

# Subsampling the Gibbs Sampler: Variance Reduction

Steven N. MacEachern and Mario Peruggia \*

Department of Statistics  
The Ohio State University  
Columbus, OH 43210-1247

December 1996

Revised, April 1999

## Abstract

Subsampling the output of a Gibbs sampler in a non-systematic fashion can improve the efficiency of marginal estimators if the subsampling strategy is tied to the actual updates made. We illustrate this point by example, approximation, and asymptotics. The results hold both for random scan and fixed scan Gibbs samplers.

**Key Words:** Bayesian analysis; Efficiency; Estimation; Markov chains; Monte Carlo; Stationary time series.

---

\*Reprint requests should be addressed to: Steven N. MacEachern, Department of Statistics, The Ohio State University, Columbus, OH 43210-1247. e-mail: [snm@stat.ohio-state.edu](mailto:snm@stat.ohio-state.edu)

# 1 Introduction

Markov chain Monte Carlo (MCMC) methods have been used to solve an extraordinary array of problems in Bayesian inference since the seminal paper by Gelfand and Smith (1990) described the techniques and illustrated the ease with which they could provide approximate solutions to a variety of intractable problems. The ease of implementation of the methods has led to an explosion of research: sophisticated Bayesian models that defy analytical solution are now routinely fit through MCMC methods. Besag and Green (1993), Smith and Roberts (1993), Besag et al. (1995), and Robert (1995) provide recent overviews of some of the research in this area.

The output of the Markov chain is a sequence of parameter vectors  $\theta$ , say  $\theta^{(1)}, \dots, \theta^{(N)}$ , belonging to  $R^p$ . If  $\theta^{(1)} \sim \pi$ , where  $\pi$  is the limiting distribution of the chain, then the distribution of each  $\theta^{(j)}$  is  $\pi$ . If  $\theta^{(1)}$  is initialized from a different distribution, the early portion of the sequence is discarded, and some later value, say  $\theta^{(B)}$ , is taken to have distribution nearly equal to  $\pi$ . The sequence of parameter values enables one to make inference about functionals of interest. Estimation is accomplished either through tabulation, for example using  $e_1 = (1/N) \sum_{j=1}^N \theta_i^{(j)}$  as an estimate of  $E[\theta_i]$ , or with the help of "Rao-Blackwellization." The standard Rao-Blackwellized estimate would be  $(1/N) \sum_{j=1}^N E[\theta_i^{(j)} | \theta_{(i)}^{(j)}]$ , where  $\theta_{(i)}^{(j)} = (\theta_1^{(j)}, \dots, \theta_{i-1}^{(j)}, \theta_{i+1}^{(j)}, \dots, \theta_p^{(j)})$ . These estimators provide consistent estimates of  $E[\theta_i]$  under very mild conditions (Tierney, 1994).

One issue of constant concern with MCMC methods is effective implementation. Usually, research on this issue is either aimed at the development of techniques for improving the rate of mixing of the Markov chain or at the development of more efficient estimators. Recent papers have proscribed subsampling the output of a Markov chain. Theoretical results (Geyer, 1992; MacEachern and Berliner, 1994) describe the effect of a *systematic* subsampling of the output of the Markov chain wherein every  $k^{th}$  parameter vector is included in the estimate. A systematic subsample would yield  $(k/N) \sum_{j=1}^{N/k} \theta_i^{(jk)}$  as an estimator of  $E[\theta_i]$ . This estimator always has larger variance than the non-sampled estimator,  $e_1$ . In this article, we investigate alternative forms of subsampling a Markov chain. In Sections 2 and 3, we

focus on a popular MCMC method, the random scan Gibbs sampler, and describe subsampling schemes where the indices of the parameters selected for updating (and the functional to be estimated) determine which of the  $\theta^{(j)}$  are included in the subsample. These estimators can be considerably more efficient than non-sampled estimators. Not merely a curiosity, the subsampling methodology is extended in Section 4 to the most common MCMC algorithm in use today, the fixed scan Gibbs sampler.

## 2 The Subsampled Estimator

The Gibbs sampler is, perhaps, the most popular MCMC method. The Markov chain for a Gibbs sampler is constructed by stringing together a sequence of conditional distributions and sampling from them. In the sequel, we assume that  $\theta$  is a  $p$ -dimensional parameter having a continuous density with respect to Lebesgue measure supported on  $R^p$ . The conditional distributions in this case are  $[\theta_i|\theta_{(i)}]$ , for  $i = 1, \dots, p$ . A fixed scan Gibbs sampler arises from a deterministic, repeated set of successive conditional distributions, say  $[\theta_1|\theta_{(1)}], \dots, [\theta_p|\theta_{(p)}]$ . A random scan Gibbs sampler chooses successive conditional distributions by means of a random process, usually a Markov chain. A particularly simple random scan Gibbs sampler that is often implemented chooses the conditional distributions by means of independent multinomial variates.

Let  $P_i$  represent the transition kernel obtained from the conditional distribution of  $[\theta_i|\theta_{(i)}]$ , for  $i = 1, \dots, p$ . For  $j = 2, 3, \dots$ , the conditional generations will be determined by a sequence of independent variates  $Z_j$ , where  $P(Z_j = i) = p_i > 0$ , for  $i = 1, \dots, p$ , and  $\sum_{i=1}^p p_i = 1$ . If  $Z_j = i$ , then the  $j^{th}$  generation will be from  $[\theta_i|\theta_{(i)}]$ . This setup induces a Markov chain for  $\{\theta^{(j)}\}$  with transition kernel  $P = \sum_{i=1}^p p_i P_i$ . For the remainder of this article, we assume that the induced Markov chain is irreducible and aperiodic. We also assume that the chain is initialized in its unique stationary distribution, resulting in a stationary Markov chain. See Roberts and Sahu (1997) for comparison of fixed scan and random scan Gibbs samplers.

Consider a subsampled estimator for a functional, say  $E[f(\theta)]$ , that tabulates the value of  $f(\theta)$  only

for those iterates where particular parameters are updated. Formally, let  $S$  be a nonempty subset of the elements  $\{1, \dots, p\}$ , introduce the indicator  $I(Z_j \in S)$ , for  $j = 2, 3, \dots$ , and define  $M = \sum_{j=1}^N I(Z_j \in S)$ , where  $I(Z_1 \in S) = 1$  by convention. Let  $\{j_k\}$  represent the indices of the  $M$  subsampled  $\theta^{(j)}$ , and write  $e_2 = (1/M) \sum_{j=1}^N f(\theta^{(j)}) I(Z_j \in S) = (1/M) \sum_{k=1}^M f(\theta^{(j_k)})$ .

The idea behind the estimator  $e_2$  is that  $f(\theta^{(j)})$ , although expressed in terms of the entire vector  $\theta$ , may depend only on a portion of  $\theta$ , say  $(\theta_1, \dots, \theta_r)$ . Since the value of  $f(\theta^{(j)})$  only changes when one of these  $r$  parameters is updated, the non-subsampled estimator  $e_1 = (1/N) \sum_{j=1}^N f(\theta^{(j)})$  tabulates the same value of  $f(\theta)$  for a set of successive iterates. The number of times that a particular value of  $f(\theta^{(j)})$  is replicated is random, depending on the  $Z_j$ . The subsampled estimator tabulates  $f(\theta^{(j)})$  only when one of  $(\theta_1, \dots, \theta_r)$  is updated, eliminating the extra variability due to the random number of replicates of a particular  $f(\theta)$ . The corollary to the next proposition shows that the estimator  $e_2$  is consistent.

**Proposition 2.1** *Assume that  $\theta^{(1)} \sim \pi$ , that  $\pi$  has a continuous density with respect to Lebesgue measure supported on  $R^p$ , and that  $S$  is a nonempty subset of  $\{1, \dots, p\}$ . Let  $Z_2, Z_3, \dots$  be i.i.d. multinomial random variates with  $P(Z_j = i) = p_i > 0$ , for  $i = 1, \dots, p$ . Conditional on  $Z_j$ , the non-subsampled chain is a Gibbs sampler with transition kernel  $P_{Z_j}: \theta^{(j-1)} \rightarrow \theta^{(j)}$ , for  $j = 2, 3, \dots$ . The subsampled chain is  $\{\theta^{(j)} \mid Z_j \in S\}_{j=1}^\infty = \{\theta^{(j_k)}\}_{k=1}^\infty$ , where  $Z_1$  always belongs to  $S$  by convention. Under the above conditions, the subsampled chain is an irreducible, stationary Markov chain.*

*Proof.* The transition kernel for the subsampled chain is

$$P = \sum_{l=1}^{\infty} \sum_{(i_1, \dots, i_l) \in T_l} p_{i_1} \dots p_{i_l} P_{i_1} \dots P_{i_l},$$

where  $T_l = \{(i_1, \dots, i_l) \mid i_1, \dots, i_{l-1} \in \{1, \dots, p\} \setminus S, i_l \in S\}$ . Since  $\theta^{(1)} \sim \pi$  and  $\pi$  is a stationary distribution for  $P$ , the subsampled chain is stationary. The support of  $\theta^{(j_n)} \mid \theta^{(1)}$  is equal to  $R^p$ , provided that there is a positive probability of updating all coordinates of  $\theta$  in the  $n - 1$  steps between  $\theta^{(1)}$  and  $\theta^{(j_n)}$ , i.e.,  $n$  must be larger than the cardinality of  $S$ . Hence, the chain is irreducible and aperiodic.

**Corollary 2.2** *The estimator  $e_2$  for the subsampled Markov chain described in Proposition 2.1 is consistent, provided  $E[|f(\theta)|] < \infty$ .*

*Proof.* Tierney (1994) provides a result that shows that the simple average,  $(1/N) \sum_{j=1}^N f(\theta^{(j)})$ , of an ergodic Markov chain converges to  $E[f(\theta)]$ . Proposition 2.1 shows that the subsampled Markov chain is ergodic.

The example in Section 3 will show that subsampling on a non-systematic basis may reduce the variance of estimators and that this reduction in variability holds up asymptotically. The next result shows that the reduction in variance is attributable to tying the subsample to the component of  $\theta$  that is generated rather than to the non-systematic nature of the sample. As in the example in Section 3, we assume that the subsampling times are given by independent geometric variates with mean  $p_1^{-1}$ , but here these variates are also assumed to be independent of the sequence used to determine which component of  $\theta$  is updated.

**Proposition 2.3** *Assume that  $\{\theta^{(j)}\}_{j=1}^\infty$  is a stationary, irreducible, aperiodic Markov chain with limiting distribution  $\pi$ , and that  $\text{Var}[f(\theta)] < \infty$ . Let a positive integer  $M$  be given, and suppose that a geometric subsample of size  $M$  is used to estimate  $E[f(\theta)]$ . Formally, let  $v_1, \dots, v_{M+1}$  be i.i.d. geometric variates with mean  $p_1^{-1}$ ,  $j_k = \sum_{i=1}^k v_i$ ,  $N = \sum_{i=1}^{M+1} v_i - 1$ ,  $e_1 = (1/N) \sum_{j=1}^N f(\theta^{(j)})$ , and  $e_3 = \frac{1}{M} \sum_{k=1}^M f(\theta^{(j_k)})$ . Then  $\text{Var}_N[e_3] \geq \text{Var}_N[e_1]$ , for every  $N \geq M$ , and  $\text{Var}[e_3] \geq \text{Var}[e_1]$ .*

*Proof.* The distribution of  $(N + 1)$  is that of a negative binomial random variable counting the number of trials until the  $(M + 1)^{\text{st}}$  success in a sequence of i.i.d. Bernoulli trials with success probability  $p_1^{-1}$ . For a particular  $N$ , each of the  $\binom{N}{M}$  sequences of indices  $1 \leq j_1 < j_2 < \dots < j_M \leq N$  has the same probability,  $p_1^{M+1}(1 - p_1)^{N-M}$ , of being observed. Hence, the  $M$  subsampled iterates form a simple

random sample from the  $N$  iterates. Since  $E_N[e_3|e_1] = e_1$ , we have

$$\begin{aligned} \text{Var}_N[e_3] &= \text{Var}_N(E_N[e_3|e_1]) + E_N(\text{Var}_N[e_3|e_1]) \\ &= \text{Var}_N[e_1] + E_N(\text{Var}_N[e_3|e_1]) \\ &\geq \text{Var}_N[e_1], \end{aligned}$$

with equality holding only when  $\text{Var}_N[e_3|e_1]$  is 0. Averaging over  $N$  yields  $\text{Var}[e_3] \geq \text{Var}[e_1]$ .

**Note:** A similar result can be obtained by fixing the total number  $N$  of iterates to be generated and letting the size  $M$  of the subsample be random.

### 3 The Multivariate Normal Distribution

We now turn to the efficiency of the subsampled estimator relative to the non-subsampled estimator. We consider examples based on multivariate normal distributions where the goal is estimation of the expected value of the first coordinate.

Suppose that  $\theta = (\theta_1, \dots, \theta_p)$  has a multivariate normal distribution,  $\pi$ , with  $E[\theta_i] = 0$ , for  $i = 1, \dots, p$ , and covariance matrix  $\Sigma$ . A random scan Gibbs sampler is used to obtain a sample from  $\pi$  as follows. Let positive probabilities  $p_i$ , with  $\sum_{i=1}^p p_i = 1$ , be given. The Markov chain is initialized in  $\pi$ . At each step of the chain, a coordinate  $i$  is selected at random according to the probabilities  $p_i$  and a random variate is generated from the conditional distribution of  $\theta_i$  given  $\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_p$ .

The Gibbs sampler is run through  $N$  iterations. Let  $\{\theta_1^{(j)}\}$ ,  $j = 1, \dots, N$ , denote the entire sequence of recorded  $\theta_1$  values and let  $\{\theta_1^{(j_k)}\}$ ,  $k = 1, \dots, M$ , denote the subsequence of the  $\theta_1$  values recorded when the first coordinate of  $\theta$  is updated. By convention, we set  $j_1 = 1$ . To facilitate the ensuing comparisons, we will condition on  $M$ , the number of times that  $\theta_1$  is updated. The structure of this Gibbs sampler implies that the times between updates of  $\theta_1$  are independent geometric variates, say  $v_k$ , each with mean  $p_1^{-1}$ .

The non-subsampled and subsampled estimators are, respectively,

$$e_1 = \frac{1}{N} \sum_{j=1}^N \theta_1^{(j)} = \sum_{k=1}^M w_k \theta_1^{(j_k)},$$

where  $w_k = v_k / \sum_{i=1}^M v_i$ , and

$$e_2 = \frac{1}{M} \sum_{k=1}^M \theta_1^{(j_k)}.$$

Exact algebraic comparisons between the variances of the two estimators are difficult to derive, but, in the case when  $\theta_1, \dots, \theta_p$  are independent standard normal random variables, direct calculations show that the asymptotic (as  $M \rightarrow \infty$ ) relative efficiency of  $e_1$  to  $e_2$  is  $\text{ARE}(e_1, e_2) = 2 - p_1$ , so that there is a gain in efficiency attributable to subsampling the Markov chain that is retained in the limit. Details of the derivation are available from the authors.

We now present the results of a small simulation study. In all cases, we chose the limiting distribution,  $\pi$ , of the Gibbs chain  $\{\theta^{(j)}\}$  to be multivariate normal of dimension  $p = 5$ , with zero mean and unit variance for the univariate marginals. We considered two types of correlation structures, the first characterized by a covariance matrix with off-diagonal elements all equal to the same constant  $\rho$ , and the second by a covariance matrix with the autoregressive pattern  $\text{Cov}[\theta_k, \theta_l] = \rho^{(l-k)}$ , for  $1 \leq k < l \leq 5$ . For both correlation structures, we examined several values of  $\rho$ .

We implemented the random scan Gibbs sampler by selecting for updating, at each step, one of the five components of the vector  $\theta$  according to the uniform probabilities  $p_i = 1/5$ , for  $i = 1, \dots, 5$ . The goal was to estimate  $E[\theta_1] = 0$ . For each normal limiting distribution,  $\pi$ , considered in the simulation, we repeated the following steps 1,000 times.

1. Draw  $\theta^{(1)}$  from  $\pi$  and set  $j_1 = 1$ .
2. Run the random scan Gibbs sampler until  $\theta_1$  has been updated 100 times, yielding  $\{\theta_1^{(j_k)}\}$ ,  $k = 1, \dots, 101$ , and  $\{\theta_1^{(j)}\}$ ,  $j = 1, \dots, N + 1$ .
3. Compute the subsampled estimate,  $e_2$ , by averaging the first 100 values of  $\{\theta_1^{(j_k)}\}$  and evaluate the square of the estimate,  $e_2^2$ .

Correlation Structure							
	Constant				Autoregressive		
$\rho$	0.0	0.2	0.4	-0.2	0.2	0.4	-0.4
$\text{RE}(e_1, e_2)$	2.02	1.66	1.30	1.36	1.90	1.60	1.54

Table 1: Relative Efficiency of the Non-Subsampled and Subsampled estimators of  $E[\theta_1] = 0$ .

4. Compute the non-subsampled estimate,  $e_1$ , by averaging the first  $N$  values of  $\{\theta_1^{(j)}\}$  and evaluate the square of the estimate,  $e_1^2$ .

As a measure of the relative efficiency of the two estimators,  $\text{RE}(e_1, e_2)$ , we then computed the ratio between the average of the 1,000 values of  $e_1^2$  and the average of the 1,000 values of  $e_2^2$ .

Table 1 summarizes the simulation results. Qualitatively, the value for the case of independence obtained in the simulation ( $\text{RE}(e_1, e_2) = 2.02$ ) is in line with the corresponding asymptotic theoretical value ( $\text{ARE}(e_1, e_2) = 1 - p_1 = 1.8$ ). In the other cases, the simulation results show that the gains in relative efficiency derived from subsampling become smaller as the strength of the dependence between the components of the vector  $\theta$  increases.

## 4 The fixed scan Gibbs sampler

Subsampling also proves useful for the fixed scan Gibbs sampler. To establish this point, we cast the fixed scan Gibbs sampler in our notation. The sequence of variates to be updated in the Markov chain is now fully deterministic. The indices of the variates to be updated are given by  $Z_i = i \pmod{p}$ , with the convention that  $Z_i = 0$  becomes  $Z_i = p$ . A typical description of the fixed scan Gibbs sampler would count the number of complete cycles through  $(\theta_1, \dots, \theta_p)$ . However, as in the previous sections, we count the number of iterates of the Markov chain as the number of variates updated. Hence,  $N$  iterates, which we take to be a multiple of  $p$ , represents only  $N/p$  complete cycles through  $\theta$ .

To illustrate the value of subsampling with a fixed scan Gibbs sampler, we consider estimation of a non-linear, multivariate function of  $\theta$ , say  $g(\theta)$ . The standard, non-subsampled estimator averages the

values of  $g$  after each complete cycle, yielding  $e_1 = (p/N) \sum_{j=1}^{N/p} g(\theta^{(jp)})$ . With the typical description of the Gibbs sampler, this estimator makes use of every iterate of the Markov chain. Our view of the Markov chain suggests that we actually have  $N$  values of  $\theta$  on which to base our estimate, and so  $e_1$  can (arguably) be described as a subsampled estimator. With this view, a non-sampled version of the estimator is  $e_2 = (1/N) \sum_{j=1}^N g(\theta^{(j)})$ . We note that for  $g(\cdot)$  which are linear in  $\theta$ ,  $e_1$  is nearly equal to  $e_2$ , with slight, asymptotically negligible differences arising from treatment of the first  $p$  iterates of the Markov chain.

Our first result describes the benefit of the finer scale estimator for non-linear  $g$ . We make the comparison to an average of estimators of the form  $e_1$ . The estimators of the form  $e_1$  differ only in which coordinate of  $\theta$  is first updated.

**Proposition 4.1** *Assume that  $\{\theta^{(j)}\}_{j=1}^N$  is an irreducible, aperiodic Markov chain, initialized with  $\theta^{(1)} \sim \pi$ , induced by a fixed scan Gibbs sampler, and having limiting distribution  $\pi$ . Define  $e_{1i} = (p/N) \sum_{j=1}^{N/p} g(\theta^{((j-1)p+i)})$ ,  $V_i = \text{Var}(e_{1i})$ , and  $C_{ii'} = \text{cov}(e_{1i}, e_{1i'})$  for  $i, i' = 1, \dots, p$ . Assume  $V_i < \infty$  for  $i = 1, \dots, p$ . Then  $\text{Var}(e_2) \leq p^{-1} \sum_{i=1}^p \text{Var}(e_{1i})$ .*

*Proof.*

$$\begin{aligned}
\text{Var}(e_2) &= \text{Var}\left(p^{-1} \sum_{i=1}^p e_{1i}\right) \\
&= p^{-2} \left[ \sum_{i=1}^p V_i + 2 \sum_{i < i'} C_{ii'} \right] \\
&\leq p^{-2} \left[ \sum_{i=1}^p V_i + 2 \sum_{i \neq i'} \frac{V_i}{2} \right] \\
&= p^{-2} \left[ \sum_{i=1}^p V_i + \sum_{i \neq i'} V_i \right] \\
&= p^{-1} \sum_{i=1}^p V_i,
\end{aligned}$$

with the inequality following from a comparison of the geometric mean of  $V_i$  and  $V_{i'}$  to their arithmetic

mean.

Proposition 4.1 demonstrates that  $e_2$  is preferable in expectation to a version of  $e_1$  created by randomly selecting the starting index of the Markov chain. It also implies that  $e_2$  is better than the worst case version of  $e_1$ , a desirable property for those concerned with minimax evaluation of estimators.

A case which is particularly attractive to study from a theoretical point of view is the situation where the Markov chain and  $g(\cdot)$  show enough symmetry that the  $V_i$  of the previous proposition are all equal. The  $C_{ii'}$  may still differ, but the following corollary provides a direct comparison of  $e_1$  to  $e_2$ .

**Corollary 4.2** *Suppose, in addition to the assumptions of Proposition 4.1, that  $V_i = V$  for  $i = 1, \dots, p$ . Then  $\text{Var}(e_2) \leq \text{Var}(e_1)$ .*

Corollary 4.1 suggests that  $e_2$ , rather than  $e_1$ , should be considered the standard for non-subsampled estimators with the fixed-scan Gibbs sampler. Having established a solid non-subsampled estimator for the basis of comparison, we turn to the issue of subsampling.

The drawback of the fine scale estimator quickly becomes apparent when an actual  $g(\cdot)$  is considered. Consider a  $p > 2$  variate problem with the goal of estimating  $E[\theta_1\theta_2]$ . Relying on the fact that  $\theta_1$  and  $\theta_2$  remain unchanged as  $\theta_3, \dots, \theta_p$  are generated, we write the fine scale estimator as

$$\begin{aligned} e_2 &= \frac{1}{N} \sum_{i=1}^N \theta_1^{(i)} \theta_2^{(i)} = \frac{1}{N} \left[ \sum_{j=1}^{N/p} \theta_1^{((j-1)p+1)} \theta_2^{((j-1)p+1)} + (p-1) \sum_{j=1}^{N/p} \theta_1^{(jp)} \theta_2^{(jp)} \right] \\ &= \frac{p}{N} \left[ \frac{1}{p} \sum_{j=1}^{N/p} \theta_1^{((j-1)p+1)} \theta_2^{((j-1)p+1)} + \frac{(p-1)}{p} \sum_{j=1}^{N/p} \theta_1^{(jp)} \theta_2^{(jp)} \right]. \end{aligned}$$

The unequal weighting of the two sums is odd. A more natural estimator is the equally weighted

$$e_3 = \frac{p}{N} \left[ \frac{1}{2} \sum_{j=1}^{N/p} \theta_1^{((j-1)p+1)} \theta_2^{((j-1)p+1)} + \frac{1}{2} \sum_{j=1}^{N/p} \theta_1^{(jp)} \theta_2^{(jp)} \right].$$

We note that  $e_3$  may be motivated by the same consideration used to construct estimators for the random scan Gibbs sampler. We tabulate values of  $g(\cdot)$  whenever the function has the possibility of changing. The next proposition demonstrates that  $e_3$  can be preferable to  $e_2$ .

**Proposition 4.3** *Assume the same conditions as in Proposition 4.1. Define  $V_i$  and  $C_{12}$  as before. If  $V_1 = V_2$ , then  $\text{Var}(e_3) \leq \text{Var}(e_2) \leq \text{Var}(e_1)$ .*

*Proof.* Calculation shows that  $\text{Var}(e_3) = V_1/4 + V_2/4 + 2C_{12}/4 = V/2 + C_{12}/2$ , and that  $\text{Var}(e_2) = (((p-1)^2 + 1)V + 2(p-1)C_{12})/(p^2)$ . The expression for  $\text{Var}(e_2)$  is minimized at  $p = 1/2$ , indicating that  $\text{Var}(e_2) \geq \text{Var}(e_3)$ . Since  $\text{Var}(e_1) \geq \text{Var}(e_2)$ , the final conclusion follows.

This last result demonstrates that subsampling the fixed scan Gibbs sampler can reduce the variability of estimators. Although the conditions under which the improvement has been demonstrated are strong, we believe that the intuitive appeal of these results suggest that an estimator of the form of  $e_3$ , which averages values of  $g(\cdot)$  whenever they may change, should be the default tabulation style estimator for the fixed scan Gibbs sampler.

## 5 Conclusions

This paper has presented a class of estimators for use with Gibbs samplers that in some instances roughly halve the variance of the standard non-subsampled estimators. The effort required to implement the subsampled estimators is negligible, and so we recommend their use as standard practice. The improvement due to subsampling is inherent in the method. It holds up asymptotically, and so is worthwhile even for extremely long runs of the Markov chain.

The discussion in this paper has been limited to examples where the choice of parameter to update is either governed by an independent and identically distributed set of variates or is deterministic. The estimators can be used with Gibbs samplers where choice of the parameter to update is governed by a Markov chain. Such extension requires a development similar to that for “generalized sampled chains” (see Meyn and Tweedie, 1993). Even approximate calculations are made difficult by a sequential dependence in the component of  $\theta$  to update, but we find that the intuition behind the estimators

remains convincing: a carefully devised subsample may reduce the variability of estimators.

The estimators find easy application to functions of more than one parameter. The simplest use, as in Section 4, is to tabulate the results any time one of the parameters upon which the function depends is updated. More sophisticated use of the estimators is based on an evaluation of the tradeoff between the variability introduced by the random replication of similar values of the function to be estimated and the variability introduced by omitting distinct (though similar) values of the function. When the function varies only slightly with a parameter, say  $\theta_1$ , the former source of variability may outweigh the latter, and so it can be better to omit 1 from the set of subsampling indices,  $S$ .

Some functions of more than one parameter can be represented as a linear combination of pieces, each of which relies on fewer parameters. Such functions can be estimated by first estimating each piece with a subsampled estimator and then combining the estimates of the pieces to create the final estimate.

A referee has suggested consideration of the random permutation Gibbs sampler, where, for each complete cycle of  $(\theta_1, \dots, \theta_p)$ , a random permutation is chosen to determine the order of updating. Such a method lies between the random scan and fixed scan Gibbs samplers described here. For this sort of Gibbs sampler, a subsampled estimator of the form of Section 4's  $e_3$  is often preferable to an estimator such as Section 4's  $e_2$ . The subsampled estimator avoids the now random weights associated with  $e_2$ . In simple examples, the superiority of  $e_3$  can be demonstrated, although, for brevity, we omit such discussion.

Finally, the idea behind these estimators can be extended to MCMC methods other than the Gibbs sampler. For more general methods such as the Metropolis-Hastings method, whenever the proposal distributions select a candidate value by changing only some of the coordinates of the parameter vector or when there is a distribution over proposal distributions, the subsampled estimator can be tied to the estimand or to the selection of the proposal distribution, providing potentially improved estimators.

## REFERENCES

- Besag, J. and Green, P. J. (1993), "Spatial Statistics and Bayesian Computation," *Journal of the Royal Statistical Society, Ser. B*, 55, 25–37.
- Besag, J., Green, P., Higdon D., and Mengersen K. (1995), "Bayesian Computation and Stochastic Systems," *Statistical Science*, 10, 3–66.
- Chung, K. L. (1974), *Probability and Mathematical Statistics*, (2nd edition), Orlando: Academic Press.
- 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.
- Geyer, C. J. (1992), "Practical Markov Chain Monte Carlo," *Statistical Science*, 7, 473–483.
- MacEachern, S.N. and Berliner, L.M. (1994), "Subsampling the Gibbs Sampler," *American Statistician*, 48, 188–190.
- Meyn, S.P. and Tweedie, R.L. (1993), *Markov Chains and Stochastic Stability*, New York: Springer-Verlag.
- Robert, C. (1995), "Convergence Control Methods for Markov Chain Monte Carlo Algorithms," *Statistical Science*, 10, 231–253.
- Roberts, G.O. and Sahu, S.K. (1997), "Updating Schemes, Correlation Structures, Blocking and Parameterization for the Gibbs Sampler," *Journal of the Royal Statistical Society, Ser. B*, 59, 291–317.
- Smith, A. F. M. and Roberts, G. O. (1993), "Bayesian Computation via the Gibbs Sampler and related Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society, Ser. B*, 55, 3–23.
- Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions," *The Annals of Statistics*, 22, 1701–1728.