BU-1360-M

Statistical Inference and Monte Carlo Algorithms

George Casella *
Cornell University

September 2, 1996

Abstract

This review article looks at a small part of the picture of the interrelationship between statistical theory and computational algorithms, especially the Gibbs sampler and the Accept-Reject algorithm. We pay particular attention to how the methodologies affect and complement each other.

*This research was supported by NSF Grant No. DMS-9625440. This is paper BU-1360-M in the Biometrics Unit, Cornell University, Ithaca, NY 14853. To be presented in Grenada, Spain, September 27, 1996. File spain2.tex

1 Introduction

Computations and statistics have always been intertwined. In particular, applied statistics has relied on computing to implement its solutions of real data problems. Here we look at another part of the relationship between statistics and computation, and examine a small part of how the theories not only are intertwined, but how they have influenced each other.

With the explosion of methods based on Monte Carlo methods, particularly those using Markov chain algorithms such as the Gibbs sampler, there has been a blurring of the distinction between the statistical model and the algorithmic model. This is particularly evident in the examples of Section 3. There, the statistical model will typically be a hierarchical model, while the computational algorithm will be based on a set of conditional distributions. We will see that the manner in which we view the model can have a large impact on the validity of the statistical inference. It is therefore important to consider the statistical model that underlies the Monte Carlo algorithm.

We can also turn things around. When one uses a Monte Carlo algorithm to do a calculation, it is common to process the output by taking an average. However, we should realize that the output from a Monte Carlo algorithm can be viewed as data, with the algorithm itself playing the part as a statistical model. As such, taking a naive average may not be the most effective way of processing the output. In Section 4 we look at this question, and investigate the effect of classical decision theory on output from the Accept-Reject algorithm. We consider these improvements as a post-simulation processing of a generated sample, which is statistically superior to the original estimator, although they may be computationally inferior in taking more computer time. However, this latter concern can also be addressed with estimators that offer statistical improvement while only requiring a slight increase in computational effort.

We also emphasize that our approach and, in particular, the optimizations involved in the derivation of some of the improved estimators, is based on statistical rather than computational principles. The overall goal of the statistician is to process samples in an optimal way, and to make the best inference possible. To do so requires treating an algorithm as a statistical model, and (as far as possible) ignoring the computational issues.

Another consideration in the interplay of statistical theory and algorithms is the prospect of using the structure of the algorithm to more efficiently construct an optimal procedure. We illustrate this in Section 5, where we look at three examples. These examples use the Gibbs sampler, and show that we can use the iterative nature of the algorithm to implement procedures that are sometimes computationally feasible and can result in an optimal inference. We end the paper with a short discussion section.

2 Synthesis

Given the audience of this presentation, a digression may be in order into the Bayes/frequentist approaches to statistics. The topic of algorithms, particularly Monte Carlo algorithms, is a prime example of an area that is best handled statistically by a mixture of the Bayesian and frequentist approaches. Moreover, it seems that to completely analyze, understand, and optimize the relationship between a statistical model, its associated inference, and the algorithm used for computations, both Bayesian and frequentist ideas must be used.

The Bayesian approach provides us with a means of constructing an estimator that, when evaluated according to its global risk performance, could result in an optimal frequentist estimator. This highlights important features of both the Bayesian and frequentist approaches. Although the Bayesian paradigm is well-suited for the construction of possibly optimal estimators, it is less well-suited for their global evaluation. The frequentist paradigm is quite complementary, as it is well-suited for global evaluations, but is less well-suited for construction.

We look at two examples, taken from Lehmann and Casella (1997).

Example 1 Rao-Blackwellizing the Gibbs Sampler. The Gibbs sampler (Geman and Geman 1984, Gelfand and Smith 1990) provides a method of computing Bayes estimators. These estimators are computed by averaging random variables and this averaging is improved if the Rao-Blackwell theorem is applied (Liu, Wong and Kong 1994, 1995). More precisely, in a typical use of the Gibbs sampler, our estimand is the actual Bayes estimator, which we are computing by generating random variables and averaging them. The validity of our method rests on the Ergodic Theorem (Law of Large Numbers). When the Rao-Blackwell theorem is applied to these averages, we get a new average with the same expectation (the actual value of the estimator) and smaller variance.

Thus, the calculation of a Bayes estimator is improved using a frequentist methodology. Moreover, monitoring convergence of the Gibbs sampler is essentially a frequentist problem, so again frequentist techniques can be used to improve Bayes estimators. ||

The preceding example shows how frequentist methods can aid a Bayesian approach. The reverse is also true.

Example 2 REML variance estimation. In the one-way random effects model

$$Y_{ij} = \beta + u_i + \epsilon_{ij} \quad (j = 1, \ldots, n_i, \quad i = 1, \ldots, k) \tag{1}$$

where β is the overall mean, u_i is a random effect, and ϵ_{ij} is error, it is often of primary interest to estimate σ^2 and σ_ϵ^2, the variance of the random effects u_i and ϵ_{ij}, respectively. Two basic problems must be overcome.

(a) Elimination of the effect of β from the estimates of σ^2 and σ_ϵ^2. As the latter are estimates of dispersion, they should not be affected by a change in the mean level.

(b) Interpretation of possibly negative estimates of variance, which can arise from some classical estimation methods (see Searle, et al. 1992, Section 3.5c).

Both (a) and (b) can be dealt with using frequentist methodologies. For example, the effect of β can be eliminated by requiring the variance estimates to be translation invariant (one derivation of the so-called REML variance estimates; see Searle et al. 1992, Section 6.6 and Chapter 9) and the negativity problem can be handled by truncation.

Alternatively, a Bayesian model can eliminate both of these problems in a straightforward way. First, the parameter β can be integrated out using a prior distribution, creating a marginal likelihood. Moreover, Bayes estimates of σ^2 and σ_ϵ^2 will never be negative.

Note that we are using the Bayesian approach to construct the estimators. The evaluation of the estimators, and establishment of any optimality properties, can still be done using a frequentist global risk approach.

||

Thus, it is important to view these two approaches as complementary rather than adversarial, as together they provide a rich set of tools and techniques for the statistician. Moreover, there are situations and problems in which one or the other approach is better-suited, or even a combination may be best, so a statistician without a command of both approaches may be less than complete.

3 Algorithms and Statistical Inference

In this section we look at how an algorithmic approach to a problem has fundamental repercussion on the statistical inference. In Section 3.1, where we mainly give details for the mixed linear model, we will see that approaching a problem through a Gibbs sampler can mask posterior impropriety. This can have a profound effect on the possible statistical inferences. In the most extreme cases, which are in no way pathological, evaluating a statistical model only through a Gibbs sampler can lead to erroneous, even nonsensical, inferences. This latter point is examined in Section 3.2.

3.1 How the Algorithm Affects the Posterior

The model equation of a general linear mixed model is given by

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon} \tag{2}$$

where $\mathbf{Y}$ is an $n \times 1$ vector of observations, $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects (parameters), $\mathbf{u}$ is a $q \times 1$ vector of random effects (random variables), $\mathbf{X}$ and $\mathbf{Z}$ are known design matrices whose dimensions are $n \times p$ and $n \times q$, respectively, and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of residual errors.

A typical set of error distributions (or priors) for the mixed model has $\boldsymbol{\epsilon}|\sigma_\epsilon^2 \sim N_n(\mathbf{0}, \mathbf{I}\sigma_\epsilon^2)$ and $\mathbf{u}|\sigma_1^2, \ldots, \sigma_r^2 \sim N_q(\mathbf{0}, \mathbf{D})$ where $\mathbf{u} = (\mathbf{u}_1' \; \mathbf{u}_2' \ldots \mathbf{u}_r')'$, $\mathbf{u}_i$ is $q_i \times 1$, $\mathbf{D} = \oplus_{i=1}^r \mathbf{I}_{q_i}\sigma_i^2$, and $\sum_{i=1}^r q_i = q$. The r subvectors of $\mathbf{u}$ correspond to the r different random factors in the experiment. It is also common to put a flat prior (Lebesgue measure) on the so-called fixed effects, represented by the vector $\boldsymbol{\beta}$. In classical mixed model inference, such an assumption is used in *REML*, or restricted maximum likelihood estimation. As is turns out, the type of prior used on $\boldsymbol{\beta}$ has no impact on what follows.

The variance components themselves, which are often the prime targets of inference, are often given power-type priors of the form

$$\pi_\epsilon(\sigma_\epsilon^2|b) \propto \left(\sigma_\epsilon^2\right)^{-(b+1)}, \quad \pi_i(\sigma_i^2|a_i) \propto \left(\sigma_i^2\right)^{-(a_i+1)}, \tag{3}$$

where the a_i's and b are known and the following conditional independence assumptions are in force: (1) given $\mathbf{u}$, $\mathbf{Y}$ is conditionally independent of $\sigma_1^2, \ldots, \sigma_r^2$, (2) given $\sigma_1^2, \ldots, \sigma_r^2$, $\mathbf{u}$ is conditionally independent of $\boldsymbol{\beta}$ and σ_ϵ^2, and (3) $\boldsymbol{\beta}$, σ_ϵ^2, and $\sigma_1^2, \ldots, \sigma_r^2$ are a priori independent.

All of these assumptions can be summarized in the *hierarchical model*

$$\mathbf{Y}|\mathbf{u}, \sigma_\epsilon^2, \boldsymbol{\beta} \sim N_n(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \mathbf{I}\sigma_\epsilon^2)$$

$$\pi(\boldsymbol{\beta}) \propto 1 \qquad \mathbf{u}|\sigma_1^2, \ldots, \sigma_r^2 \sim N_q(\mathbf{0}, \mathbf{D}) \qquad \pi_\epsilon(\sigma_\epsilon^2|b) \propto \left(\sigma_\epsilon^2\right)^{-(b+1)} \tag{4}$$

$$\pi_i(\sigma_i^2|a_i) \propto \left(\sigma_i^2\right)^{-(a_i+1)}.$$

With the increased popularity of Monte Carlo algorithms such as the Gibbs sampler, the experimenter tends to pay less attention to the model specified by (4), and rather concentrates on the set of full conditionals, which make up the input into the Gibbs Markov chain. For our mixed model, these conditionals are given by

$$f\left(\sigma_i^2|\boldsymbol{\sigma}_{-i}, \mathbf{y}, \mathbf{u}, \sigma_\epsilon^2, \boldsymbol{\beta}\right) = IG\left(a_i + \frac{q_i}{2}, \frac{2}{\mathbf{u}_i'\mathbf{u}_i}\right) \tag{5}$$

$$f\left(\sigma_\epsilon^2|\boldsymbol{\sigma}, \mathbf{y}, \mathbf{u}, \boldsymbol{\beta}\right) = IG\left(b + \frac{n}{2}, 2\left\{(\mathbf{y} - (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}))'(\mathbf{y} - (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}))\right\}^{-1}\right)$$

$$f\left(\mathbf{u}|\boldsymbol{\sigma}, \mathbf{y}, \sigma_\epsilon^2, \boldsymbol{\beta}\right) = N_q\left(\left(\mathbf{Z}'\mathbf{Z} + \sigma_\epsilon^2\mathbf{D}^{-1}\right)^{-1}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \; \sigma_\epsilon^2\left(\mathbf{Z}'\mathbf{Z} + \sigma_\epsilon^2\mathbf{D}^{-1}\right)^{-1}\right)$$

$$f\left(\beta|\sigma\mathbf{y},\sigma_\epsilon^2,\mathbf{u}\right)=N_p\left(\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{y}-\mathbf{Z}\mathbf{u}\right),\ \sigma_\epsilon^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)$$

where $\boldsymbol{\sigma}=(\sigma_1^2,\ldots,\sigma_r^2)$, $\boldsymbol{\sigma}_{-i}=(\sigma_1^2,\ldots,\sigma_{i-1},\sigma_{i+1},\ldots,\sigma_r^2)$, IG stands for inverted gamma and we say that $X\sim IG(r,s)$ if $f_X(t)\propto t^{-r-1}\exp(-1/st)$ for positive t.

If $2a_i\leq -q_i$ for some i or $2b\leq -n$, then at least one of the conditionals is improper, since the inverted gamma density is defined only when both parameters are positive (Berger 1985, p. 561). Clearly, one improper conditional implies an improper posterior.

Although it may be tempting to assume that propriety of the conditionals in (5) implies propriety of the posterior distribution, this is false. Indeed, there are many values of the vector $(a_1,a_2,\ldots a_r,b)$ which simultaneously yield proper conditionals ($2a_i>-q_i\ \forall i$ and $2b>-n$) and an improper posterior. Thus, in general, if one incorrectly assumes propriety of a posterior and writes down a (false) proportionality statement like

$$\pi(\sigma_1^2,\ldots,\sigma_r^2,\sigma_\epsilon^2,\mathbf{u},\beta|\mathbf{y})\propto f(\mathbf{y}|\mathbf{u},\sigma_\epsilon^2,\beta)f(\mathbf{u}|\sigma_1^2,\ldots,\sigma_r^2)\pi(\beta)\pi_\epsilon(\sigma_\epsilon^2|b)\prod_{i=1}^{r}\pi_i(\sigma_i^2|a_i) \tag{6}$$

where f is used to represent a generic density, it may happen that the Gibbs conditionals are all proper densities. Such a situation is very dangerous because, if the output from the Gibbs sampler fails to warn the user that the posterior is improper (which seems to be the common situation), the result could be an inference about a nonexistent posterior distribution. We will return to this point in Section 3.2.

We now state a theorem that will insure the propriety of posterior distributions coming from the model. This theorem is similar, in spirit, to those given in Ibrahim and Laud (1991), who consider the use of Jeffreys's prior in generalized linear models (GLM's), Dey, Gelfand and Peng (1994), who discuss the use of improper priors in overdispersed GLM's, and Natarajan and McCulloch (1995), who deal with mixed models for binomial responses. Another related paper is Zeger and Karim (1991) who discuss the use of improper priors and Gibbs sampling in GLM's. For a proof of the theorem see Hobert and Casella (1996).

Theorem 1 *Let* $t=rank(\mathbf{P}_X\mathbf{Z})=rank\left(\mathbf{Z}'\mathbf{P}_X\mathbf{Z}\right)\leq q$ *where we define* $\mathbf{P}_X=\left(\mathbf{I}-\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right)$. *There are two cases:*

1. *If* $t=q$ *or if* $r=1$ *then conditions (i), (ii), and (iii) below are necessary and sufficient for the propriety of the posterior distribution of model (4).*

2. *If $t < q$ and $r > 1$ then conditions (i), (ii), and (iii) below are sufficient for the propriety of the posterior distribution of model (4) while necessary conditions result when (ii) is replaced with (ii') $q_i > -2a_i$.*

$$(i)\ a_i < 0$$
$$(ii)\ q_i > q - t - 2a_i$$
$$(iii)\ n + 2\sum a_i + 2b - p > 0.$$

Thus, we see that it is relatively easy to check if the posterior distributions are proper, being merely a matter of counting categories. Also, conditions *(i)-(iii)* are intuitively reasonable, and can be interpreted as requiring that we have enough observations, in particular enough observations on the variance components σ_i^2, to adequately control the tails of the posterior (large enough q_i).

3.2 How the Algorithm Affects the Inference

In this section, we look at what can happen to the inference if one uses a set of Gibbs conditionals, all of which are proper, that do not correspond to a proper posterior. This situation was investigated in detail by Hobert and Casella (1995), and we will discuss a few of their findings.

A set of conditional densities such as those in (5) may, or may not, result in a proper posterior. However, the fact that may obscure the impropriety of the posterior is the *functional compatibility* of the set of densities. First consider the following simple example from Casella and George (1992).

Example 3 The pair of exponential conditionals densities

$$f_1(x|y) = ye^{-yx} \text{ and } f_2(y|x) = xe^{-xy}.$$

appear to be a pair of conditional densities, but there is no joint density function which will yield f_1 and f_2 as conditional densities. If such a joint density did exist, the pair f_1 and f_2 would be *compatible.* As one does not exist, this pair is incompatible. However, the non-integrable function $g(x, y) = \exp(-xy)$, if treated as a joint density, does yield f_1 and f_2 as its "conditionals". In such a case, where no proper $g(\cdot)$ exists, but an improper one does, we say that f_1 and f_2 are *functionally compatible.* This is the dangerous case, as f_1 and f_2 appear to be a set of conditional densities. This is exactly what can happen in (5) if the conditions of Theorem 1 are not satisfied. ||

When there are more than two variables, the definitions of compatibility and functional compatibility become more involved, but the idea is the same. Compatibility of a set of densities was investigated by Besag (1974), Arnold

and Press (1989), and Gelman and Speed (1993). They tended to focus on conditions under which a set of conditional densities could be used to uniquely determine the joint density, assuming that such a density existed. In our case, however, we cannot assume that such a joint density exists.

The major concern for a user of a Gibbs sampler based on a set of functionally compatible densities that are not compatible (that is, for which not proper joint density exists), is what inference can be made from the resulting Markov chain? This is the question investigated in detail by Hobert and Casella (1995), and the results are quite negative. They prove the following theorem.

Theorem 2 *Let $f_1, \ldots, f_m$ be a set of conditional densities on which a Gibbs sampler is based. The resulting Markov chain Φ is positive recurrent if and only if $f_1, \ldots, f_m$ are compatible.*

Thus, a set of densities that are only functionally compatible will not result in a positive recurrent Markov chain. Hence, there cannot be any stationary probability distribution for the chain to converge to. Moreover, there is virtually no reasonable inference that can be made. Under some additional technical conditions (which are satisfied for most typical Gibbs samplers), it can be shown that if $t : A \to \Re_+$ is a bounded measurable function for which, given $\epsilon > 0$, there exists a compact set $C \in A$ such that $t(y) \leq \epsilon \; \forall \, y \in C^c$, then

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} t(\Phi_i) = 0 \quad \text{a.s.} \tag{7}$$

In a typical Gibbs sampling application, one might estimate a posterior density $\pi(\theta|\mathbf{y})$ with an average of conditional posterior densities, say $\pi(\theta|\mathbf{y}) \approx (1/m) \sum_{i=1}^{m} \pi(\theta|\mathbf{y}, \lambda_i)$. It will often be the case that the densities $\pi(\theta|\mathbf{y}, \lambda_i)$ satisfy the conditions on the function t above. Hence, the only place the average $(1/m) \sum_{i=1}^{m} \pi(\theta|\mathbf{y}, \lambda_i)$ can converge to is 0; or else it will not converge.

Gibbs samplers based on a set of densities that are not compatible result in Markov chains that are *null*, that is, they are either null recurrent or transient. In either case, there is no limiting probability distribution. However, output from the Gibbs sampler may produce nice looking pictures of the supposed marginal posterior densities, particularly when the posterior density is computed as an average of conditional densities. But there can be no actual distribution to which the Gibbs picture corresponds. This was the problem with the Gibbs-based conclusions of Wang et al. (1993, 1994) and Gelfand et al. (1990) as they used models for which a posterior distribution did not exist.

An insidious feature of this situation is that a null Gibbs chain may be undetectable to the practitioner, that is, the resulting Monte Carlo approximations appear completely reasonable. Moreover, not only do the Gibbs averages look reasonable, but the actual output from the Markov chain may appear reasonable. (Consider Geyer 1992, who published what he first believed to be proper

Gibbs output, but later found that it corresponded to an improper posterior. He noted, in proof, that, "...(the model) produces an *improper* posterior, so the Gibbs sampler apparently converged when there was no stationary distribution for it to converge to. A run of one million iterations gave no hint of lack of convergence...") Thus, it is not surprising that a practitioner can be fooled into believing that the Gibbs chain is giving a reasonable inference.

In order to demonstrate just how reasonable some of these null Gibbs chains can appear, we give an example.

Example 4 The one-way random effects model (1) with a typical set of priors is

$$\begin{gathered} y_{ij}|\beta, \mathbf{u}, \sigma_\epsilon^2 \sim N(\beta + u_i, \sigma_\epsilon^2) \\ \beta \sim d\beta \qquad \mathbf{u} \sim N_k\left(\mathbf{0}, \mathbf{I}\sigma^2\right) \qquad \sigma_\epsilon^2 \sim (\sigma_\epsilon^2)^{-(b+1)} \\ \sigma^2 \sim (\sigma^2)^{-(a+1)}. \end{gathered} \tag{8}$$

For a simulation study we set $k = 7$, $n_i = n = 5$, $\sigma^2 = 5$, $\sigma_\epsilon^2 = 2$, and $\beta = 10$. The vector $(u_1, \ldots, u_7)$ was simulated by generating seven iid $N(0, 5)$ random variables and the vector $(\epsilon_{11}, \ldots, \epsilon_{75})$ was simulated by generating 35 iid $N(0, 2)$ random variables. We also set $a = b = 0$, which yields an improper posterior. A Gibbs chain was constructed using the conditionals given in (5). We denote the chain by $\left(\sigma^{2(j)}, \sigma_\epsilon^{2(j)}, \mathbf{u}'^{(j)}, \beta^{(j)}\right)$, $j \geq 1$. At the start, all parameters were set to one, except for the overall mean, β, which was set to eight. The chain was first allowed to run for 15,000 iterations; keep in mind that the word "burn-in" is not appropriate for these initial iterations because the chain is null and is therefore not converging (in the usual sense). The sole purpose of these initial iterations was to provide the chain with ample opportunity to misbehave and alert us that something may be wrong; it never did. We chose 15,000 because a typical burn-in would probably be in the hundreds (see Gelfand et al. 1990 and Wang et al. 1993) so that if our chain did not misbehave during the burn-in stage, neither would that of an unknowing experimenter.

After the initial 15,000 iterations, the output from the 15,001st through the 16,000th was collected. Figure 1 is a histogram of the 1,000 effect variances from the null Gibbs chain, that is, $\sigma^{2(j+15{,}000)}$, $j = 1, 2, \ldots, 1000$, with a Monte Carlo approximation of the supposed marginal posterior density superimposed. Figure 2 is the analog of Figure 1 for the error variance component. The density approximations in Figures 1 and 2 were calculated using the usual "average of conditional densities" approximation . All of these plots appear perfectly reasonable even though the posterior distribution is improper and the Monte Carlo density approximations have almost sure pointwise limits of zero or no limit at all. Clearly, if one were unaware of the impropriety, plots like these could lead to seriously misleading conclusions.

This particular posterior is improper due to an infinite amount of mass near $\sigma^2 = 0$. One might suspect that if the starting value of σ^2 were near zero, the

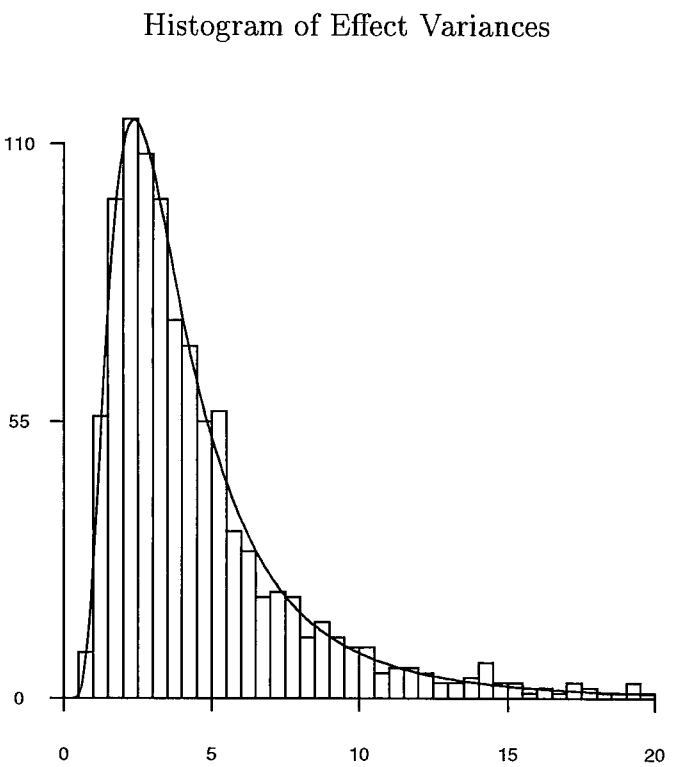

Figure 1: Histogram of the 1000 values of the effect variance from the null Gibbs chain, that is, a histogram of $\sigma^{2(j+15{,}000)}$ for $j = 1, 2, \ldots, 1000$. Superimposed is the approximate (supposed) marginal posterior density of σ^2. An appropriately scaled version of $\hat{\pi}_{\sigma^2|\mathbf{y}}(t|\mathbf{y})$ is on the ordinate with t on the abscissa. (Actually, 15 of the 1,000 values of the effect variance, ranging from 21.0 to 45.1, were not included in the histogram.)

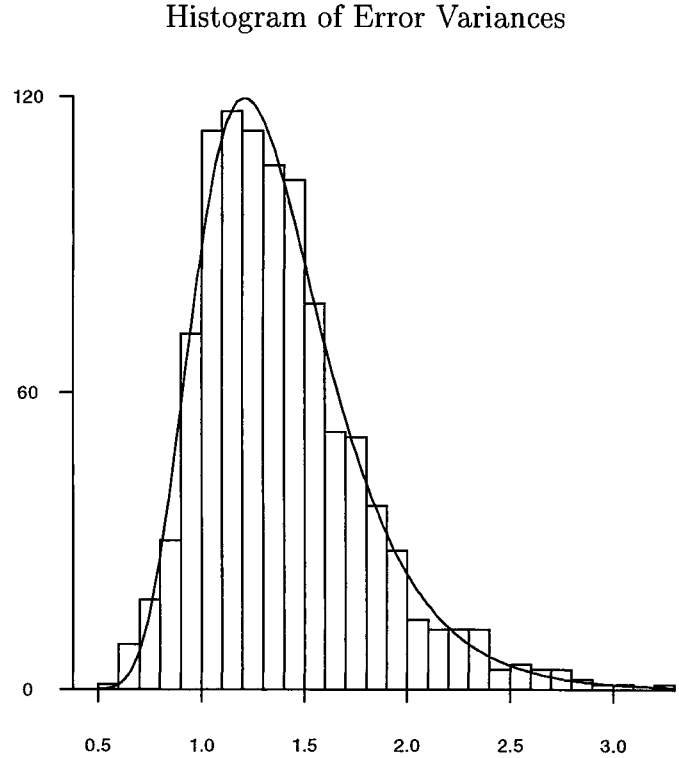

Figure 2: Histogram of the 1000 values of the error variance from the null Gibbs chain, that is, a histogram of $\sigma_\epsilon^{2(j+15{,}000)}$ for $j = 1, 2, \ldots, 1000$. Superimposed is the approximate (supposed) marginal posterior density of σ_ϵ^2. An appropriately scaled version of $\hat{\pi}_{\sigma_\epsilon^2|\mathbf{y}}(t|\mathbf{y})$ is on the ordinate with t on the abscissa.

σ^2 component of the Gibbs chain would be absorbed at 0. This is not the case, however. In fact, the σ^2 component and the random effects components move towards zero, but eventually they all return to a reasonable part of the space. For example, we started the chain with $\sigma^2 = 10^{-50}$ and after 20,000 iterations the σ^2 component was approximately 10^{-122} and the largest magnitude of any of the random effects components was about 10^{-60}. The chain was allowed to run for a total of one million iterations, after which all of the components were back in a reasonable part of the parameter space. This Gibbs chain behaves somewhat like one constructed with the exponential conditionals of Example 3 in that it leaves the "center" of the space for long periods of time, but eventually returns. Such behavior is consistent with null recurrence. ||

Lastly, we note that it seems virtually impossible to detect a null chain with a diagnostic measure. Standard "convergence diagnostics" proposed in the MCMC literature (see, for example, Raftery and Banfield 1991, Gelman and Rubin 1992, Roberts 1992, Tanner 1993, and Robert 1995) assume that the chain is positive recurrent and use the output to provide information about when Monte Carlo approximations are "close enough" to the true values. They are not designed to detect if the Gibbs chain converges (positive recurrence), nor even when the Gibbs chain has converged; as it never does. Thus, one should not count on "convergence diagnostics" to detect an improper posterior.

4 Decision Theory and Algorithms

Now that we have looked at the effect of the algorithm on the statistical inference, we will somewhat turn things around and look at the effect of statistical theory on the output from the algorithm. We can consider a Monte Carlo algorithm as outputting data about an underlying process, with the goal being the construction of an estimate of some feature of the process. In this light, we can ask how to best process the data, and answer that question by applying statistical principles. In what follows, we apply one of the simplest principles, that of Rao-Blackwellization, to the output of an Accept-Reject Algorithm. For more details, including applications to the Metropolis-Hastings Algorithm, see Casella and Robert (1995, 1996abc).

4.1 The Accept-Reject Algorithm as a Statistical Model

The Accept-Reject algorithm is based on the following lemma.

Lemma 1 *If f and g are two densities, and there exists $M < \infty$ such that $f(x) \leq Mg(x)$ for every x, the random variable X provided by the algorithm*

1. *Simulate $Y \sim g(y)$;*

2. *Simulate* $U \sim \mathcal{U}[0,1]$ *and take* $X = Y$ *if* $U \leq f(Y)/Mg(Y)$*; otherwise, repeat step 1.*

Then X *is distributed according to* f*.*

When viewed statistically, we have the following description of the algorithm. A sequence $Y_1, Y_2, \ldots$ of independent random variables is generated from g along with a corresponding sequence $U_1, U_2, \ldots$ of uniform random variables. Given a function h, the Accept-Reject estimator of $\tau = E\{h(X)\}$, based upon a sample $X_1, \ldots, X_t$ generated according to Lemma 1, is given by

$$\delta_{AR} = \frac{1}{t} \sum_{i=1}^{t} h(X_i). \tag{9}$$

Note that, conditional on the value t, the random variables $X_1, \ldots, X_t$ represent an *iid* sample from the distribution f. The Accept-Reject algorithm is usually implemented with a prespecified value of t, and the number of generated Y_i's is a random integer N satisfying

$$\sum_{i=1}^{N} I(U_i \leq w_i) = t \qquad \text{and} \qquad \sum_{i=1}^{N-1} I(U_i \leq w_i) = t - 1,$$

where we define $w_i = f(Y_i)/Mg(Y_i)$.

When we evaluate δ_{AR} as an estimator of τ, we see an estimator that

1. *Is based on extraneous information (the uniform random variables).*
2. *Is, in fact, a randomized estimator, that scourge of statistics.*

Classical statistical theory tells us that

1. *We need an estimator that does not depend on the observed values of the uniform random variables.*
2. *If an estimator is constructed by averaging over the uniform random variables, such an estimator will dominate* δ_{AR} *by the Rao-Blackwell theorem.*

It is straightforward to "Rao-Blackwellize" δ_{AR} by noting that it can be written

$$\delta_{AR} = \frac{1}{t} \sum_{i=1}^{N} I(U_i \leq w_i) h(Y_i), \tag{10}$$

so the conditional expectation

$$\delta_{RB} = \frac{1}{t} E \left\{ \sum_{i=1}^{N} I(U_i \leq w_i) h(Y_i) \middle| N, Y_1, \ldots, Y_N \right\} \tag{11}$$

improves upon (10) by the Rao-Blackwell Theorem.

Details of this calculation are carried out in Casella and Robert (1996a), where it is established that

$$\delta_{RB} = \frac{1}{t}\sum_{i=1}^{n} \rho_i h(Y_i) \tag{12}$$

where, for $i = 1, \cdots, n-1$, ρ_i satisfies

$$\begin{aligned} \rho_i &= \mathrm{P}(U_i \leq w_i | N = n, Y_1, \ldots, Y_n) \\ &= w_i \frac{\sum_{(i_1,\ldots,i_{t-2})} \prod_{j=1}^{t-2} w_{i_j} \prod_{j=t-1}^{n-2}(1-w_{i_j})}{\sum_{(i_1,\ldots,i_{t-1})} \prod_{j=1}^{t-1} w_{i_j} \prod_{j=t}^{n-1}(1-w_{i_j})}, \end{aligned} \tag{13}$$

while $\rho_n = 1$. The numerator sum is over all subsets of $\{1, \ldots, i-1, i+1, \ldots, n-1\}$ of size $t-2$, and the denominator sum is over all subsets of size $t-1$. The resulting estimator δ_{RB} is an average over all the possible permutations of the realized sample, the permutations being weighted by their probabilities. The Rao-Blackwellized estimator is then a function only of $(N, Y_{(1)}, \ldots, Y_{(N-1)}, Y_N)$, where $Y_{(i)}$ denotes the order statistics.

Because of the identity

$$\mathrm{var}(\delta) = \mathrm{var}[E\{\delta(U,Y)|Y\}] + E[\mathrm{var}\{\delta(U,Y)|Y\}]. \tag{14}$$

we see that the improvement that δ_{RB} brings over δ_{AR} is related to the size of $E[\mathrm{var}\{\delta(U,Y)|Y\}]$. This latter quantity can be interpreted as measuring the average variance in the estimator that is due to the auxiliary randomization, that is, the variance that is due to the uniform random variables. In some cases this quantity can be substantial.

Example 5 The target distribution is a Gamma distribution $\mathcal{G}(\alpha, \beta)$ with $\alpha > 1$. We set $\beta = 2\alpha$ so that the mean of the distribution is $1/2$. The candidate distribution we select is the Gamma $\mathcal{G}(a, b)$ distribution with $a = [\alpha]$ and $b = \beta a/\alpha$. We require $a < \alpha$ in order for M in Lemma 1 to be finite. The choice $b = 2a$ improves the fit between the two distributions since both means match. We consider two cases which reflect different acceptance rates for the Accept-Reject algorithm. In Case 1 we set $\alpha = 2.434$, $a = 2$ and $1/M = 0.9$ and, in Case 2, $\alpha = 20.62$, $a = 2$ and $1/M = 0.3$.

For each case we estimate the mean, chosen to be $1/2$ using both the simple Accept-Reject algorithm and its Rao-Blackwellized version. We also include mean squared error estimates for the Accept-Reject estimator and the improvement brought by Rao-Blackwellizing. This improvement is measured by the percentage decrease in mean squared error. From the table, it can be seen that the Rao-Blackwellisation provides a substantial decrease in mean squared error, reaching 60% in the case where the acceptance rate of the algorithm is

Table 1: Estimation of a gamma mean, chosen to be 1/2, using the Accept-Reject Algorithm, based on 7,500 simulations.

		Acceptance rate .9		
AR Sample Size	AR Estimate δ_{AR}	RB Estimate δ_{RB}	AR MSE	Percent Decrease in MSE
10	.5002	.5007	.0100	17.02
25	.5001	.4999	.0041	18.64
50	.4996	.4997	.0020	20.81
100	.4996	.4997	.0010	21.45
		Acceptance rate .3		
AR Sample Size	AR Estimate δ_{AR}	RB Estimate δ_{RB}	AR MSE	Percent Decrease in MSE
10	.5005	.5004	.0012	52.85
25	.4997	.5000	.0005	58.62
50	.4998	.5001	.0002	60.49
100	.4995	.5001	.0001	61.60

0.3. The improvement is better at the lower Accept-Reject acceptance rate partially because the Rao-Blackwellized sample is about three times bigger, with approximately two thirds of the sample being discarded by the Accept-Reject algorithm. Another interesting observation is that the percent improvement in mean squared error remains constant as the Accept-Reject sample size increases, implying that the variance of the original Accept-Reject estimator does not approach the variance of the Rao-Blackwellized estimator even as the sample size increases. We will return to this point in Section 5.2.

||

Computation of the ρ_i's of (13) can be accomplished with a recursion relation, and will typically require a calculation of order n^2. This may represent, to some, an unacceptable increase in computation time given the size of the anticipated decrease in mean squared error. To somewhat address this point, in Casella and Robert(1996b) we considered a simpler version of the Rao-Blackwell strategy that led to (12). Notice that, in what follows, we will simultaneously decrease computational complexity and increase statistical complexity.

4.2 Termwise Rao-Blackwellization

Starting from the Accept-Reject estimator (10), rather than calculating the full conditional expectation, we can instead calculate the termwise conditional expectation. This accomplishes the goal of removing the uniform random variables but retains computational simplicity.

To calculate the termwise conditional expectation of (10), conditioning the ith term on (N, Y_i), we need the conditional distribution of $U_i|, Y_i, N = n$. Although the original random variables are independent, the Accept-Reject algorithm stopping rule introduces a dependence in the sample. For example, for $i = 1, \ldots, n-1$ the marginal distribution of Y_i is

$$m(y) = \frac{t-1}{n-1} f(y) + \frac{n-t}{n-1} \frac{g(y) - \frac{1}{M} f(y)}{1 - \frac{1}{M}} \tag{15}$$

and Y_n has marginal distribution $f(y)$. It then can be shown that the resulting estimator, δ_{TRB} is given by

$$\begin{aligned} \delta_{TRB} &= \frac{1}{t} \sum_{i=1}^{n} E[I(U_i \leq w_i)|Y_i] h(Y_i) \qquad (16) \\ &= \frac{1}{t} \left(h(y_n) + \sum_{i=1}^{n-1} b(y_i) h(y_i) \right), \end{aligned}$$

where

$$b(y_i) = E[I_{U_i \leq \omega_i}|y_i] = \frac{t-1}{n-1} \frac{f(y_i)}{m(y_i)}, \quad i = 1, \cdots, n-1. \tag{17}$$

See the Appendix for details of these calculations.

We now have a seemingly reasonable estimator that is not complicated to compute, but its statistical properties are not as easy to establish as the full Rao-Blackwellized estimator (12). In fact, the Rao-Blackwell theorem does not apply to the estimator (16) because we did not condition on the entire estimator. To establish dominance of δ_{TRB} of (16) over δ_{AR} of (9), we must calculate the variance of δ_{TRB}, which involves $n(n-1)/2$ covariance terms. Moreover, it can easily be seen that δ_{TRB} cannot dominate δ_{AR} in mean squared error. This is because the sum of the weights in (17) is random, and if the target function $h(\cdot)$ is a nonzero constant function, δ_{TRB} will not estimate it correctly, while δ_{AR} will. This major difficulty is also common to some importance sampling schemes and prohibits uniform domination results there. A solution to this drawback is to force the estimators to estimate constant functions correctly, which can be achieved by dividing the weights $b(y_i)$ by their sum, thus replacing δ_{TRB} by its *rescaled* version

$$\delta_{Tr} = \frac{1}{t}(h(y_n) + \frac{t-1}{t} \left(\sum_{i=1}^{n-1} \frac{b(y_i)}{\sum_{j=1}^{n-1} b(y_j)} h(y_i) \right). \tag{18}$$

Table 2: Estimation a gamma mean, chosen to be $1/2$, using rescaled estimators from the Accept-Reject Algorithm, based on $7,500$ simulations.

		Acceptance rate .9		
AR Sample Size	% Dec. in MSE δ_{TRB}	% Dec. in MSE δ_{Tr}	% Dec. in MSE δ_{ISr}	% Dec. in MSE δ_{RB}
10	14.01	16.88	20.27	17.03
25	14.67	18.45	20.04	18.64
50	17.48	20.77	21.68	20.81
100	18.11	21.37	21.50	21.45
		Acceptance rate .3		
AR Sample	% Dec. in MSE	% Dec. in MSE	% Dec. in MSE	% Dec. in MSE
Size	δ_{TRB}	δ_{Tr}	δ_{ISr}	δ_{RB}
10	-259.62	53.76	54.07	52.85
25	-277.80	59.04	59.23	58.62
50	-272.18	60.73	60.78	60.49
100	-281.77	61.82	61.91	61.82

Such rescalings seem common in practice, despite any concern about the effect of introducing a bias in the estimator. Such concerns need not cause worry, however, as the bias induced by this rescaling is of an higher order than the variance (Casella and Robert 1996b). The following theorem can then be established.

Theorem 3 *For every function h, δ_{Tr} asymptotically dominates δ_{AR} in terms of quadratic risk. More precisely, as $t \to \infty$, if $N = O_p(t)$ then,*

$$E[(\delta_{Tr} - \tau)^2] \leq E[(\delta_{AR} - \tau)^2],$$

where $\tau = E[h(X)]$.

Moreover, the size of the improvement brought about by the rescaled estimator is truly impressive.

Example 6 (Continuation of Example 5) . Table 2 gives MSE reductions for the rescaled estimator δ_{Tr}, along with a rescaled importance sampling estimator and the full Rao-Blackwellized estimator (12). For comparison, we included in Table 2 a rescaled *importance sampling* estimator, derived as follows. A typical importance sampling estimator is of the form

$$\delta_{IS} = \frac{1}{n}\sum_{i=1}^{n} \frac{f(y_i)}{g(y_i)} h(y_i), \tag{19}$$

which would be unbiased under a random sampling scheme. However, the Accept-Reject Algorithm renders (19) biased. More importantly, (19) is not correct for constants, and will suffer from the same problems as δ_{TRB}. We thus want to rescale δ_{IS}, which results in the rescaled importance sampling estimator

$$\delta_{ISr} = \frac{1}{t}(h(y_n) + \frac{t-1}{t}\left(\sum_{i=1}^{n-1} \frac{f(y_i)/g(y_i)}{\sum_{j=1}^{n-1} f(y_j)/g(y_j)} h(y_i)\right). \tag{20}$$

The last observation comes from the correct density, and doesn't have to be reweighted. The remaining $n-1$ terms are rescaled. As it turns out, this estimator performs quite well in our simulation studies. This is really no surprise, as it is very close to the rescaled termwise Rao-Blackwell estimate.

There are a number of interesting points to notice about Table 2. First, termwise conditional expectation can actually make things worse, as δ_{TRB} increases the MSE over δ_{AR}. Although we knew that δ_{TRB} could not dominate for constant functions, the numerical example shows that even for more variable functions there may not be dominance.

The second striking thing to notice is that the improvement from the rescaled estimators δ_{Tr} and δ_{ISr} is actually better than that of the Rao-Blackwellized estimator δ_{RB}. This, no doubt, represents a favorable variance/bias trade-off, but is still quite startling. The decrease in computation time of δ_{Tr} and δ_{ISr} over δ_{RB} can be quite substantial, and the fact that mean squared error is improved really underscores the power of rescaling. ||

It is interesting to note that the rescaling idea, making the weights sum to one, arose naturally as "the right thing to do", especially in light of the performance of the estimators when $h(\cdot)$ is constant. Many times we notice, or intuit, empirical adjustments that help in certain cases. We can use the structure of decision theory to formalize our intuition, and see if the empirical improvements will, in fact, be useful in a wide variety of cases. Here we see that the value of the rescaling is confirmed by the decision-theoretic calculation of Theorem 3 and a simulation study. We thus have a nice interplay between using our intuition to construct an what we think is an improved estimator, and using theory to establish that we have, in fact, done so.

5 Other Considerations

In this section we review some recent work that further explores the structure of Monte Carlo algorithms, particularly the Gibbs sampler. The goals of these investigations are to understand how to better, or even optimally, process the

output of the algorithm, and also to use the structure of the algorithm to help construct optimal procedures. It is interesting to note that both frequentist and Bayesian inferences benefit in the following examples. Unfortunately, these illustrations are somewhat less detailed, as some of the work is still in progress

5.1 Constructing the Inference from the Algorithm

An endpoint of a Gibbs sampler is typically a sample from a posterior distribution $\pi(\theta|\mathbf{y})$, a distribution which may itself be intractable to work with. If a confidence set, or more specifically, a credible set, for θ is desired, we may have to solve a difficult integral equation where the integrand may not be expressible in closed form. Specifically, suppose that we have a pair of conditional posterior densities $\pi(\theta|\mathbf{y},\lambda)$ and $\pi(\lambda|\mathbf{y},\theta)$ in a Gibbs sampler Markov chain, and we are interested in inferences about $\pi(\theta|\mathbf{y})$. If we use the Gibbs sampler to generate the pairs $(\theta_i,\lambda_i), i = 1,2,\cdots$, then, from the ergodic theorem, $\pi(\theta|\mathbf{y}) = \lim_{m\to\infty}(1/m)\sum_{i=1}^{m}\pi(\theta|\mathbf{y},\lambda_i)$. Suppose that, for a specified value of α, we are interested in finding the value a^* such that $\int_{-\infty}^{a^*}\pi(\theta|\mathbf{y})d\theta = \alpha$, a lower confidence bound. A first approach would be to solve for a^* in

$$\frac{1}{m}\int_{-\infty}^{a^*}\sum_{i=1}^{m}\pi(\theta|\mathbf{y},\lambda_i)d\theta = \alpha.$$

As this calculation could be quite involved, we ask if the value a^* can be constructed from the Gibbs sequence (θ_i,λ_i) in any simple way?

A first approach on the problem, developed in Eberly (1997), is the following. Writing $\Pi(\cdot)$ for a distribution function, for example, $\Pi(a|\mathbf{y}) = \int_{-\infty}^{a}\pi(\theta|\mathbf{y})d\theta$, calculate for each λ_i a value a_i such that $\Pi(a_i|\mathbf{y},\lambda_i) = \gamma$, where the value of γ will be determined shortly. (Note that in a typical Gibbs sampler, the full conditionals are usually very nice densities, so solving for the a_is should be very quick.) Now $\frac{1}{m}\sum_{i=1}^{m}a_i = \bar{a} \to a'$, for some value a', but it is not necessarily the case that $a' = a^*$. However, expanding $\Pi(a_i|\mathbf{y},\lambda_i)$ in a Taylor series around $\bar{a}$ yields

$$\Pi(a_i|\mathbf{y},\lambda_i) \approx \Pi(\bar{a}|\mathbf{y},\lambda_i) + (a_i - \bar{a})\pi(\bar{a}|\mathbf{y},\lambda_i).$$

Now sum both sides, and remember that $\Pi(a_i|\mathbf{y},\lambda_i) = \gamma$ to get

$$\gamma \approx \frac{1}{m}\sum_{i=1}^{m}\Pi(\bar{a}|\mathbf{y},\lambda_i) + \frac{1}{m}\sum_{i=1}^{m}(a_i - \bar{a})\pi(\bar{a}|\mathbf{y},\lambda_i).$$

It can be established that $\frac{1}{m}\sum_{i=1}^{m}\Pi(\bar{a}|\mathbf{y},\lambda_i) \to \Pi(a'|\mathbf{y})$, so we have the approximation

$$\Pi(a'|\mathbf{y}) \approx \gamma - \frac{1}{m}\sum_{i=1}^{m}(a_i - \bar{a})\pi(\bar{a}|\mathbf{y},\lambda_i),$$

which suggests setting $\gamma = \alpha + \frac{1}{m}\sum_{i=1}^{m}(a_i - \bar{a})\pi(\bar{a}|\mathbf{y}, \lambda_i)$, with the hope that $\bar{a} \to a' \approx a^*$.

This linear approximation seems to perform adequately in some situations, but can be improved upon by a quadratic Taylor series approximation. Further work, in understanding the value and limitations of this approximation, and thoroughly developing the theory, is presently being done.

5.2 The Effect of Rao-Blackwellization

In Section 4.1 we alluded to the fact that Rao-Blackwellization will always result in an appreciable variance reduction, even as the sample size (or the number of Monte Carlo iterations) increases. To address this point more precisely, consider the work of Levine (1996), who formulated this problem in terms of the asymptotic relative efficiency (ARE) of $\delta_0 = (1/m)\sum h(X_i)$ with respect to its Rao-Blackwellized version $\delta_1 = (1/m)\sum E[h(X_i)|Y_i]$, where the pairs (X_i, Y_i) are generated from a Gibbs sampler with $X_i \sim f(x|Y_{i-1})$ and $Y_i \sim f(y|X_i)$. (Levine 1996 considers more complex Gibbs samplers, but we will only use this simple case for illustration. The key property that the sampler need have is reversibility.) The ARE is a ratio of the variances of the limiting distribution for the two estimators, which are given by

$$\sigma^2_{\delta_0} = \text{var}(h(X)) + 2\sum_{k=1}^{\infty}\text{cov}(X_0, X_k) \tag{21}$$

and

$$\sigma^2_{\delta_1} = \text{var}(E[h(X)|Y]) + 2\sum_{k=1}^{\infty}\text{cov}(E[X_0|Y_0], E[X_k|Y_k]). \tag{22}$$

Levine then proves the following theorem.

Theorem 4 *If a sample $\{(X_i, Y_i)\}_{i=0}^{n}$ is generated by the bivariate Gibbs sampler, then for all $h(\cdot)$ with finite variance, the ratio $\sigma^2_{\delta_0}/\sigma^2_{\delta_1} \geq 1$, with equality if and only if $var(h(X)) = var(E[h(X)|Y]) = 0$.*

To see the amount of possible improvement, consider the following example.

Example 7 Let

$$(X, Y)' \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

where $-1 < \rho < 1$. Assume interest lies in estimating $\mu = E(X)$. The Gibbs sampler can obtain samples from the bivariate normal distribution by alternately drawing random variables from

$$\begin{aligned} X \mid Y &\sim N(\rho Y,\ 1 - \rho^2) \\ Y \mid X &\sim N(\rho X,\ 1 - \rho^2). \end{aligned}$$

It can be shown that $\text{cov}(X_1, X_k) = \rho^{2k}$, for all k, and

$$\sigma^2_{\delta_0}/\sigma^2_{\delta_1} = \frac{1}{\rho^2} > 1.$$

So, if δ_1 is less than $1/\rho^2$ times more complex that δ_0, then δ_1 should be used. Since $\text{E}(X \mid Y) = \rho Y$, it takes $n+2$ floating point operations (flops) to compute $\delta_1 = (1/n)\sum_{k=0}^{n} E(X \mid Y_k)$ as compared to $n+1$ flops to compute $\delta_0 = (1/n)\sum_{k=0}^{n} X_k$. Therefore, the cost of computation, in terms of flops, is essentially the same, but there can be a vast gain in precision by using δ_1. ||

5.3 Minimax Gibbs Samplers

An interesting example of the interplay between decision theory and Monte Carlo algorithms is given by the problem of optimizing the random scan Gibbs sampler (see, for example, Rosenthal 1995, Amit 1996, Roberts and Sahu 1996) The random scan Gibbs sampler is characterized by selection probabilities $\alpha_1, \ldots, \alpha_d$. These probabilities determine the percentage of visits to a specific site or component of the $d \times 1$ vector of interest $\mathbf{X} = (X_1, \ldots, X_d)$ during a run of the sampler. A standard approach is to choose the selection probabilities to provide the sampling strategy with the smallest convergence rate. However, choosing the selection probabilities according to such a criterion may be undesirable in practice. For example, the convergence rate is not only typically difficult to compute and possibly mathematically intractable, but also may also ignore important features of the target distribution necessary for determining the optimal random scan, as we will see below.

Levine (1996) considers an alternative measure derived from statistical decision theoretic considerations, which seems to provide an attractive criterion for choosing an appropriate random scan. Assume a random $d \times 1$ vector $\mathbf{X}$ is generated by a random scan Gibbs sampler which generates a Markov chain $\{\mathbf{X}^{(i)}\}_{i=1}^{n}$ with stationary distribution π. Suppose interest lies in estimating $\mu = E_\pi(h(\mathbf{X}))$ where $\text{var}(h(\mathbf{X})) < \infty$. If we estimate μ with the sample mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} h(\mathbf{X}^{(i)})$, the optimal mean squared error scan is the one that minimizes the risk

$$R^{(n)}(\boldsymbol{\alpha}, h) = E_\pi\left[(\hat{\mu} - \mu)^2\right] = \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} h(\mathbf{X}^{(i)})\right). \tag{23}$$

Alternatively, we may consider the asymptotic risk

$$\begin{aligned} R(\boldsymbol{\alpha}, h) &= \lim_{n\to\infty} nR^{(n)}(\boldsymbol{\alpha}, h) \\ &= \text{var}\left(h(\mathbf{X})\right) + 2\lim_{n\to\infty}\sum_{i=1}^{n-1}\left(\frac{n-i}{n}\right)\text{cov}\left(h(\mathbf{X}^{(0)}), h(\mathbf{X}^{(i)})\right) \end{aligned} \tag{24}$$

$$= \quad \text{var}\left(h(\mathbf{X})\right) + 2 \lim_{n\to\infty} \sum_{i=1}^{\infty} \text{cov}\left(h(\mathbf{X}^{(0)}), h(\mathbf{X}^{(i)})\right)$$

as a basis for choosing a random scan.

We note that the convergence rate of the random scan, the norm of the forward operator, can be expressed as

$$\lambda_{RS}^2 = \sup_h \text{cov}\left(h(\mathbf{X}^{(0)}), h(\mathbf{X}^{(2)})\right),$$

where the supremum is over all functions with finite variance. Thus we see that, when compared to (24), the convergence rate contains less information about the variance and covariances of the chain. It is in this sense that we feel that (24) is a better optimality measure.

To use (24) as a criterion for selecting a scan, we would like it to produce a reasonable scan for any function h. This suggests that we might want to protect against the worst possible function h, with finite variance, by minimizing the maximum risk $sup_h R(\boldsymbol{\alpha}, h)$. Levine (1996) develops a method for doing this, implementing an adaptive scan of the state space. That is, at each iteration the selection probabilities are updated via a sequence of sample points from the previous iteration, and may even use information from past iterations (which could destroy the Markov nature of the chain). However, the chain does converge, approaching the optimal random scan according to (24). Levine also discusses examples where this procedure can be implemented, however full implementation in a general setting is presently too computationally intensive to be useful. Approximations are being investigated for these cases.

6 Discussion

Even though we have covered a lot of ground in understanding the interplay between statistical theory and computational algorithms, there is an enormous amount of work that we have not mentioned. We only alluded to the fundamental papers of Liu, Wong and Kong (1994, 1995), which provide an elegant and comprehensive treatment of the structure of the Gibbs sampler. Other work, such as Tanner and Wong (1987), Liu (1994), Tierney (1994) or Robert (1995), illustrates how statistical theory interfaces with Monte Carlo algorithms, most notably the Gibbs sampler and the Metropolis algorithm.

The other body of work we have not discussed is that which deals with missing data problems, using techniques such as the EM algorithm. Although EM and Gibbs share a similar underpinning, (see Casella and Berger 1995 for a view of the EM algorithm as a Gibbs sampler) they tend to be used in somewhat different ways. However, research in these methods, which also combines statistical theory with the computational algorithms, continues to flourish; see for

example Smith and Roberts (1993), Meng and Rubin (1993), Liu and Rubin (1994), Meng (1994), Besag et al. (1995) and Meng and van Dyk (1996).

The message of this paper, which by now may be obscured in these sometime incoherent ramblings, is one that bears repeating. What we have done is to approach a new methodology, that of iterative Monte Carlo calculation, with the standard tools of the theoretical statistician. What resulted are procedures whose output and performance have been optimized from a statistical view. It sometimes may happen, as with the Rao-Blackwellized Estimator of (12), or Section 5.3, that a statistically optimal answer may result in a difficult, or even prohibitive computational burden. In such cases, statistical theory, in particular decision theory, can still provide answers. It them becomes a matter of specifying an alternate optimality criterion, or loss function, to take these other matters into account.

7 Appendix: The Termwise Weights

To calculate the weights for the termwise Rao-Blackwellized estimator (16), it is necessary to derive the distribution of the uniform random variable conditional on the generated value of the candidate random variable. This is a rather straightforward exercise in distribution theory, and is only made complicated by the stopping rule of the Accept-Reject Algorithm.

From the Accept-Reject Algorithm of Lemma 1, we get a sequence $Y_1, Y_2, \ldots$ of independent random variables generated from g along with a corresponding sequence $U_1, U_2, \ldots$ of uniform random variables. For a fixed sample size t, i.e. for a fixed number of accepted random variables, the number of generated Y_i's is a random integer N. The joint distribution of $(N, Y_1, \ldots, Y_N, U_1, \ldots, U_N)$ is given by

$$\begin{aligned} &P(N = n, Y_1 \le y_1, \ldots, Y_n \le y_n, U_1 \le u_1, \ldots, U_n \le u_n) \\ &= \int_{-\infty}^{y_n} g(t_n)(u_n \wedge w_n)dt_n \int_{-\infty}^{y_1} \cdots \int_{-\infty}^{y_{n-1}} g(t_1)\ldots g(t_{n-1}) \qquad (25) \\ &\quad \times \sum_{(i_1,\ldots,i_{t-1})} \prod_{j=1}^{t-1}(w_{i_j} \wedge u_{i_j}) \prod_{j=t}^{n-1}(u_{i_j} - w_{i_j})^+ dt_1 \ldots dt_{n-1}, \end{aligned}$$

where $w_i = f(y_i)/Mg(y_i)$ and the sum is over all subsets of $\{1, \ldots, n-1\}$ of size $t-1$.

We next want to get the joint distribution of $(Y_i, U_i)|N = n$, for any $i = 1, \cdots, n-1$. Since this distribution is the same for each of these values of i, we can just derive it for (Y_1, U_1). Recall that $Y_n \sim f$.

If we set $y_1 = y$, $u_1 = u$, $y_2 = y_3 = \cdots = y_n = \infty$ and $u_2 = u_3 = \cdots = u_n = 1$, we can derive the joint distribution of (N, Y_1, U_1). Assume, without loss of generality, that $\lim_{y\to\infty} f(y)/g(y) = 1$. (If this is not the case, we just have to

adjust the constant M in what follows). Then, aside from the pair (w_1, u_1), we have $(w_{i_j} \wedge u_{i_j}) = \frac{1}{M}$ and $(u_{i_j} - w_{i_j})^+ = \left(1 - \frac{1}{M}\right)$, hence

$$\begin{aligned}
&\sum_{(i_1,\ldots,i_{t-1})} \prod_{j=1}^{t-1} (w_{i_j} \wedge u_{i_j}) \prod_{j=t}^{n-1} (u_{i_j} - w_{i_j})^+ = \\
&= (w_1 \wedge u_1) \binom{n-2}{t-2} \left(\frac{1}{M}\right)^{t-2} \left(1 - \frac{1}{M}\right)^{n-t} \\
&\quad + (u_1 - w_1)^+ \binom{n-2}{n-t-1} \left(\frac{1}{M}\right)^{t-1} \left(1 - \frac{1}{M}\right)^{n-t-1}.
\end{aligned} \tag{26}$$

Noting that

$$\binom{n-2}{t-2} = \frac{t-1}{n-1}\binom{n-1}{t-1}, \quad \binom{n-2}{n-t-1} = \frac{n-t}{n-1}\binom{n-1}{t-1},$$

and $\int_{-\infty}^{y_n} g(t_n)(u_n \wedge w_n) dt_n = \int_{-\infty}^{\infty} g(t_n) \left(\frac{1}{M}\right) dt_n = \frac{1}{M}$, we have

$$\begin{aligned}
&P(N = n, Y_1 \le y, U_1 \le u) = \\
&= \binom{n-1}{t-1} \left(\frac{1}{M}\right)^{t-1} \left(1 - \frac{1}{M}\right)^{n-t-1} \\
&\quad \times \left[\frac{t-1}{n-1}(w_1 \wedge u_1)\left(1 - \frac{1}{M}\right) + \frac{n-t}{n-1}(u_1 - w_1)^+ \left(\frac{1}{M}\right)\right] \\
&\quad \times \int_{-\infty}^{y} g(t_1) dt_1.
\end{aligned} \tag{27}$$

From (27) we can immediately get the negative binomial marginal distribution of N, $P(N = n) = \binom{n-1}{t-1} \left(\frac{1}{M}\right)^t \left(1 - \frac{1}{M}\right)^{n-t}$, the marginal distribution of Y_1, $m(y)$ of (15) and, most importantly, we get the conditional distribution of $U_1 | Y_1, N$ and can calculate

$$P(U_1 \le w(y) | Y_1 = y, N = n) = \frac{g(y) w(y) M \frac{t-1}{n-1}}{m(y)}, \tag{28}$$

which is the same as $b(y_i)$ of (17).

References

1. Amit, Y. (1996). Convergence Properties of the Gibbs Sampler for Perturbations of Gaussians. To appear in *The Annals of Statistics.*

2. Arnold, B. C., and Press, S. J. (1989). Compatible Conditional Distributions. *Journal of the American Statistical Association,* **84**, 152–156.

3. 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.

4. Besag, J., Green, P., Higdon, D. and Mengersen, K. (1995). Bayesian Computation and Stochastic Systems (with discussion). *Statistical Science* **10** 1-66.

5. Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis, Second Edition.* New York: Springer-Verlag.

6. Casella, G., and Berger, R. L. (1994). Estimation with Selected Binomial Information or, Do You Really Believe that Dave Winfield is Batting .471? *Journal of the American Statistical Association,* **89** 1080–1090.

7. Casella, G., and George, E. I. (1992). 'Explaining the Gibbs Sampler. *The American Statistician,* **46** 167–174.

8. Casella, G. and Robert, C. P. (1995). Une Implémentation de Théorème de Rao-Blackwell en Simulation avec Rejet. *C. R. Acad. Sci. Paris* **322** *Série 1* 571-576.

9. Casella, G. and Robert, C. P. (1996a). Rao-Blackwellization of Sampling Schemes. *Biometrika* **83** 81-94.

10. Casella, G. and Robert, C. P. (1996b). Post-Processing Accept-Reject Samples: Recycling and Rescaling. Technical Report BU-1311-M, Biometrics Unit, Cornell University, Ithaca, NY, and Document de Travail 9625, Centre de Recherche en Économie et Statistique, INSEE, Paris.

11. Casella, G. and Robert, C. P. (1996c). Recycling Rejected Values in Accept-Reject Methods. *C. R. Acad. Sci. Paris* **321** *Série 1* 1621-1626.

12. Dey, D. K., Gelfand, A. E., and Peng, F. (1994. Overdispersed Generalized Linear Models. Technical Report, University of Connecticut, Dept. of Statistics.

13. Eberly, L. E. (1997). Constructing Confidence Statements from the Gibbs Sampler. Unpublished PhD Thesis, Biometrics Unit and Statistics Center, Cornell University, Ithaca NY.

14. Gelfand, A. E., Hills, S. E., Racine-Poon, A., and Smith, A. F. M. (1990). Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling. *Journal of the American Statistical Association* **85** 972–985.

15. 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.

16. Gelman, A., and Speed, T. P. (1993). Characterizing a Joint Probability Distribution by Conditionals. *Journal of the Royal Statistical Society*, Ser. B, **55** 185–188.

17. Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6 (6)** 721-741.

18. Geyer, C. (1992). Practical Markov Chain Monte Carlo. *Statistical Science* **7** 473-483.

19. Hastings, W. K. (1970). Monte Carlo Sampling using Markov Chains and Their Application. *Biometrika* **57** 77-109.

20. Hill, B. M. (1965). Inference about Variance Components in the One-Way Model. *Journal of the American Statistical Association* **60** 806–825.

21. Hobert, J. P. and Casella, G. (1995). Functional Compatibility, Markov Chains, and Gibbs Sampling with Improper Posteriors. Technical Report BU-1280-M, Biometrics Unit, Cornell University, Ithaca, NY.

22. Hobert, J. P. and Casella, G. (1996). The Effect of Improper Priors on Gibbs Sampling in Hierarchical Models. To appear in *Journal of the American Statistical Association.*

23. Ibrahim, J. G. and Laud, P. W. (1991). On Bayesian analysis of generalized linear models using Jeffreys's prior. *Journal of the American Statistical Association* **86**, 981-986.

24. Lehmann, E. L. and Casella, G. *Theory of Point Estimation, Second Edition.* New York: Springer-Verlag.

25. Levine, R. A. (1996). Optimizing Convergence Rates and Variances in Gibbs Sampling Schemes. Unpublished PhD Thesis, Biometrics Unit and Statistics Center, Cornell University, Ithaca NY.

26. Liu, J. (1994). The collapsed Gibbs sampler in Bayesian computation with application to a gene regulation problem. *Journal of the American Statistical Association* **89** 958-966.

27. Liu, C. and Rubin, D. B. (1994). The ECME algorithm: A Simple Extension of EM and ECM with Fast Monotone Convergence. *Biometrika* **81** 633-648.

28. Liu, J., 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.

29. Liu, J., Wong, W. H., and Kong, A. (1995). Correlation Structure and Convergence Rate of the Gibbs Sampler with Various Scans. *Journal of the Royal Statistical Society B* **57** 157-169.

30. Meng, X-L. (1994). On the Rate of Convergence of the ECM Algorithm. *Annals of Statistics* **22** 326-339.

31. Meng, X-L. and Rubin, D.B. (1993). Maximum Likelihood estimation via the ECM Algorithm: A General Framework. *Biometrika* **80** 267-278.

32. Meng, X-L. and van Dyk, D. (1996). (1993). THe EM Algorithm - An Old Folk Song Sung to a Fast New Tune. To appear in *Journal of the Royal Statistical Society Series B.*

33. Metropolis, M. Rosenbluth, A., Rosenbluth, m., Teller, A., and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics* **21** 1087-1092.

34. Natarajan, R., and McCulloch, C. E. (1995). A Note on the Existence of the Posterior Distribution for a Class of Mixed Models for Binomial Responses. *Biometrika* **82** 639–643.

35. Raftery, A. E., and Banfield, J. D. (1991). Stopping the Gibbs Sampler, the Use of Morphology, and Other Issues in Spatial Statistics. *Annals of the Institute of Statistical Mathematics* **43**, 32–43.

36. Robert, C. (1995). Convergence Control Methods for Markov Chain Monte Carlo Algorithms. *Statistical Science* **10** 231-253.

37. Roberts, G. O. (1992). Convergence Diagnostics of the Gibbs Sampler. In *Bayesian Statistics 4* (eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), Oxford: Oxford University Press, 775–782.

38. Roberts, G. O. and Sahu, S. K. (1996). Updating Schemes, Correlation Structure, Blocking and Parameterisation for the Gibbs Sampler. Technical Report, Statistical Laboratory, University of Cambridge.

39. Rosenthal, J. S. (1995). Rates of Convergence for Gibbs Sampling for Variance Component Models. *Journal of the American Statistical Association* **23** 740-761.

40. Searle, S.R., Casella, G. and McCulloch, C. E. (1992). *Variance Components.* New York: John Wiley.

41. 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 B* **55** 3-23.

42. Tanner, M. A. (1993). *Tools for Statistical Inference.* Springer-Verlag, New York.

43. Tanner, M. A. and Wong, W. (1987). The Calculation of Posterior Distributions by Data Augmentation (with discussion). *Journal of the American Statistical Association,* **82** 805-811.

44. Tierney, L. (1994). Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics* **22** 1701-1762.

45. Wang, C. S., Rutledge, J. J., and Gianola, D. (1993). Marginal Inferences about Variance Components in a Mixed Linear Model using Gibbs Sampling. *Genetique, Selection, Evolution* **25** 41–62.

46. Wang, C. S., Rutledge, J. J., and Gianola, D. (1994). Bayesian Analysis of Mixed Linear Models via Gibbs Sampling with an Application to Litter Size of Iberian Pigs. *Genetique, Selection, Evolution* **26** 1–25.

47. Zeger, S. L., and Karim, M. R. (1991). Generalized Linear Models With Random Effects; A Gibbs Sampling Approach. *Journal of the American Statistical Association* **86** 79–86.