Federal Reserve Bank of Minneapolis
Research Department Staff Report 148

December 1991

EVALUATING THE ACCURACY OF SAMPLING-BASED APPROACHES TO THE CALCULATION OF POSTERIOR MOMENTS

John Geweke*

Federal Reserve Bank of Minneapolis

and University of Minnesota

ABSTRACT

Data augmentation and Gibbs sampling are two closely related, sampling-based approaches to the calculation of posterior moments. The fact that each produces a sample whose constituents are neither independent nor identically distributed complicates the assessment of convergence and numerical accuracy of the approximations to the expected value of functions of interest under the posterior. In this paper methods for spectral analysis are used to evaluate numerical accuracy formally and construct diagnostics for convergence. These methods are illustrated in the normal linear model with informative priors, and in the Tobit-censored regression model.

Keywords and phrases: Data augmentation, Gibbs sampling, Mixed estimation, Monte Carlo integration, Tobit model

*This paper was prepared as an invited presentation at the Fourth Valencia International Meeting on Bayesian Statistics, Peñiscola, Spain, April 15-20, 1991. Financial support from National Science Foundation Grant SES-8908365 and research assistance from Zhenyu Wang are gratefully acknowledged. Software and data may be requested by electronic mail addressed to geweke@atlas.socsci.umn.edu.

1. Introduction

Gibbs sampling (Geman and Geman, 1984) in conjunction with data augmentation (Tanner and Wong, 1987) constitutes a very useful tool for the solution of many important Bayesian multiple integration problems (Gelfand and Smith, 1990). The method is complementary to other numerical approaches, and is attractive and competitive in many statistical models. This is consistent with the current rapid growth in the application of the method to interesting problems by many Bayesian statisticians. It produces samples of functions of interest whose distribution converges to the posterior distribution, but whose constituents are in general never independently distributed and are identically distributed only in the limit. In this context treatment of the compelling questions of convergence and the accuracy of approximation has been informal.

The research reported here undertakes to place these matters on a more formal footing, and in so doing render the Gibbs sampler with data augmentation a more systematic, reliable, and replicable tool for solution of the Bayesian multiple integration problem. Its thesis is that the failure of the Gibbs sampling process to produce independently distributed realizations of the functions of interest is no inhibitor to learning about the means of these sequences. Indeed, the problem of inference for the mean of a stationary process is one of long-standing in time series analysis, with a number of solutions. We apply one of these solutions here, showing how it yields systematic assessments of the numerical accuracy of approximations to the expected values of functions of interest under the posterior. With a formal distribution theory for the approximations in hand, it is straightforward to construct diagnostics for convergence.

The paper begins in Section 2 with a brief exposition of the Gibbs sampler and data augmentation, in the next section. The main methodological results are presented in Section 3, together with a simple example constructed to provide some insight into the typical serial correlation structure of the Gibbs sampler. In Section 4 the proposed solution is applied to the multiple normal linear regression model with a proper normal prior on a subset of its coefficients. The Gibbs sampler proves to be very efficient in treating this problem, which is important in its own right in econometrics and is a key building block in applying the Gibbs sampler with data augmentation to Bayesian inference in many other statistical models. In Section 5 this application is extended to the Tobit censored regression model. The formal treatment of convergence and numerical accuracy appears essential to producing reliable results with the Gibbs sampler and data augmentation in this model. Conclusions, together with conjectures about future research, are presented in the final section.

2. The Gibbs Sampler and Data Augmentation

The generic task in applied Bayesian inference is to obtain the expected value of a function of interest under a posterior density. In standard notation, this object is expressed

$$E[g(\theta)] = \int_{\Theta} g(\theta)\pi(\theta)L(\theta; X)d\theta \,/\, \int_{\Theta} \pi(\theta)L(\theta; X)d\theta,$$

where θ is the finite-dimensional vector of parameters whose domain is a subset of Euclidean space Θ; X is the observed data; $g(\theta)$ is the function of interest; $\pi(\theta)$ is proportional to a proper or improper prior density; and $L(\theta; X)$ is proportional to the likelihood function. In what follows, we shall suppress the dependence of the likelihood function on the observed data, and denote the kernel of the posterior density $p(\theta) = \pi(\theta)L(\theta)$. We shall refer to the calculation of $E[g(\theta)]$ as the Bayesian multiple integration problem.

2.1 Other Approaches

In some cases $p(\theta)$ and $g(\theta)p(\theta)$ are of sufficiently simple form that it is possible to obtain an exact analytical evaluation of $E[g(\theta)] = \int_{\Theta} g(\theta)p(\theta)d\theta / \int_{\Theta} p(\theta)d\theta$, but the class of such cases is much smaller than the class of problems routinely studied in statistics and econometrics. Zellner (1971) provides a treatment of many of these cases, and the class has not grown much in the 20 years since that volume was written. However, a rich variety of methods of approximating $E[g(\theta)]$ have emerged, and continues to broaden. Series expansions of $p(\theta)$ and $g(\theta)p(\theta)$ provide one basis for these methods, including LaPlace's method (Tierney and Kadane, 1986) and marginal inference (Leonard, Hsu and Tsu, 1989). Monte Carlo sampling from the parameter space Θ provides another line of attack, including importance sampling (Kloek and van Dijk, 1978; Geweke, 1989) and antithetic acceleration (Geweke, 1988). The various methods for solving the Bayesian multiple integration problem tend to be complementary. For example, series expansion methods lead to essentially instantaneous computations but provide approximations whose accuracy cannot easily be improved or systematically evaluated. Monte Carlo methods produce approximations whose accuracy is easily assessed and can be improved by increasing the number of iterations, but the computations may be quite time consuming. Both methods require preliminary analytical work, in the form of carrying out the series expansions or constructing the importance sampling density, that can be tedious.

The objective of this research is to facilitate the application of two other closely related Monte Carlo methods, which have become known as Gibbs sampling, and Gibbs sampling

with data augmentation. The utility of the Gibbs sampler, as proposed by Geman and Geman (1984), for the generic task in applied Bayesian inference was recognized by Gelfand and Smith (1990). The use of data augmentation in the calculation of posterior densities was proposed by Tanner and Wong (1987). The potential of combining the two is immediately evident.

2.2 The Gibbs Sampler

The Gibbs sampler provides a method for sampling from a multivariate probability density, employing only the densities of subsets of vectors conditional on all the others. The method is easily described in the context of our generic Bayesian problem. Suppose the parameter vector is partitioned, $\theta' = (\theta_{(1)}', \theta_{(2)}', \ \dots \ , \theta_{(s)}')$. Further suppose that the conditional distributions,

$$\theta_{(j)} \mid \{\theta_{(1)}, \dots , \theta_{(j-1)}, \theta_{(j+1)} \dots , \theta_{(s)}\} \sim p_{(j)}[\theta_{(1)}, \dots , \theta_{(j-1)}, \theta_{(j+1)}, \dots , \theta_{(s)}] \quad (j = 1, \dots , s)$$

are known, and are of a form that synthetic i.i.d. random variables can be generated readily and efficiently from each of the $p_{(j)}(\cdot)$. Now let $\theta^{(0)\prime} = (\theta_{(1)}^{(0)\prime}, \ \theta_{(2)}^{(0)\prime}, \ \dots \ , \theta_{(s)}^{(0)\prime})$ be an arbitrary point in Θ. Generate successive synthetic random subvectors,

$$\theta_{(j)}^{(1)} \mid \{\theta_{(1)}^{(1)}, \dots , \theta_{(j-1)}^{(1)}, \theta_{(j+1)}^{(0)} \dots , \theta_{(s)}^{(0)}\} \sim p_{(j)}[\theta_{(1)}^{(1)}, \dots , \theta_{(j-1)}^{(1)}, \theta_{(j+1)}^{(0)} \dots , \theta_{(s)}^{(0)}] \qquad (j = 1, \dots , s).$$

For subsequent reference, denote the composition of the vector after step j of this conditional sampling process by $\theta^{(1,j)\prime} = (\theta_{(1)}^{(1)\prime}, \dots , \theta_{(j)}^{(1)\prime}, \theta_{(j+1)}^{(0)\prime} \dots , \theta_{(s)}^{(0)\prime})$, and denote its composition after the last step by $\theta^{(1)} = \theta^{(1,s)}$. We shall refer to each of the conditional samplings as a *step*. We shall refer to the completion of the first s steps, resulting in the vector $\theta^{(1)}$, as the first *pass* through the vector θ.

The second and successive passes are performed similarly. At the i'th step of the j'th pass,

$$\theta_{(i)}^{(j)} \mid \{\theta_{(1)}^{(j)}, \dots , \theta_{(i-1)}^{(j)}, \theta_{(i+1)}^{(j-1)} \dots , \theta_{(s)}^{(j-1)}\} \sim p_{(j)}[\theta_{(1)}^{(j)}, \dots , \theta_{(i-1)}^{(j)}, \theta_{(i+1)}^{(j-1)} \dots , \theta_{(s)}^{(j-1)}],$$

and the composition of the vector is

$$\theta^{(j,i)\prime} = [\theta_{(1)}^{(j)\prime}, \ldots, \theta_{(i)}^{(j)\prime}, \theta_{(i+1)}^{(j-1)\prime} \ldots, \theta_{(s)}^{(j-1)\prime}];$$

at the end of the j'th pass the composition of the vector is

$$\theta^{(j)\prime} = [\theta_{(1)}^{(j)\prime}, \ldots, \theta_{(s)}^{(j)\prime}].$$

Under weak conditions outlined in Gelfand and Smith (1990), which amount to the assumption that Θ is connected, $\theta^{(j)}$ converges in distribution to the limiting density $p(\theta)$. Moreover, the rate of convergence is geometric in the L_1 norm. To obtain these convergence results it is necessary only to assume that each subvector $\theta^{(j)}$ is visited infinitely often. Thus, many variants on the cyclical scheme outlined here are possible. For most applications the simplicity of the cyclical scheme seems to be compelling, but we shall return to this question in the last section of the paper.

Since $\theta^{(j)}$ converges in distribution to the posterior distribution $p(\cdot)$, the limiting distribution of $g(\theta^{(j)})$ is the same as the distribution of $g(\theta)$ under the posterior. Given independent realizations of $\theta^{(j)}$, the strong law of large numbers would at once motivate an approximation to $E[g(\theta)]$ using sample averages. Of course, successive drawings are not independent and in the applications reported in the literature, the Gibbs sampling process is typically restarted many times, quasi-independently, in order to achieve a sufficiently good approximation to independence. However, creation of an approximation to independence is neither necessary nor desirable, and we shall return to this point in the next section.

The Gibbs sampler is an attractive solution of the Bayesian multiple integration problem when the conditional densities are simple and easy to obtain. In the special simple case $s = 1$, one is sampling directly from the posterior density and convergence trivially obtains in the first pass. This case is not inherently interesting, but it suggests that Gibbs sampling schemes with small s may have convergence and computational efficiency properties that are attractive relative to those with large s. At the other extreme, one can take s to be equal to the dimension of the parameter vector, and use one of several generic procedures for generating univariate random variables from an arbitrary distribution. Such schemes are likely to be impractical, since the integrand changes at each step of each pass.

The Gibbs sampler is a competitive solution of the Bayesian multiple integration problem when the form of the posterior density renders other methods awkward or inefficient. For example, the derivation of series expansions or importance sampling densities may be cumbersome, while at the same time the conditional densities $p^{(j)}(\cdot)$ are

trivial. In the two examples taken up in this paper, series expansions and importance sampling densities can be constructed, but the Gibbs sampler is much simpler and computations with it are very fast.

2.3 Data Augmentation

In many instances the posterior density $p(\theta)$ does not immediately decompose into subvectors with convenient conditional densities. However, there always exists the formal possibility that one can reexpress the posterior density

$$p(\theta) = \int_{Y^*} \tilde{r}(\theta|y^*)q(y^*)dy^* = \int_{Y^*} \tilde{q}(y^*|\theta)r(\theta)dy^*,$$

and the conditional densities $\tilde{r}(\theta|y^*)$ and $\tilde{q}(y^*|\theta)$ may be well suited to the Gibbs sampling scheme. (Of course, this may involve more than two-step passes: i.e., it may be necessary to further decompose θ, $\tilde{a}$, or both.) The introduction of $\tilde{a}$, proposed by Tanner and Wong (1987), is known as data augmentation. The key to its utility is that the construction of $\tilde{a}$ is frequently natural rather than artificial. Indeed, in many signal extraction problems and latent variable models, difficulties with the posterior density arise precisely because of the need to perform this integration. In these cases, it is often easy to draw θ and $\tilde{a}$ successively; it is not even necessary to write the posterior density explicitly. We shall return to such an example, previously studied using the Gibbs sampler by Chib (1990), in Section 5. In what follows, when we refer to the Gibbs sampler we shall implicitly include the possibility of data augmentation.

3. Assessing Numerical Accuracy and Convergence

The Gibbs sampler, with or without data augmentation, suffers from the complications that the sequences produced are neither independent nor identically distributed. To date, the literature has dealt with these problems in ways that are informal and computationally inefficient. Here, we suggest a careful and systematic treatment of the problem. This treatment has three attractions. (1) It is computationally efficient, using virtually all the sample evidence from the Gibbs sampling scheme. The example taken up in Section 5 yields drastic improvements in computational efficiency over that reported elsewhere, and there is sound reason to believe that is the case generally. (2) Using standard techniques in spectral analysis, the suggested treatment provides a standard error for the approximation of $E[g(\theta)]$ by corresponding sample averages of $g(\cdot)$ taken over the passes and steps of

the Gibbs sampler. (3) Based on these distributional results, a diagnostic for nonconvergence of the Gibbs sampling scheme is constructed.

3.1 Serial Correlation and the Efficient Use of Information

We take up the dependence and convergence problems in succession. To begin, ignore the convergence problem and assume that the sequence $\theta^{(j)}$ is identically but not independently distributed. In general, a fully efficient use of the realizations of the Gibbs sampling process might entail the computation of the corresponding value of $g(\cdot)$ at each pass and step. To maintain this level of generality, consider the $s \times 1$ stochastic process

$$G(j) = (g(\theta_{(1)}^{(j)}), g(\theta_{(2)}^{(j)}), \ldots, g(\theta_{(s)}^{(j)}))' \qquad (j = 1, 2, \ldots p).$$

The problem is to estimate the mean of $G(j)$, subject to the constraint that each mean of the $s \times 1$ vector is the same. Assume that the Gibbs sampling process, and the importance function $g(\cdot)$, jointly imply the existence of a spectrum for $\{G(j)\}$, and the existence of a spectral density $S_G(\omega)$ with no discontinuities at the frequency $\omega = 0$. The asymptotically efficient (in p) estimator of this mean is simply the grand sample average of all of the $g(\theta_{(i)}^{(j)})$; in our notation, it is $(ps)^{-1}\sum_{j=1}^{p} \iota' G(j)$, where ι denotes an $s \times 1$ vector of 1's, and we shall refer to this estimator as $\bar{g}_p$. The asymptotic variance of this esimator is $(ps)^{-1}\iota' S_G(0)\iota$ (Hannan, 1970, pp. 207-210). We may obtain a standard error of numerical approximation for the estimator by estimating $S_G(0)$ in conventional fashion. Moreover, estimation of the full spectral density may yield insights into the nature of the stochastic process implicit in the Gibbs sampling scheme, as will be suggested in some of the specific examples taken up subsequently.

While this method extracts the most information about $E[g(\theta)]$ given the realizations $\theta_{(i)}^{(j)}$, the pertinent practical decision is how often to compute the function(s) of interest $g(\cdot)$ relative to the steps and passes of the Gibbs sampling process. The best decision would reflect the relative costs of drawing $\theta_{(i)}^{(j)}$ and computing $g(\cdot)$, and the degree of serial correlation in the process $\{G(j)\}$. It is clear how this problem could be set up and the solution incorporated in sophisticated software, but we do not enter into these issues here. Instead, we conjecture that typically it will be satisfactory to compute $g(\theta_{(i)}^{(j)})$ at the end of each pass. Since there are no computations within passes, $\{G(j)\}$ becomes a univariate stochastic process and the asymptotically efficient estimator of $E[g(\theta)]$ is $\bar{g}_p =$

$p^{-1}\sum_{j=1}^{p} G(j)$, whose asymptotic variance is $p^{-1}S_G(0)$. The corresponding *numerical standard error (NSE)* of the estimate is $[p^{-1}\hat{S}_G(0)]^{1/2}$. In all results reported in this paper, $\hat{S}_G(0)$ is formed from the periodogram of $\{G(j)\}$ using a Daniell window of width $2\pi/M$, $M = (.3p^{1/2})$.

This method for assessing numerical accuracy can be applied to many variants on the basic sampling scheme for the $\theta_{(i)}^{(j)}$. Many of the applications in which the sampling process is restarted many times can be analyzed in exactly the same way. For example, if every m'th pass is used in an effort to induce quasi-independence in the computed G(j), the estimated spectral density may still be used to provide a measure of numerical accuracy. In fact this computation, or an equivalent computation, would appear necessary in verifying a claim that $\{\theta^{(j)}\}$ is a stochastic process which is essentially serially uncorrelated. Given the methods proposed here, of course, there is no analytical constraint requiring the construction of such a sequence in the first place.

3.2 Assessing Convergence

This formulation of the dependence problem also provides a practical perspective on the convergence problem. Given the sequence $\{G(j)\}$, comparison of values early in the sequence with those late in the sequence is likely to reveal failure of convergence. Let

$$\bar{g}_p^A = p_A^{-1}\sum_{j=1}^{p_A} G(j), \qquad \bar{g}_p^B = p_B^{-1}\sum_{j=p^*}^{p} G(j) \quad (p^* = p - p_B + 1),$$

and let $\hat{S}_G^A(0)$ and $\hat{S}_G^B(0)$ denote consistent spectral density estimates for $\{G(j), j = 1, \ldots, p_A\}$ and $\{G(j), j = p^*, \ldots, p)$, respectively. If the ratios p_A/p and p_B/p are fixed, with $(p_A + p_B)/p < 1$, then as $p \to \infty$,

$$(\bar{g}_p^A - \bar{g}_p^B) / [p_A^{-1}\hat{S}_G^A(0) + p_B^{-1}\hat{S}_G^B(0)] \Rightarrow N(0, 1) \qquad (3.1)$$

if the sequence $\{G(j)\}$ is stationary. We shall refer to the left side of this expression as the *convergence diagnostic (CD)*. This application of a standard central limit theorem exploits not only the increasing number of elements of each sample average, but also the limiting independence of $\bar{g}_p^A$ and $\bar{g}_p^B$ owing to $(p_A + p_B)/p < 1$. These two conditions need to be kept in mind when using this diagnostic, as must considerations of power. In work to date we have taken $p_A = .1p$ and $p_B = .5p$. These choices meet the assumptions

underlying (3.1), while attempting to provide diagnostic power against the possibility that the {G(j)} process was not fully converged early on.

3.3 Preliminary Passes

In practice, one is free to choose the start of the sampling process. From the initial and possibly arbitrary point $\theta^{(0)}$, initial iterations may proceed before the sampling that enters into the computation of $\bar{g}_p$ begins. Indeed, a subsequent example will suggest that this process is critical to $\bar{g}_p$ whose numerical accuracy is reliably known. Computation of $g(\theta)$ is not required at this stage. In this paper the number of such presampling passes is treated as a subjectively chosen parameter of the experimental design. With some foundation of experience in this sort of exercise it should be possible to design algorithms for terminating the presample passes and initiating the computation of $\bar{g}_p$, based on (3.1) or similar computations.

3.4 Relative Numerical Efficiency

Variants of the Gibbs sampling procedure can be compared with each other and with other solutions of the Bayesian multiple integration problem by means of a convenient benchmark. Had the problem been solved by making p independent, identically distributed Monte Carlo drawings $\{\theta_1, \ldots, \theta_p\}$ directly from the posterior density, and $E[g(\theta)]$ estimated as the sample average of the $g(\theta_i)$ over these drawings, then the variance of this estimate would be $var[g(\theta)]/p$, where $var[g(\theta)]$ is the posterior variance of $g(\theta)$. By contrast the variance of the Gibbs sampler is $S_G(0)/p$. Following Geweke (1989), we shall refer to the ratio of the former to the latter, $var[g(\theta)]/S_G(0)$, as the *relative numerical efficiency (RNE)* of the Gibbs sampling estimator of $E[g(\theta)]$. This quantity is of great practical interest for two reasons. First, it may be approximated by means of routine side computations in the Gibbs sampling process itself. While we cannot in fact construct i.i.d. drawings from the posterior density, the Gibbs sampling estimate $\hat{var}[g(\theta)]$ of $var[g(\theta)]$ can be formed in the same way that the Gibbs sampling estimate of $E[g(\theta)]$ is formed. The ratio $\hat{var}[g(\theta)]/\hat{S}_G(\theta)$ then approximates relative numerical efficiency. The other reason for considering RNE is its immediate relation to computational efficiency. The number of drawings required to achieve a given degree of numerical accuracy is inversely related to the relative numerical efficiency of the Gibbs sampling process for the function of interest: were RNE doubled then the number of drawings required would be halved, and so on.

It is worth noting that the RNE of the Gibbs sampling process is solely a function of the serial correlation characteristics of the process $\{G(j)\}$:

$$\text{RNE} = \text{var}[g(\theta)]/S_G(0) = (2\pi)^{-1} \int_{-\pi}^{\pi} S_G(\omega)d\omega \,/\, S_G(0).$$

This formulation makes it clear that the relative numerical efficiency of the Gibbs sampling process depends on the power of the spectral density of $\{G(j)\}$ at $\omega = 0$, relative to the distribution of its spectral density across other frequencies. Thus, relative numerical efficiency may be quite different for different functions of interest. Furthermore, RNE is not bounded above by one: in principle, efficiency many times that achieved by i.i.d. sampling directly from the posterior density can be achieved by the Gibbs sampling estimator of $E[g(\theta)]$. Heuristically, positive serial correlation of $\{G(j)\}$ renders the Gibbs sampling estimator less efficient, and negative serial correlation in $\{G(j)\}$ renders it more efficient.

3.5 A Constructed Example

It may be helpful to illustrate these ideas in a constructed example simple enough that an analytical approach is possible. Consider the case of a bivariate normal posterior density for $\theta = (\theta_1, \theta_2)'$, with zero mean and $\text{var}(\theta_i) = \sigma_{ii}$, $\text{cov}(\theta_1, \theta_2) = \sigma_{12}$. Denote the stochastic process corresponding to the Gibbs sampler by $\{\tilde{\theta}\}$, and suppose that the Gibbs sampling design is

$$\tilde{\theta}_1^j = (\sigma_{12}/\sigma_{22})\tilde{\theta}_2^{j-1} + \varepsilon_{1j} \qquad \text{var}(\varepsilon_{1j}) = \sigma_{11} - \sigma_{12}^2 / \sigma_{22}$$

$$\tilde{\theta}_2^j = (\sigma_{12}/\sigma_{11})\tilde{\theta}_1^j + \varepsilon_{2j} \qquad \text{var}(\varepsilon_{2j}) = \sigma_{22} - \sigma_{12}^2 / \sigma_{11}$$

The spectral density of this bivariate process, at frequency $\omega = 0$, is

$$S_{\tilde{\theta}}(0) = (1-r^2)^{-1} \begin{bmatrix} \sigma_{11}(1+r^2) & 2\sigma_{12} \\ 2\sigma_{12} & \sigma_{22}(1+r^2) \end{bmatrix}, \qquad r^2 \equiv \sigma_{12}^2 / \sigma_{11}\sigma_{22}.$$

By comparison, the posterior variance of θ may be written

$$\Sigma = (1-r^2)^{-1} \begin{bmatrix} \sigma_{11}(1-r^2) & \sigma_{12}(1-r^2) \\ \sigma_{12}(1-r^2) & \sigma_{22}(1-r^2) \end{bmatrix},$$

and their difference is

$$S_{\tilde{\theta}}(0) - \Sigma = (1-r^2)^{-1} \begin{bmatrix} 2\sigma_{11}r^2 & \sigma_{12}(1+r^2) \\ \sigma_{12}(1+r^2) & \sigma_{22}r^2 \end{bmatrix}, \qquad (3.3)$$

which is not positive semidefinite. Consequently, the relative numerical efficiency of some functions of interest of the parameters θ_1 and θ_2 would exceed one, while others would be less than one, if the functions of interest are evaluated at the end of each pass. If the functions of interest are evaluated after each step, then the pertinent difference is

$$2S_{\tilde{\theta}}(0) - \Sigma = (1-r^2)^{-1} \begin{bmatrix} \sigma_{11}r^2(1+3r^2) & \sigma_{12}(3+r^2) \\ \sigma_{12}(3+r^2) & \sigma_{22}(1+3r^2) \end{bmatrix}.$$

Since this matrix is positive definite, the relative numerical efficiency for any linear function of interest of the parameters must then be less than one.

To examine these results in more detail, let $\theta_{11} = \theta_{22} = 1$, $\theta_{12} = .5^{1/2}$. Then the eigenvalues of $S_{\tilde{\theta}}(0) - \Sigma$ are 4.1213 and -.1213, with respective corresponding eigenvectors $(.5^{1/2}, .5^{1/2})$ and $(.5^{1/2}, -.5^{1/2})$. The Gibbs sampling process induces positive correlation between $\tilde{\theta}_1^j$ and $\tilde{\theta}_2^j$. This raises the variance of $\tilde{\theta}_1^j + \tilde{\theta}_2^j$ and lowers the variance of $\tilde{\theta}_1^j - \tilde{\theta}_2^j$, relative to the case of independence. In the case of $\tilde{\theta}_1^j - \tilde{\theta}_2^j$, the reduction more than offsets the positive serial correlation in $\tilde{\theta}^j$.

Table 1 exhibits the numerical standard errors (NSE's) and relative numerical efficiencies (RNE's) of four alternative functions of interest. The population values are derived from $S_{\tilde{\theta}}(0)$ and Σ. The results presented for $p = 400$ and $p = 10{,}000$ are based on a starting value chosen from the posterior density (possible in this constructed example, but not possible generally) and no preliminary passes. The results for $p = 10{,}000$ agree quite well with the population values. Those for $p = 400$ agree well, except that the NSE for the function of interest $.5(\theta_1 - \theta_2)$ is somewhat too high and the corresponding RNE is therefore somewhat too low. This discrepancy can be traced to the smoothing of the

periodogram in the formation of the estimate $\hat{S}_G(\omega)$: this function has a minimum at $\omega = 0$, and when $p = 400$ the estimate at $\omega = 0$ is an average of periodogram ordinates extending from $-\pi/3$ to $\pi/3$, which raises its value. When $p = 10{,}000$ the average extends from $-\pi/15$ to $\pi/15$ and this effect is negligible.

Table 2 provides some estimated spectral density ordinates from the case $p = 10{,}000$. The computations at frequencies other than $\omega = 0$ are not essential to the procedure, of course. However, they are easy to produce and are given here to further illustrate the differences in NSE's and RNE's for the different functions of interest. Note, in particular, that for the first three functions of interest power is greatest at $\omega = 0$. Whenever this is true, RNE must be less than one. For the fourth function of interest power is smallest at $\omega = 0$, implying that RNE must exceed one.

4. Inference in the Linear Model with an Informative Prior

We turn now to a simple but important application of Gibbs sampling, the multiple normal linear regression model, with a proper normal prior on a subset of the coefficients. The example is simple because there is only a single parameter that prevents analytical solution of the whole problem, in the case of linear functions of interest of the coefficients, or integration by Monte Carlo sampling directly from the entire posterior density, in the case of nonlinear functions of interest of the coefficients. The example is important in itself, because the model is widely applied in many disciplines and informative normal priors are frequently a reasonable representation of prior knowledge. It is also important because the model and prior occur repeatedly as a key conditional distribution when more difficult problems are attacked using Gibbs sampling. (An example is provided in the next section.) The solution of this problem may then be applied in those cases. The strategy of constructing such "building blocks" seems well suited to the research program of constructing Gibbs samplers for many standard econometric models.

4.1 The Model and the Prior

To establish notation, write the multiple normal linear regression model,

$$y_i = x_i'\beta + \varepsilon_i \qquad \varepsilon_i \sim \text{IIDN}(0, \sigma^2) \qquad (i = 1, \ldots, n),$$

where x_i is the i'th observation on a $k \times 1$ vector of explanatory variables. Alternatively the model can be expressed

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 I_n),$$

with each row of the $n \times 1$ vector y and $n \times k$ matrix X corresponding to a single observation. The likelihood function is

$$L(\beta, \sigma) \propto \sigma^{-n} \exp[-\sum_{i=1}^{n}(y_i - x_i'\beta)'(y_i - x_i'\beta)/2\sigma^2] = \sigma^{-n} \exp[(y - X\beta)'(y - X\beta)/2\sigma^2].$$

Since the likelihood function is the essential representation of the model, the developments reported in this section apply to any model that generates this likelihood function, including one with stochastic explanatory variables. The vector of unknown parameters is $\theta' = (\beta', \sigma)$.

The prior density is of the form $\pi(\beta, \sigma) = \pi_1(\beta)\pi_2(\sigma)$. The prior density for β is expressed in terms of m $(\leq k)$ linear combinations of β,

$$R\beta \sim N(r, T) \quad \Leftrightarrow \quad \pi_1(\beta) \propto \exp\{-\tfrac{1}{2}(R\beta - r)'T^{-1}(R\beta - r)\}$$

or

$$Q\beta \sim N(q, I_m) \quad \Leftrightarrow \quad \pi_1(\beta) \propto \exp\{-\tfrac{1}{2}(Q\beta - q)'(Q\beta - q)\},$$

where Q is a factorization of T^{-1}, $Q'Q = T^{-1}$, and $q = Qr$. When $m < k$ this prior is improper, but of course may be constructed as a limit of proper priors. The conjugate uninformative prior $\pi_2(\sigma) \propto \sigma^{-1}$ is assumed for σ, although what follows could easily be replicated for any one of the family of proper, inverted gamma priors for σ, of which $\pi_2(\sigma)$ is a limit.

The posterior density may be written

$$p(\beta, \sigma) \propto \sigma^{-(n+1)}\exp\{[\beta - \hat{\beta}(\sigma)]'[V(\sigma)]^{-1}[\beta - \hat{\beta}(\sigma)]\}, \tag{4.1}$$

with

$$\hat{\beta}(\sigma) = (X'X + \sigma^2Q'Q)^{-1}(X'y + \sigma^2Q'q) \tag{4.2}$$

and

$$V(\sigma) = \sigma^2(X'X + \sigma^2Q'Q)^{-1}. \tag{4.3}$$

4.2 Previous Approaches

Analytical integration of the posterior density is not possible, even for linear functions of interest of the coefficients β. Most practical approaches, including the one taken here, rely on the observation that conditional on σ, the posterior density for β is multivariate normal. Theil and Goldberger (1961) suggested that σ^2 be fixed at $s^2 = (y - Xb)'(y - Xb)/(n - k)$, a procedure they termed "mixed estimation" because the mean and variance of the posterior density can then be computed immediately using standard least

squares regression software, appending the n entries in y with the m elements of sq, and the n rows of X with the m rows of sQ. Denote the point estimator

$$\hat{\beta}_{TG} = (X'X + s^2Q'Q)^{-1}(X'y + s^2Q'q)$$

and the corresponding variance

$$\hat{var}(\hat{\beta}_{TG}) = s^2(X'X + s^2Q'Q)^{-1}.$$

In part because of this convenient description, the Theil-Goldberger mixed estimator has proved popular and has been used in many applications.

Tiao and Zellner(1964) took up the problem of Bayesian inference for two normal linear regression models with the same coefficients but unequal variances in their disturbances. If one of the variances is known and the other is not, then the essentials of this problem are the same as the one posed here. They show that an asymptotic normal expansion of the posterior density yields the mean $\hat{\beta}_{TG}$ and the variance $\hat{var}(\hat{\beta}_{TG})$.

4.3 The Gibbs Sampler

We construct a two-step Gibbs sampler, based on the distribution of σ conditional on β, and the distribution of β conditional on σ. The posterior density of σ conditional on β is

$$p(\sigma|\beta) \propto \sigma^{-(n+1)}\exp[-(y-X\beta)'(y-X\beta)/2\sigma^2).$$

If we define $SSR(\beta) = (y-X\beta)'(y-X\beta)$, then

$$(SSR(\beta)/\sigma^2)|\beta \sim \chi^2(n).$$

The posterior density of β conditional on σ is normal as indicated by (4.1) - (4.3). However, it is computationally inefficient to invert the k x k matrix $X'X + \sigma^2Q'Q$ in each pass of the Gibbs sampling algorithm. Instead, let L be a factor of $(X'X)^{-1}$, $LL' = (X'X)^{-1}$. Let $L'Q'QL$ have diagonalization $P\Lambda P'$: i.e., Λ is a diagonal matrix of eigenvalues of $L'Q'QL$, and the columns of P are the corresponding, ordered eigenvectors normalized so that $P'P = PP' = I_k$. Finally, let $H = LP$. (These computations are only performed once, prior to the Gibbs sampling passes.) Then the variance of β conditional on σ may be expressed

$$(\sigma^{-2}X'X + Q'Q)^{-1} = H(\sigma^{-2}I_k + \Lambda)^{-1}H'.$$

This leads to a simple construction for β given σ. Construct $\varepsilon \sim N(0, I_k)$; scale ε_i by $(\sigma^{-2} + \lambda_i)^{-1/2}$ to form the vector ζ; form $\eta = H\varepsilon$; and then add the mean vector

$$(\sigma^{-2}X'X + Q'Q)^{-1}(\sigma^{-2}X'y + Q'q) = H(\sigma^{-2}I_k + \Lambda)^{-1}H'(X'y + \sigma^2Q'q),$$

using the right hand side of this expression to perform the computations. The number of multiplications required is proportional to k^3, the same as for matrix inversion, but the computations are nearly three times as fast for this method, suggesting that the factor of proportionality must be about one-third that for direct inversion.

4.4 A Numerical Example

This Gibbs sampling algorithm for Bayesian inference in the normal linear model with informative normal linear priors on the coefficients was coded in Fortran-77 using the IMSL Math/Library and IMSL Stat/Library. The results reported here were executed on a Sun Sparcstation 4/40 (IPC), in 64-bit arithmetic. We report results using 400 passes and 10,000 passes. For 400 passes, Gibbs sampling time was 0.32 seconds and the time required to form the periodograms and compute spectral density estimates at 21 ordinates was 4.2 seconds. For 10,000 passes, Gibbs sampling time averaged 7.78 seconds and the spectral computations averaged 17.28 seconds.

The results reported here are based on artificial data from a model with $k = 3$ regressors. One regressor is an intercept term and the other two are orthogonal standard normal variates. The disturbance term is also standard normal. The coefficients are all 1.0. (Since the vector of coefficients β is generated jointly, conditional on σ, in our Gibbs sampling algorithm, results will not depend on the structure of the design matrix $X'X$; the orthogonal structure taken here is simply a convenient one.) The sample size is $n = 100$. Hence, with an uninformative prior the posterior distribution for β would be a multivariate Student-t distribution with 97 degrees of freedom, mean approximately (1, 1, 1), and variance approximately proportional to $(.01)I_3$. The informative prior employed is $\beta \sim N(r, (.01)I_3)$. Thus, the prior and the data are equally informative in the sense that the precision matrices associated with each are about the same.

The initial values for the algorithm are taken from the least squares estimates, one presample value was generated and discarded, and then the successive passes with computation of functions of interest at the end of each pass were initiated. Hence, $\beta^{(-1)} = b = (X'X)^{-1}X'y$ and $\sigma^{(-1)} = (s^2)^{1/2}$, where $s^2 = (y-Xb)'(y-Xb)/(n-k)$; $\beta^{(0)}$ and $\sigma^{(0)}$ are discarded. This was done for the five alternative settings of the prior mean, r, reported in Table 3. When the prior mean is many standard deviations away from the sample mean,

the initial values are violently unrepresentative of the posterior. Nevertheless, the diagnostics indicate no problems with convergence. Examination of the actual early values generated shows that the $\sigma^{(j)}$ sequence moves immediately from the neighborhood of s^2, which is far too low when r is far from (1, 1, 1), to values consistent with the mass of the posterior for σ. The contrast with the mixed estimates, in this regard, is striking. The variance matrix (4.3) associated with this estimate implicitly takes s^2 as representative of σ^2, and does not reflect the larger plausible values which are implied when the sample and prior means are far apart. In the context of the asymptotic expansion of Tiao and Zellner (1964), that approximation is good to the extent that the data dominate the prior, a circumstance markedly uncharcteristic of situations in which the sample and prior means are far apart in the metric of sample precision.

The results in Table 3 strongly suggest that the Gibbs sampling algorithm provides an adequate solution to the problem of Bayesian inference in the normal linear model with an informative normal prior on the coefficients. Three aspects of the results support this conclusion. First, convergence beginning with the least squares estimates is essentially instantaneous, even when these esimates provide parameter values at which the posterior density is quite low. Second, the relative numerical efficiency (RNE) for all parameters exceeds .5 and is often near 1. This might have been anticipated as an asymptotic result, since β and σ become independent as sample size increases without bound. In the examples studied here they are not, especially for values of r far from the sample mean. That the algorithm works so well in these circumstances is encouraging. Consistent with (but not implied by) the RNE values between .5 and 1.2, spectral densities of all parameters appear nearly flat, and are not reported here. Third, computation times are reasonable. The structure of the problem and the experience with this example indicate computation time of about $(8.5 \times 10^{-7})npk^3$ seconds. This implies reasonable desktop computing times for most of the econometric applications of this model.

5. Inference in the Tobit Censored Regression Model

Limited dependent variable models constitute one of the principal tools of applied econometrics. In some of these models the dependent variable is dichotomous, reflecting a decision to purchase or not purchase a durable good, whether or not to retire, etc. The probit and logit models are often used in these situations. In other cases decisions are of the form, "whether or not, and if so then how much?". This characterizes the form of many investment decisions, like the construction of a plant, and consumption decisions, like the purchase of an automobile. The Tobit model, introduced by Tobin (1958) is probably the most widely applied model in these situations. The monograph of Maddala (1983), and the three chapters of Grilliches and Intriligator (1984) provide a thorough discussion of these and related limited dependent variable models.

5.1 The Model and the Prior

To establish notation, write the Tobit censored regression model,

$$y_i^* = x_i'\beta + \varepsilon_i, \quad \varepsilon_i \sim \text{IIDN}(0, \sigma^2) \qquad (i = 1, .., n), \tag{5.1}$$

$$y_i = \begin{cases} y_i^*, & \text{if } y_i^* \geq 0 \\ 0, & \text{if } y_i^* < 0. \end{cases} \tag{5.2}$$

We observe $\{x_i, y,\}_{i=1}^n$; the x_i's are k x 1 vectors; the y_i's are scalars; and the y_i^*'s are unobserved. For notational convenience, order the observations so that the c censored observations (those for which $y_i = 0$) come first, followed by the n - c uncensored observations. Let $y_2 = (y_{c+1}, \ldots, y_n)'$, and let $X_2' = [x_{c+1}, \ldots, x_n]$. Then the kernel of the likelihood function is

$$\prod_{i=1}^{c}[1 - \Phi(x_i'\beta/\sigma)]\sigma^{-(n-c)}\exp\{-(y_2-X_2\beta)'(y_2-X_2\beta)/2\sigma^2\}.$$

The prior density is exactly the same as that employed in the previous section: $\pi(\beta, \sigma) = \pi_1(\beta)\pi_2(\sigma)$, with the prior density for β expressed as m ($\leq$ k) linear combinations of β,

$$R\beta \sim N(r, T) \quad \Leftrightarrow \quad \pi_1(\beta) \propto \exp\{-\tfrac{1}{2}(R\beta - r)'T^{-1}(R\beta - r)\}$$

or

$$Q\beta \sim N(q, I_m) \quad \Leftrightarrow \quad \pi_1(\beta) \propto \exp\{-\tfrac{1}{2}(Q\beta - q)'(Q\beta - q)\},$$

where Q is a factorization of T^{-1}, $Q'Q = T^{-1}$, and $q = Qr$; $\pi_2(\sigma) \propto \sigma^{-1}$. The posterior density may be written

$$p(\beta, \sigma) \propto \prod_{i=1}^{c}[1 - \Phi(x_i'\beta/\sigma)]\sigma^{-(n-c)}\exp\{[\beta - \hat{\beta}(\sigma)]'[V(\sigma)]^{-1}[\beta - \hat{\beta}(\sigma)]\},$$

with

$$\hat{\beta}(\sigma) = (X_2'X_2 + \sigma^2Q'Q)^{-1}(X_2'y_2 + \sigma^2Q'q)$$

and

$$V(\sigma) = \sigma^2(X_2'X_2 + \sigma^2Q'Q)^{-1}.$$

5.2 Previous Approaches

Maximum likelihood is well established as the principal method of frequentist inference in the Tobit censored regression model. Maddala (1983) and Amemiya (1984) provide thorough surveys. Bayesian inference in a closely related model that arises in biomedical applications has been discussed by Carriquiry et al. (1987) and Sweeting (1987). The most thorough Bayesian treatment of this model is provided by Chib (1990), who has implemented and compared Monte Carlo integration with importance sampling, Laplace approximations, and Gibbs sampling with data augmentation. The implementation of Gibbs sampling and data augmentation reported here extends Chib's treatment in three dimensions. First, the methodology of Section 3 is used to adduce evidence on convergence and the accuracy of the numerical approximations. Second, an informative prior is permitted, whereas Chib (1990) uses only uninformative priors. Third, the research reported here entails a controlled study of the effect of alternative starting values and convergence. The outcome suggests caution in interpreting informal diagnostics for convergence.

5.3 The Gibbs Sampler with Data Augmentation

Construction of a three-step Gibbs sampler with data augmentation is straightforward. Conditional on β and σ, y_i^* has a truncated normal distribution, constructed from $N(x_i'\beta, \sigma^2)$ truncated above at 0; i.e., the p.d.f. of y_i^* is

$$f(y_i^* \mid \beta, \sigma) = \begin{cases} [1 - \Phi(x_i'\beta/\sigma)]^{-1}\exp[-(y_i^* - x_i'\beta)^2/2\sigma^2)], & y_i^* \le 0; \\ 0, & y_i^* > 0; \end{cases}$$

for $i = 1, \ldots, c$; $y_i^* = y_i$ for $i = c + 1, \ldots, n$. An algorithm for generating from the truncated univariate normal distribution, described in Geweke (1991) and considerably faster than either naive rejection methods or the conventional construction of Devroye (1986), is employed. This data augmentation constitutes the first step of the Gibbs sampler.

Conditional on $\{y_i^*\}_{i=1}^n$, the problem reduces to precisely the one set forth and solved in Section 4. This constitutes the second and third steps of the Gibbs sampler with data augmentation.

5.4 A Numerical Example

An artificial sample of size 200 was constructed, using a data generating process similar to Wales and Woodland (1980):

$$x_{i3},\ z_{i2},\ z_{i3} \text{ IID Uniform } (-1, 1);$$

$$x_{i2} = z_{i2} + z_{i3} + u_i, \text{ with } u_i \sim N(0, 1.312);$$

$$\varepsilon_i \sim N(0, .6428);$$

$$y_i^* = -1.1854 + 1.0x_{i2} + 10.0x_{i3} + \varepsilon_i;$$

$$y_i = \begin{cases} y_i^*, & \text{if } y_i^* \geq 0 \\ 0, & \text{if } y_i^* < 0. \end{cases}$$

The five random variables x_{i3}, z_{i2}, z_{i3}, u_i, ε_i are mutually and serially independent. Of the generated sample, 114 observations are censored.

A small, full-factorial experiment was conducted. The total number of passes was taken to be either 400 or 10,000; no preliminary passes were taken, or the number of preliminary passes was set equal to the total number of passes; and four alternative initial values for $\theta = (\beta_1, \beta_2, \beta_3, \sigma^2)'$ were used. The four alternative starting values were:

(i) Uncensored ordinary least squares. $\beta^{(0)}$ is the least squares vector and $\sigma^{2(0)}$ is the corresponding value of s^2, from application of least squares to the full set of 200 observations with $y_i = 0$ for censored observations;

(ii) Censored ordinary least squares. $\beta^{(0)}$ is the least squares vector and $\sigma^{2(0)}$ is the corresponding value of s^2, from application of least squares to the 86 uncensored observations.

(iii) The augmented posterior mode. The augmented posterior density is a function of 118 variables. A computationally efficient method of finding this mode is to apply the Gibbs sampling algorithm with data augmentation, except that each vector is set to its conditional modal value rather than generating from the conditional posterior density.

(iv) The augmented posterior mode with censored $\sigma^{2(0)}$. This is a hybrid of (ii) and (iii), $\beta^{(0)}$ from (ii) and $\sigma^{2(0)}$ from (iii).

Estimated posterior means and convergence diagnostics for all 16 cells are provided in Table 4, along with numerical values for the initial vectors. Table 5 provides greater detail, for some selected cells in which convergence diagnostics were satisfactory. Table 6 provides the estimated spectral densities of the sampled parameters from two of the cells. As a final check on the results reported here, the last cell in Table 5 was reexecuted, but using 1,000,000 preliminary passes rather than 10,000: the results were within the range anticipated from the NSE's for that cell. The computation times reported in Table 5 were realized on a Sun Sparcstation 4/40 (IPC), with software written in double precision Fortran-77 using the IMSL Math/Library and IMSL Stat/Library. These times correspond roughly to a 20-fold increase in speed over the similar computations of Chib (1990), who used a 16 Mhs 3 MB personal computer with the Gauss programming language.

The results of these experiments can be organized in several dimensions.

(1) 400 passes are generally insufficient for convergence. Only one of the eight cells, with augmented posterior mode initial values and 400 preliminary passes, performs satisfactorily. Poor convergence diagnostics correspond to estimated posterior means that are up to one-half posterior standard deviation from the values reported in Table 5. Thus, reasonable but unconverged values could be significantly misleading.

(2) Consistent with this finding, preliminary passes are important in producing reliable results. In many cells, preliminary passes are necessary to produce satisfactory convergence diagnostics.

(3) The augmented posterior mode exhibits strikingly better performance as an initial value than do the other initial values. From one perspective this is surprising. At the augmented posterior mode most of the y_i^* lie on the regression plane $x_i'\beta$ -- i.e., the corresponding ε_i^* are set to zero. Since 114 observations are censored this produces a very low value of $\sigma^{2(0)}$, and increasing the value of $\sigma^{2(0)}$ to a more

reasonable value, which was done in the final cell of the experimental design in this dimension (see (iv), above) makes matters worse rather than better. The corresponding values of $\beta^{(0)}$ are no closer to the posterior mean than the other initial values in the experiment. From another perspective, this outcome is not surprising. At the augmented posterior mode the 118-dimensional density relevant for Gibbs sampling with data augmentation is high (by definition), and given the smoothness inherent in this problem movement in the various dimensions is more likely and hence convergence is more rapid starting from this point.

(4) There are quite substantial differences in the serial correlation properties of the sampled parameters, as indicated in Table 6. The parameters β_1 and β_3 exhibit very strong positive serial correlation, to the extent that a poor picture of the pattern emerges with only 400 observations. Virtually without exception, these are the only parameters that exhibit poor convergence diagnostics. With a small number of observations it is impossible to distinguish between nonstationarity and high power at low frequencies. For our purposes the distinction is uninteresting, since either will lead to unreliable approximations of posterior moments if the number of passses is too small.

(5) As is necessarily the case, these parameters exhibit poor RNE's. The Gibbs sampler with data augmentation requires about 20 times as many passes as direct Monte Carlo sampling from the posterior would require independent samples (were that possible). From this perspective, the effective number of passes in the experiment is either 20 (when p = 400) or 500 (when p = 10,000) for β_1 and β_3. This is consistent with our inability to obtain satisfactory results for these parameters when p = 400.

6. Conclusions

Gibbs sampling with data augmentation is an attractive solution of the Bayesian multiple integration problem whenever the parameters of the (augmented) posterior density can be grouped in such a way that the distribution of any group conditional on the others is of a standard form from which synthetic sampling is straightforward. As our ingenuity in expressing posterior densities in this form increases, many standard models become ammenable to Bayesian treatment using this method. It now appears that most of the standard applied econometric models can be cast in a form appropriate for Gibbs sampling. Even more promising, awareness of the Gibbs sampler and data augmentation are likely to

suggest new models that are ammenable to Bayesian inference, that are intractable using analytical or other numerical methods.

Yet one must be cautious. A great attraction of this approach, not to be minimized, is that it is in general straightforward to apply relative to other numerical methods. This means that less of the investigator's time is spent on arcane numerical issues, mistakes are less likely to be made, and incorporation into interactive software is more practical. But there is no small distance between having a method justified solely by a convergence result, and one which reliably produces approximations of integrals whose accuracy can be reliably assessed. Formal assessment of convergence and numerical accuracy are essential to rendering the Gibbs sampler a tool of replicable scientific studies, because of the pseudo-randomness inherent in the method.

It is risky to speculate on productive avenues of future research based on the limited collective experience with the Gibbs sampler in Bayesian inference, but three seem clear. First, it is practical to write software that varies many aspects of the experimental design implicit in the Gibbs sampler, so as to produce satisfactory outcomes and economize on machine time. For example, preliminary passes can be used to compute convergence diagnostics, measure relative computation times for Gibbs sampling (on the one hand) and computation of functions of interest (on the other), and get at least a rough estimate of the spectral densities of the sampled processes at the zero frequency. This would determine the number of subseqent passess needed to attain desired numerical accuracy, and the appropriate points for computing functions of interest. Second, it would be very helpful to obtain additional insight on the relation of the internal structure of problems to the stochastic properties of the Gibbs sampler. The results presented here are suggestive, but serve mainly to raise questions. Can we use information about serial correlation to structure the pattern of steps within passes to achieve greater computational efficiency? Is the use of the augmented posterior mode, which proved very helpful in the example in Section 5.4, an attractive starting point for most problems? Finally, the Gibbs sampler with data augmentation is inherently an asyncrhonously parallel algorithm, in which nodes compute inner products. This distinguishes it from other approaches, like Monte Carlo integration which is inherently distributed and series expansions which are inherently serial. Thus, there is at least the possibility that this method could prove practical and highly competitive in an environment of parallel architectures.

References

Amemiya, T. 1984: "Tobit Models: A Survey," *Journal of Econometrics* **24**, 3 - 61.

Carriquiry, A.L., D. Gianola, and R.L. Fernando, 1987: "Mixed-Model Analysis of a Censored Normal Distribution with Reference to Animal Breeding," *Biometrics* **43**, 929-939.

Chib, S., 1990: "Bayes Inference in the Tobit Censored Regression Model," University of Missouri Department of Economics working paper. *Journal of Econometrics*, forthcoming.

Devroye, L., 1986: *Non-Uniform Random Variate Generation.* New York: Springer-Verlag.

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

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

Geweke, J., 1988: "Antithetic Acceleration of Monte Carlo Integration in Bayesian Inference," *Journal of Econometrics* **38**, 73-90.

Geweke, J., 1989: "Bayesian Inference in Econometric Models Using Monte Carlo Integration," *Econometrica* **57**, 1317-1339.

Geweke, J., 1991: "Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints," University of Minnesota working paper. *Computing Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface*, forthcoming.

Grilliches, Z., and M. Intriligator, 1984: *Handbook of Econometrics.* Amsterdam: North-Holland.

Hannan, E.J., 1970: *Multiple Time Series.* New York: Wiley.

Kloek, T., and H.K. Van Dijk, 1978: "Bayesian Estimates of Equation System Parameters: An Application of Integration by Monte Carlo," *Econometrica* **46**, 1-20.

Leonard, T., J.S.J. Hsu, and K.-W. Tsu, 1989: "Bayesian Marginal Inference," *Journal of the American Statistical Association* **84**, 1051-1058.

Maddala, G.S., 1983: *Limited-Dependent and Qualitative Variables in Econometrics.* Cambridge: Cambridge University Press.

Sweeting, T.J., 1987: "Approximate Bayesian Analysis of Censored Survival Data," *Biometrika* **74**, 809-816.

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

Theil, H., and A.S. Goldberer, 1961: "On Pure and Mixed Statistical Estimation in Economics," *International Economic Review* **2**, 65-78.

Tiao, G. C., and A. Zellner, 1964: "Bayes' Theorem and the Use of Prior Knowledge in Regression Analysis," *Biometrika* **51**, 219-230.

Tierney, L., and J.B. Kadane, 1986: "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association* **81**, 82-86.

Tobin, J., 1958: "Estimation of Relationships for Limited Dependent Variables," *Econometrica* **26**, 24-36.

Wales, T.J., and A.D. Woodland, 1980: "Sample Selectivity and the Estimation of Labor Supply Functions," *International Economic Review* **21**, 437-468.

Zellner, A., 1971: *An Introduction to Bayesian Inference in Econometrics*. New York: Wiley.

Table 1

Gibbs Sampling in the Constructed Bivariate Normal Example of Section 3.6

Function of Interest, $g(\theta)$:	θ_1	θ_2	$.5(\theta_1+\theta_2)$	$.5(\theta_1-\theta_2)$
Population values:				
$E[g(\theta)]$	.000	.000	.000	.000
$sd[g(\theta)]$	1.000	1.000	.924	.383
NSE	$1.732/\sqrt{p}$	$1.732/\sqrt{p}$	$1.707/\sqrt{p}$	$.293/\sqrt{p}$
RNE	.333	.333	.293	1.707
Gibbs Sampling, 400 passes (.09 seconds):				
$\hat{E}[g(\theta)]$	-.046	-.017	-.032	.016
$\hat{sd}[g(\theta)]$	.961	.998	.901	.384
NSE	.081	.088	.083	.016
RNE	.352	.318	.293	1.389
CD	-.137	.674	.291	-1.513
Gibbs Sampling, 10,000 passes (1.89 seconds):				
$\hat{E}[g(\theta)]$	.014	.005	.009	.004
$\hat{sd}[g(\theta)]$	1.019	1.021	.945	.384
NSE	.0171	.0173	.0169	.0029
RNE	.355	.350	.311	1.810
CD	.147	.375	.267	-.588

Table 2

Estimated Spectral Density Ordinates for the Constructed Bivariate Normal Example of Section 3.6

Computations based on 10,000 passes

Function of Interest, $g(\theta)$:	θ_1	θ_2	$.5(\theta_1+\theta_2)$	$.5(\theta_1-\theta_2)$
Frequency, ω:				
$.0\pi$	2.916	2.969	2.861	.081
$.1\pi$	2.677	2.590	2.533	.100
$.2\pi$	1.890	2.000	1.826	.119
$.3\pi$	1.171	1.257	1.070	.144
$.4\pi$	.765	.742	.603	.150
$.5\pi$	.594	.602	.438	.159
$.6\pi$	.488	.484	.325	.160
$.7\pi$	.405	.414	.254	.155
$.8\pi$	.349	.390	.210	.159
$.9\pi$	.363	.353	.179	.178
π	.326	.342	.163	.171

Table 3

Linear Model with an Informative Prior

(See Section 4.4 for description)

	Mixed Estimate		-------Posterior, p = 400--------					-------Posterior, p = 10,000------				
Parameter	Mean	s.d.	Mean	s.d.	NSE	RNE	CD	Mean	s.d.	NSE	RNE	CD
r = (1, 1, 1)												
β_1	.937	.073	.939	.065	.0032	1.028	-.631	.936	.066	.0007	.904	-.123
β_2	1.001	.073	.999	.069	.0033	1.075	-.152	1.001	.066	.0007	.970	-.555
β_3	1.045	.075	1.055	.064	.0034	.881	-.548	1.046	.068	.0007	.969	.251
σ^2			.823	.118	.0060	.968	.025	.818	.121	.0013	.851	-.457
r = (1.5, 1.5, 1.5)												
β_1	1.153	.073	1.185	.072	.0039	.830	1.017	1.185	.074	.0009	.720	-.041
β_2	1.222	.073	1.245	.070	.0036	.913	.521	1.249	.072	.0008	.736	-.202
β_3	1.288	.075	1.306	.072	.0034	1.111	.892	1.311	.072	.0007	1.069	-.913
σ^2			1.020	.155	.0095	.663	.857	1.019	.164	.0021	.590	-.498
r = (2, 2, 2)												
β_1	1.370	.073	1.638	.094	.0063	.557	-1.625	1.647	.095	.0012	.615	-.927
β_2	1.443	.073	1.685	.085	.0048	.775	-1.570	1.691	.092	.0011	.663	-.873
β_3	1.530	.075	1.746	.091	.0060	.579	-2.142	1.752	.089	.0010	.797	-1.754
σ^2			2.292	.409	3.146	.423	-2.991	2.333	.439	.00620	.502	-.929
r = (6, 6, 6)												
β_1	3.101	.073	5.930	.110	.0059	.859	-.403	5.926	.102	.0011	.868	.208
β_2	3.210	.073	5.928	.098	.0051	.928	-.102	5.931	.100	.0010	1.067	-2.729
β_3	3.468	.075	5.949	.096	.0055	.759	1.216	5.943	.098	.0010	.982	-.459
σ^2			73.50	11.05	528.6	1.093	-1.451	74.33	10.74	10.65	1.017	-2.001
r = (11, 11, 11)												
β_1	5.266	.073	10.962	.105	.0050	1.086	-.242	10.963	.103	.0010	1.114	-.305
β_2	5.420	.073	10.961	.095	.0045	1.119	.108	10.966	.099	.0010	.896	1.211
β_3	5.891	.075	10.975	.093	.0051	.847	-.773	10.972	.097	.0010	.960	-.757
σ^2			296.3	43.66	250.7	.758	-1.163	300.3	44.08	41.64	1.121	-.280

Table 4

Convergence Diagnostics, Tobit Censored Regression Model

Initial value: Uncensored OLS ($\beta_1 = 1.871$, $\beta_2 = .801$, $\beta_3 = 4.082$, $\sigma^2 = 2.357$)

	400 passes				10,000 passes			
	0 preliminary		400 preliminary		0 preliminary		10,000 preliminary	
	Mean	CD	Mean	CD	Mean	CD	Mean	CD
β_1	-1.272	3.098	-1.328	-3.719	-1.322	4.711	-1.318	3.663
β_2	.929	-1.242	.936	1.037	.936	-.682	.936	-1.517
β_3	9.941	-2.877	10.038	3.641	10.028	-4.487	10.025	-3.817
σ^2	.785	1.415	.714	1.358	.718	.901	.717	.617

Initial value: Censored OLS ($\beta_1 = -.898$, $\beta_2 = .928$, $\beta_3 = 9.423$, $\sigma^2 = .677$)

	400 passes				10,000 passes			
	0 preliminary		400 preliminary		0 preliminary		10,000 preliminary	
	Mean	CD	Mean	CD	Mean	CD	Mean	CD
β_1	-1.255	4.212	-1.281	-1.491	-1.336	2.083	-1.311	1.151
β_2	.935	-.648	.931	3.226	.936	-1.002	.935	1.600
β_3	9.912	-3.591	9.943	1.368	10.051	-1.754	10.010	-1.160
σ^2	.757	1.455	.718	.390	.721	.672	.715	.125

Initial value: Augmented posterior mode ($\beta_1 = 1.043$, $\beta_2 = .928$, $\beta_3 = 9.624$, $\sigma^2 = .293$)

	400 passes				10,000 passes			
	0 preliminary		400 preliminary		0 preliminary		10,000 preliminary	
	Mean	CD	Mean	CD	Mean	CD	Mean	CD
β_1	-1.231	4.138	-1.304	-.650	-1.320	-1.176	-1.326	.817
β_2	.930	-2.532	.937	.380	.935	.966	.934	1.159
β_3	9.874	-4.021	9.998	-.693	10.025	.947	10.038	-1.071
σ^2	.763	1.570	.711	.507	.717	1.032	.714	1.613

Initial value: Augmented posterior mode, censored σ^2
($\beta_1 = 1.043$, $\beta_2 = .928$, $\beta_3 = 9.624$, $\sigma^2 = .677$)

	400 passes				10,000 passes			
	0 preliminary		400 preliminary		0 preliminary		10,000 preliminary	
	Mean	CD	Mean	CD	Mean	CD	Mean	CD
β_1	-1.298	5.920	-1.305	-3.194	-1.327	2.780	-1.314	-1.946
β_2	.927	-.616	.932	.431	.933	-1.363	.935	1.611
β_3	10.000	-4.781	10.005	3.891	10.038	-2.855	10.017	1.573
σ^2	.787	1.503	.694	-.543	.719	.602	.717	.615

Table 5

Bayesian Inference, Tobit Censored Regression Model

Censored OLS initial value; 10,000 passes; 10,000 preliminary passes

Execution times: Preliminary passes, 96.49; Gibbs sampling, 96.06; Spectral, 17.26

	Passes 1 - 1,000		Passes 5,001-10,000		All passess				
	Mean	St Dev	Mean	St Dev	Mean	St Dev	NSE	RNE	CD
β_1	-1.295	.165	-1.316	.186	-1.311	.179	.0075	.057	1.151
β_2	.940	.062	.934	.059	.935	.059	.0011	.313	1.600
β_3	9.985	.275	10.018	.309	10.010	.300	.0124	.059	-1.160
σ^2	.715	.111	.714	.113	.715	.111	.0023	.226	.125

Augmented mode initial value; 10,000 passes; 0 preliminary passes

Execution times: Preliminary passes, 0.00; Gibbs sampling, 98.00; Spectral, 17.32

	Passes 1 - 1,000		Passes 5,001-10,000		All passess				
	Mean	St Dev	Mean	St Dev	Mean	St Dev	NSE	RNE	CD
β_1	-1.334	.235	-1.308	.202	-1.320	.201	.0088	.052	-1.176
β_2	.937	.063	.933	.060	.935	.060	.0011	.279	.966
β_3	10.040	.418	10.004	.340	10..025	.340	.0146	.054	.947
σ^2	.752	.630	.711	.111	.717	.225	.0047	.231	1.032

Augmented mode initial value; 10,000 passes; 10,000 preliminary passes

Execution times: Preliminary passes, 97.11; Gibbs sampling, 100.43; Spectral, 18.30

	Passes 1 - 1,000		Passes 5,001-10,000		All passess				
	Mean	St Dev	Mean	St Dev	Mean	St Dev	NSE	RNE	CD
β_1	-1.316	.190	-1.332	.195	-1.326	.195	.0086	.052	.817
$\beta 2$	.937	.061	.933	.061	.934	.060	.0011	.325	1.159
$\beta 3$	10.014	.313	10..049	.326	10.038	.327	.0143	.053	-1.071
$\sigma 2$	.728	.120	.715	.111	.714	.111	.0023	.235	1.613

Augmented mode with censored σ^2 initial value; 10,000 passes; 10,000 preliminary passes

Execution times: Preliminary passes, 96.71; Gibbs sampling, 97.18; Spectral, 18.03

	Passes 1 - 1,000		Passes 5,001-10,000		All passess				
	Mean	St Dev	Mean	St Dev	Mean	St Dev	NSE	RNE	CD
β_1	-1.350	.189	-1.312	.187	-1.314	.188	.0083	.052	-1.946
β_2	.940	.062	.934	.061	.935	.060	.0011	.291	1.611
β_3	10.068	.319	10.017	.317	10.017	.317	.0138	.053	1.575
σ^2	.722	.113	.718	.116	.717	.115	.0025	.215	.615

Table 6

Estimated Spectral Densities of Sampled Parameters, Tobit Censored Regression Model

	400 passes and preliminary passes				10,000 passes and preliminary passes			
Frequency	β_1	β_2	β_3	σ^2	β_1	β_2	β_3	σ^2
.00π	.1315	.0090	.4047	.0367	.7360	.0110	2.0270	.0522
.05π	.1305	.0084	.4061	.0368	.2244	.0089	.6178	.0406
.10π	.1252	.0077	.3921	.0333	.0486	.0077	.1369	.03351
.15π	.0956	.0076	.2967	.0256	.0201	.0070	.0557	.0263
.20π	.0314	.0070	.1045	.0225	.0121	.0063	.0382	.0187
.25π	.0130	.0057	.0382	.0147	.0092	.0054	.0285	.0146
.30π	.0085	.0050	.0262	.0121	.0073	.0041	.0223	.0122
.35π	.0056	.0037	.0203	.0104	.0059	.0036	.0175	.0100
.40π	.0044	.0032	.0180	.0085	.0046	.0027	.0153	.0083
.45π	.0039	.0025	.0143	.0074	.0040	.0027	.0138	.0070
.50π	.0039	.0020	.0142	.0062	.0036	.0025	.0119	.0060
.55π	.0036	.0018	.0123	.0052	.0035	.0021	.0105	.0057
.60π	.0031	.0017	.0117	.0046	.0030	.0017	.0107	.0050
.65π	.0027	.0016	.0103	.0038	.0028	.0016	.0105	.0041
.70π	.0029	.0014	.0089	.0036	.0026	.0016	.0098	.0038
.75π	.0029	.0013	.0095	.0032	.0026	.0017	.0098	.0037
.80π	.0030	.0012	.0092	.0036	.0025	.0016	.0091	.0034
.85π	.0026	.0011	.0084	.0034	.0025	.0013	.0092	.0036
.90π	.0023	.0010	.0076	.0030	.0028	.0013	.0084	.0034
.95π	.0020	.0007	.0060	.0024	.0025	.0012	.0084	.0034
1.00π	.0016	.0057	.0047	.0019	.0012	.0006	.0042	.0016