}, Y)$. In the same manner, one can proceed to show that $\rho(X_{[-1]}, Y) = \rho(X_{[-1]}, Y_{[-d]})$.

Proof of the three-schemes theorem

Proof: The difference between the two operators $\mathbf{F}_c$ and $\mathbf{F}_g$ is illustrated by diagram (6). By Lemmas 2 and 4, it is enough to compare the two maximal correlations between consecutive states. Let $X = (x_1, \ldots, x_{d-1}, x_d)$, and $X^- = X_{[-d]}$. Then for the two schemes,

$$\|\mathbf{F}_c\| = \rho(X^-_{[-1]}, Y^-_{[-\{d-1\}]})$$

and

$$\|\mathbf{F}_g\| = \rho(X_{[-1]}, Y_{[-\{d-1,d\}]}).$$

From the scheme arrangement, it is seen that $X_{[-1]} = (X^-_{[-1]}, x_d)$ and $Y^-_{[-\{d-1\}]} = Y_{[-\{d-1,d\}]}$. By Lemma 1, it is concluded that $\|\mathbf{F}_c\| \le \|\mathbf{F}_g\|$.

To compare $\mathbf{F}_g$ and $\mathbf{F}_s$, we let $X^* = (x_1, \ldots, \{x_{d-1}, x_d\})$ and let Y^* be a consecutive follower of X^* in the chain. Then $X_{[-1]} = X^*_{[-1]}$, and $Y_{[-d]} = (Y^*_{[-\{d-1,d\}]}, y_{d-1})$. By Lemmas 1 and 4,

$$\|\mathbf{F}_g\| = \rho(X^*_{[-1]}, Y^*_{[-d]}) \le \rho(X_{[-1]}, Y_{[-d]}) = \|\mathbf{F}_s\|.$$

The theorem is proved.

Comment. Monotonicity of the norms corresponding to the three operators can not guarantee the same result for the spectral radii. An example of this was provided by Liu, Wong and Kong (1994).

[Received April 1992. Revised September 1993.]

REFERENCES

Aitchison, J., and Dunsmore, I. R. (1975), *Statistical Prediction Analysis*, New York: Cambridge University Press.

Antoniak, C. E. (1974), "Mixtures of Dirichlet Processes With Applications to Bayesian Nonparametric Problems," *The Annals of Statistics*, 2, 1152–1174.

Berger, J. O., and Sellke, T. (1987), "Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence" (with discussion), *Journal of the American Statistical Association*, 82, 112–122.

Besag, J. (1974), "Spatial Interaction and the Statistical Analysis of Lattice Systems" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 36, 192–236.

Besag, J., and Green, P. J. (1993), "Spatial Statistics and Bayesian Computation" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 55, 24–35.

Box, G. E. P., and Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis*, Reading, MA: Addison-Wesley.

Chen, R., and Liu, J. S. (1993), "Bayesian Classification via Switching Regression," submitted to *ASA Proceeding of Bayesian Statistical Science*.

Csàki, P., and Fischer, J. H. (1960), "Contributions to the Problem of Maximal Correlation," *Matematikao Kotato Intezet, Kozlemenyei*, 5, 325–337.

Dawid, A. P., and Lauritzen, S. L. (1993), "Hyper Markov Laws In The Statistical Analysis of Decomposable Graphical Models," *The Annals of Statistics*, 21, 1272–1317.

Dempster, A. P., Laird, N., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data Via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

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

Diaconis, P., and Strook, D. (1991), "Geometric Bounds for Eigenvalues of Markov Chains," *The Annals of Applied Probability*, 1, 36–61.

Escobar, M. D. (1994), "Estimating Normal Means with a Dirichlet Process Prior," *Journal of the American Statistical Association*, 89, 268–277.

Fill, J. A. (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.

Ferguson, T. S. (1974), "Prior Distribution on Space of Probability Measures," *The Annals of Statistics*, 2, 615–629.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Kong, A., Liu, J. S., and Wong, W. H. (1994), "Sequential Imputations and Bayesian Missing Data Problems," *Journal of the American Statistical Association*, 89, 278–288.

Lancaster, H. O. (1958), "The Structure of Bivariate Distributions," *The Annals of Statistics*, 29, 719–736.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A., and Wooton, J. (1993), "Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment," *Science*, 262, 208–214.

Lawrence, C. E., and Reilly, A. A. (1990), "An Expectation Maximization Algorithm for the Identification and Characterization of Common Sites in Unaligned Biopolymer Sequences," *PROTEINS*, 7, 41–51.

Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis with Missing Data*, New York: John Wiley.

Liu, J. S., Wong, W. H., and Kong, A. (1994), "Covariance Structure of the Gibbs Sampler With Applications to the Comparisons of Estimators and Augmentation Schemes, *Biometrika*, 81, 27–40.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), "Equations of State Calculations by Fast Computing Machines," *Journal of Chemical Physics*, 21, 1087–1091.

Morris, C. (1983), "Natural Exponential Families With Quadratic Variance Functions: Statistical Theory," *The Annals of Statistics*, 11, 515–529.

Murray, G. D. (1977), Comment on "Maximum Likelihood from Incomplete Data Via the EM Algorithm" by A. P. Dempster, N. Laird, and D. B. Rubin, *Journal of the Royal Statistical Society*, Ser. B, 39, 27–28.

Rubin, D. B., and Schafer, J. (1990), "Efficiently Creating Multiple Imputations for Incomplete Multivariate Normal Data," *ASA Proceeding of Statistical Computing Section*, 83–88.

Rubinstein, R. Y. (1981), *Simulation and the Monte Carlo Method*, New York: John Wiley.

Smith, A. F. M., and Roberts, G. O. (1993), "Bayesian Computation Via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 55, 3–23.

Spiegelhalter, D. J., and Lauritzen, S. L. (1990), "Sequential Updating of Conditional Probabilities on Directed Graphical Structures," *Networks*, 20, 579–605.

Swendsen, R. H., and Wang, J. S. (1987), "Nonuniversal Critical Dynamics in Monte Carlo Simulations," *Physical Review Letters*, **58**, 86–88.

Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation" (with discussion), *Journal of the American Statistical Association*, 82, 528–550.