Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling

ALAN E. GELFAND, SUSAN E. HILLS, AMY RACINE-POON, and ADRIAN F. M. SMITH*

The use of the Gibbs sampler as a method for calculating Bayesian marginal posterior and predictive densities is reviewed and illustrated with a range of normal data models, including variance components, unordered and ordered means, hierarchical growth curves, and missing data in a crossover trial. In all cases the approach is straightforward to specify distributionally and to implement computationally, with output readily adapted for required inference summaries.

KEY WORDS: Marginalization; Variance components; Order-restricted inference; Hierarchical models; Missing data; Nonlinear parameters; Density estimation.

1. INTRODUCTION

Technical difficulties arising in the calculation of marginal posterior densities needed for Bayesian inference have long served as an impediment to the wider application of the Bayesian framework to real data. In the last few years there have been a number of advances in numerical and analytic approximation techniques for such calculations—see, for example, Naylor and Smith (1982, 1988), Smith, Skene, Shaw, Naylor, and Dransfield (1985), Smith, Skene, Shaw, and Naylor (1987), Tierney and Kadane (1986), Shaw (1988), and Geweke (1988)—but implementation of these approaches typically requires sophisticated numerical or analytic approximation expertise and possibly specialist software. In a recent article, Gelfand and Smith (1990) described sampling-based approaches for such calculations, which, by contrast, are essentially trivial to implement, even with limited computing resources. In this previous article, we entered caveats regarding the computational efficiency of such sampling-based approaches, but our continuing investigations have shown that adaptive, iterative sampling achieved through the Gibbs sampler (Geman and Geman 1984) is, in fact, surprisingly efficient, converging remarkably quickly for a wide range of problems. That said, our advocacy of the approach rests essentially on its simplicity and universality and not on any claim that it is the most efficient procedure for any given problem.

Our objective in this article is to provide illustrations of a range of applications of the Gibbs sampler in order to demonstrate its versatility and ease of implementation in practice. We begin by briefly reviewing the Gibbs sampler in Section 2. In Section 3, based upon computational experience with a variety of problems, we comment on the problem of assessing the convergence of this iterative algorithm. In Section 4 we begin our illustrative analysis with a variance components model applied to a data set introduced in Box and Tiao (1973), whose Bayesian analysis therein involved elaborate exact and asymptotic methods. In addition, we illustrate the ease with which inferences for functions of parameters, such as ratios, can be made using Gibbs sampling. In Section 5, we take up the k-sample normal means problem in the general case of unbalanced data with unknown population variances. In particular, we show that the previously inaccessible case where the population means are ordered is straightforwardly handled through Gibbs sampling. Application is made to an unbalanced generated data set from normal populations with known ordered means and severely nonhomogeneous variances. In Section 6, we look at a population linear growth curve model, as an illustration of the power of the Gibbs sampler in handling complex hierarchical models. We analyze data on the responses over time of two groups of 30 rats to control and treatment regimes, involving a total of 66 parameters in the hierarchical model specification for each group. In Section 7, we analyze a two-period crossover design involving the comparison of two drug formulations, in order to illustrate the ease with which the Gibbs sampler deals with complications arising from missing data in an originally balanced design. A summary discussion is provided in Section 8.

2. GIBBS SAMPLING

In the sequel, densities will be denoted generically by square brackets, so that joint, conditional, and marginal forms appear, respectively, as $[X, Y]$, $[X \mid Y]$, and $[Y]$. The usual marginalization by integration procedure will be denoted by forms such as $[X] = \int [X \mid Y] * [Y]$. Throughout, we shall be dealing with collections of random variables for which it is known (see, for example, Besag 1974) that specification of all full conditional distributions uniquely determines the full joint density. More precisely, for such a collection of random variables $U_1, U_2, \ldots, U_k$, the joint density $[U_1, U_2, \ldots, U_k]$ is uniquely determined by the full conditional densities $[U_s \mid U_r, r \neq s]$, $s = 1, 2, \ldots, k$. Our interest is in the marginal distributions $[U_s]$, $s = 1, 2, \ldots, k$.

* Alan E. Gelfand is Professor, University of Connecticut, Storrs, CT 06269-3120. Susan E. Hills is Statistician, University of Nottingham. Amy Racine-Poon is Statistician, CIBA-GEIGY AG, Basle, Switzerland. Adrian F. M. Smith is Professor, Imperial College of Science, Technology and Medicine. This research was partly supported by the UK Science and Engineering Research Council's Complex Stochastic Systems Initiative, which, in particular, provided travel support to the first author and supported the second author. Reviewers substantially improved our exposition by insisting on clarifications and more attention to substantive applications.

An algorithm for extracting marginal distributions from the full conditional distribution was formally introduced as the Gibbs sampler in Geman and Geman (1984), although its essence dates at least to Hastings (1970). [Substitution sampling as in Tanner and Wong (1987) is closely related; see Gelfand and Smith (1990).] The algorithm requires all the full conditional distributions to be "available" for sampling, where "available" is taken to mean that, for example, samples of U_s can be generated straightforwardly and efficiently given specified values of the conditioning variables, U_r, $r \neq s$.

Gibbs sampling is a Markovian updating scheme that proceeds as follows. Given an arbitrary starting set of values $U_1^{(0)}, \ldots, U_k^{(0)}$, we draw $U_1^{(1)}$ from $[U_1 \mid U_2^{(0)}, \ldots, U_k^{(0)}]$, then $U_2^{(1)}$ from $[U_2 \mid U_1^{(1)}, U_k^{(0)}]$, and so on up to $U_k^{(1)}$ from $[U_k \mid U_1^{(1)}, \ldots, U_{k-1}^{(1)}]$ to complete one iteration of the scheme. After t such iterations we would arrive at $(U_k^{(t)}, \ldots, U_k^{(t)})$. Geman and Geman showed under mild conditions that $U_s^{(t)} \xrightarrow{d} U_s \sim [U_s]$ as $t \to \infty$. Thus, for t large enough we can regard $U_s^{(t)}$ as a simulated observation from $[U_s]$.

Independently replicating this process m times produces m iid k-tuples $U_{1j}^{(t)}, \ldots, U_{kj}^{(t)}$, $j = 1, \ldots, m$. For any s, the collection $U_{s1}^{(t)}, \ldots, U_{sm}^{(t)}$ can be viewed as a simulated sample from $[U_s]$. The marginal density is then estimated by the finite mixture density

$$[\hat{U}_s] = m^{-1} \sum_{j=1}^{m} [U_s \mid U_r = U_{rj}^{(t)}, r \neq s]. \tag{1}$$

[See Gelfand and Smith (1990) for further discussion.] Since the expression (1) can be viewed as a "Rao–Blackwellized" density estimator, relative to the more usual kernel density estimators based upon $U_{sj}^{(t)}$, $j = 1, \ldots, m$, the estimation is high.

Suppose interest centers on the marginal distribution for a variable V that is a function $g(U_1, \ldots, U_k)$ of $U_1, \ldots, U_k$. We note that evaluation of g at each of the $(U_{1j}^{(t)}, \ldots, U_{kj}^{(t)})$ provides samples of V, so that an ordinary kernel density estimate can readily be calculated (see Section 4 for an illustration of this). A density estimate of the form (1) can also be obtained. Simply choose any U_s that is an argument of g and then make the transformation from $[U_s \mid U_r, r \neq s]$ to $[V \mid U_r, r \neq s]$.

All applications we consider in this article are within the Bayesian framework, where the U_s are unobservable, representing either parameters or missing data (and V can thus be a function of the parameters in which we are interested). All distributions will be viewed as conditional on the observed data, whence marginal distributions become the marginal posteriors needed for Bayesian inference or prediction.

In addition, we note that all the applications in this article assume a hierarchical model with conjugate priors and hyperpriors, the latter being typically arbitrarily vague. Such hierarchical structure, to some extent, mitigates robustness concerns associated with the use of the conjugate first-stage prior (Berger 1985, p. 231). However, it is important to note that while conjugacy simplifies the implementation of the Gibbs sampler, it is not an essential element. In *any* hierarchical Bayes model the full conditional distribution of any parameter is always identifiable from the joint density of the data and the parameters modulo normalizing constant. Using more sophisticated random variate generation approaches, such as the ratio of uniforms method (Devroye 1986), we can sample from arbitrary nonnormalized densities, although, of course, fine tuning of the sampling methodology, including "clever" reparameterization, may be required in order to avoid highly inefficient random variate generation. For the present, we ignore such refinements and concentrate on a clear exposition of the basic methodology in a range of familiar, but widely applicable, settings.

3. CONVERGENCE ISSUES

Complete implementation of the Gibbs sampler requires that a determination of t be made and that, across iterations, choice(s) of m be specified. In this regard it is important to distinguish the assessment of convergence for any individual data application from the broader goal of developing on-line, automated, interactive software to determine satisfactory convergence.

Our extensive experience with a wide range of particular applications suggests that accomplishing the former is not a problem. We note that appropriate values for t and m depend upon the particular application and cannot be specified in advance. All of the examples discussed in this article, however, were handled with $t \leq 50$ and $m \leq 1{,}000$. Since random variate generation is generally inexpensive, we expect to experiment with different settings. Indeed, since interest focuses heavily on the application of this sampling-based methodology to previously inaccessible problems where we often have no benchmarks or alternatives with which to compare our results, such experimentation seems necessary.

The following discussion describes a means of assessing convergence that, though naive and less rigorously defined than might be desired, has been successful in a considerable number of applications including those in the subsequent sections. We monitor the generated data in a univariate fashion, allowing the sampler to run until we feel that the marginal posterior distributions for each parameter of interest are converged. We do this in an elementary manner. For a fixed m we increase t, overlay plots of the resulting estimated densities (1), and see if the estimates are visually indistinguishable. Similarly, we also increase m to assess stability of the density estimate. In our experimentation with a wide range of problems and both real and simulated data sets, we have never required $t > 50$. We tend to hold m somewhat small (often as small as 25 and at most 200) until convergence is indicated, at which point, for a final iteration, we typically increase m by an order of magnitude to obtain our density estimate (1). This final sampling is achieved, in the context of say, $[U_s]$, by systematic drawing with replacement from among the observed vectors $\{U_{rj}, r \neq s\}$, $j = 1, \ldots, m$. For each such draw an observation is taken from the resulting full con-

ditional distribution for U_s. Univariate plots are drawn by selecting between 40 and 100 equally spaced points in the effective domain of the variable. We then evaluate the density estimate [of the form (1)] at these points and a spline-smoothed curve is drawn through these values. By effective domain, we mean the interval where, say, 99% of the mass lies. We occasionally require several passes to determine this domain but rarely require more than 100 points to obtain a satisfying plot. Clearly, this plotting method could be refined. In this regard, we also recommend a convenient check on calculations by performing a simple trapezoidal integration on the collection of estimated density values associated with these points to verify that the result is very close to 1.

Issues of higher-dimensional convergence require further study. A related question concerning the relationship between the rate of convergence and the dimensionality k also has no simple answer. In practical situations the number of hierarchical levels, the extent of exchangeability, the degree of agreement between the data and the prior, and so forth, all influence any conclusion.

Finally, the development of automated convergence assessment is a much harder problem. We make no present claim as to the most effective convergence diagnostics. Our ongoing work in attempting to automate the entire Gibbs sampler has led us to a large-scale investigation of a wide range of measures. These include empirical Q–Q plots and nonparametric two-sample tests and involve such extremes as pooling successive samples to "robustify" the process and comparing iterations, say, 5–10 cycles apart to "reduce dependence."

4. VARIANCE COMPONENTS PROBLEMS

Random effects models are very naturally modeled within the Bayesian framework. Nonetheless, calculation of the marginal posterior distributions of variance components and functions of variance components has proved a challenging technical problem. Box and Tiao (1973) reported a substantial amount of detailed, sophisticated approximation work, both analytic and numerical. Skene (1983) considered purpose-built numerical techniques. The methods described by Smith et al. (1985, 1987) require careful reparameterization dependent upon both the data and the choice of prior. In a similar spirit, Achcar and Smith (1990) discussed parameter transformations for successful implementation of Laplace's method (Tierney and Kadane 1986). By comparison, the Gibbs sampling approach is remarkably simple.

We shall illustrate the approach with a model involving only two variance components, but it will be clear that the development for more complicated models is no more difficult. Consider, then, the variance components model defined by

$$Y_{ij} = \theta_i + e_{ij}, \qquad i = 1, \ldots, k; j = 1, \ldots, J, \quad (2)$$

where, assuming conditional independence throughout, $[\theta_i \mid \mu, \sigma^2] = N(\mu, \sigma_\theta^2)$ and $[e_{ij} \mid \sigma_e^2] = N(0, \sigma_e^2)$. Assume that $\theta = (\theta_1, \ldots, \theta_K)$ and $Y = (Y_{11}, \ldots, Y_{KJ})$ and that μ, σ^2, and σ_e^2 are independent with priors $[\mu] = N(\mu_0, \sigma_0^2)$, $[\sigma_\theta^2] = \mathrm{IG}(a_1, b_1)$, and $[\sigma_e^2] = \mathrm{IG}(a_2, b_2)$ (e.g., Hill 1965), where IG denotes the inverse gamma distribution and μ_0, σ_0^2, a_1, b_1, a_2, and b_2 are assumed known. It is then straightforward to verify that the Gibbs sampler is specified by

$$[\sigma_\theta^2 \mid Y, \mu, \theta, \sigma_e^2] = \mathrm{IG}(a_1 + \tfrac{1}{2}K, b_1 + \tfrac{1}{2}\Sigma(\theta_i - \mu)^2)$$

$$[\sigma_e^2 \mid Y, \mu, \theta, \sigma_\theta^2] = \mathrm{IG}(a_2 + \tfrac{1}{2}KJ, b_2 + \tfrac{1}{2}\Sigma\Sigma(Y_{ij} - \theta_i)^2)$$

$$[\mu \mid Y, \theta, \sigma_\theta^2, \sigma_e^2] = N\left(\frac{\sigma_\theta^2\mu_0 + \sigma_o^2\Sigma\theta_i}{\sigma_\theta^2 + K\sigma_o^2}, \frac{\sigma_\theta^2\sigma_o^2}{\sigma_\theta^2 + K\sigma_o^2}\right)$$

$$[\theta \mid Y, \mu, \sigma_\theta^2, \sigma_e^2] = N\left(\frac{J\sigma_\theta^2}{J\sigma_\theta^2 + \sigma_e^2}\bar{Y} + \frac{\sigma_e^2}{J\sigma_\theta^2 + \sigma_e^2}\mu 1, \frac{\sigma_\theta^2\sigma_e^2}{J\sigma_\theta^2 + \sigma_e^2} I\right), \quad (3)$$

where $\bar{Y} = (\bar{Y}_1, \ldots, \bar{Y}_K)$, $\bar{Y}_i = \Sigma_j Y_{ij}/J$, 1 is a $K \times 1$ column vector of 1s, and I is a $K \times K$ identity matrix. In particular, in (3) we can allow the a_i and/or b_i to be equal to 0, representing a range of conventional improper priors for σ_θ^2 and σ_e^2.

Box and Tiao (1973, sec. 5.1.3) introduced two data sets for which the model (2) is appropriate. The second and more difficult set is generated from random normal deviates in which $\mu = 5$, $\sigma_\theta^2 = 4$, and $\sigma_e^2 = 16$. The resultant data, summarized in Table 1, are badly behaved, in that the standard (analysis of variance based) unbiased estimate of σ_θ^2 is negative, rendering inference about σ_θ^2 dif-

Table 1. Generated Data

	$K = 6, J = 5, \bar{Y}_{..} = 5.6656$					
Batch	*1*	*2*	*3*	*4*	*5*	*6*
$\bar{Y}$	6.2268	4.6560	7.5212	5.6848	6.0796	3.8252
S^2	8.8650	25.4900	25.6359	7.0935	14.3590	8.2691

Source	*SS*	*df*	*MS*
Between batches	41.6816	5	8.3363
Within batches	358.7014	24	14.9459
Total	400.3830	29	

$\hat{\sigma}_e^2 = 14.9459$, $\hat{\sigma}_\theta^2 = -1.3219$

NOTE: *SS* denotes sum of squares; *MS* denotes mean squares.
Source: Box and Tiao (1973, p. 247).

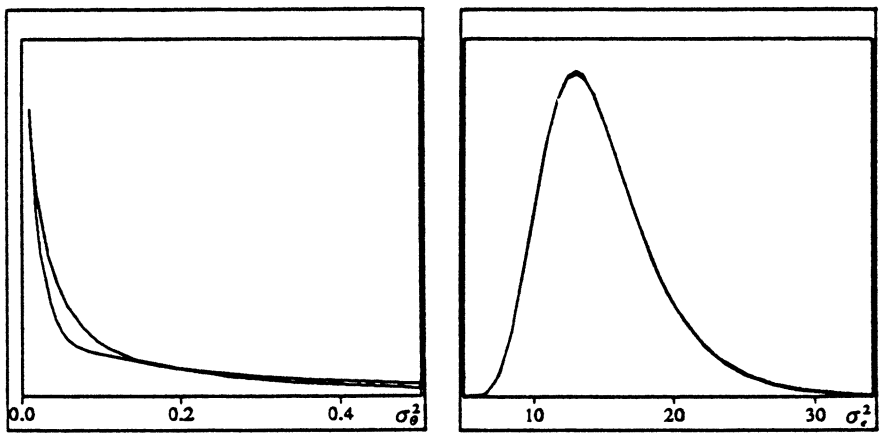

Figure 1. Convergence of Estimated Densities of Variance Components Under Prior Specification I.

ficult. I shall use this example to provide a challenging low-dimensional test of the Gibbs sampler.

For illustrative purposes, we provide a Bayesian analysis based on the prior specification $[\sigma_e^2] = \text{IG}(0, 0)$, $[\mu] = N(0, 10^{12})$, together with either I: $[\sigma_\theta^2] = \text{IG}(0, 0)$ or II: $[\sigma_\theta^2] = \text{IG}(\frac{1}{2}, b_1)$. Under I, we have the improper prior for $(\sigma_\theta^2, \sigma_e^2)$ suggested by Hill (1965), which is a naive two-dimensional extension of the familiar noninformative prior for a variance. Under II, we have a proper weak independent inverse chi-squared prior for σ_θ^2 which, depending on b_1, "supports" or "differs from" the data [see Skene (1983) for further detailed discussion]. The two priors for σ_θ^2 differ considerably. Under I, $[\sigma_\theta^2]$ is one-tailed, giving strong weight to the assertion that σ_θ^2 is near 0. Since this is weakly confirmed by the data, the marginal posterior (Fig. 1a) reflects this prior. Under II, $[\sigma_\theta^2]$ is two-tailed, having mode at $2b_1/3$. Interestingly, experimentation with b_1 varying up to 6 leads to an outcome similar to that under I. For all such b_1, the prior is virtually reproduced as the posterior (see Fig. 2a for the case $b_1 = 1$). The data provide very little information about σ_θ^2.

Our experience with Gibbs sampling in this context is very encouraging. Under both I and II (with $b_1 = 1$), the iterative approach had no difficulty with the extreme skewness in the resultant posterior of σ_θ^2. In repeated experiments, overall convergence was achieved under I within, at most, 20 iterations using $m = 100$ and under II within, at most, 10 iterations using $m = 100$. We demonstrate this in Figure 1, which, for case I, compares density estimates after 20 and 40 iterations for σ_e^2 and σ_θ^2. Figure 2 presents the corresponding curves for case II after 10 and 20 iterations.

The variance ratio, $\sigma_\theta^2/\sigma_e^2$, or perhaps the intraclass correlation coefficient, $\sigma_\theta^2/(\sigma_\theta^2 + \sigma_e^2)$, is often a quantity of interest. Remarks at the end of Section 2 show that obtaining the marginal posterior distribution for such variables is easily accomplished by, for example, taking σ_e^2 fixed and making a one-to-one transformation from σ_θ^2. Figure 3 shows the estimated density for the variance ratio under both I and II obtained after 20 iterations when $m = 1{,}000$, the untypically large value of m arising from the awkward shape of the posterior. A density estimator with normal kernels was used with window width suggested by Silverman (1986, p. 48).

As we indicated at the beginning of this section, there are several alternative ways of implementing a Bayesian analysis in this context, and we make no claim that the Gibbs sampler method is the most efficient. Note, however, that the "efficiency" of the other approaches is at the expense of detailed sophisticated applied analysis (Box

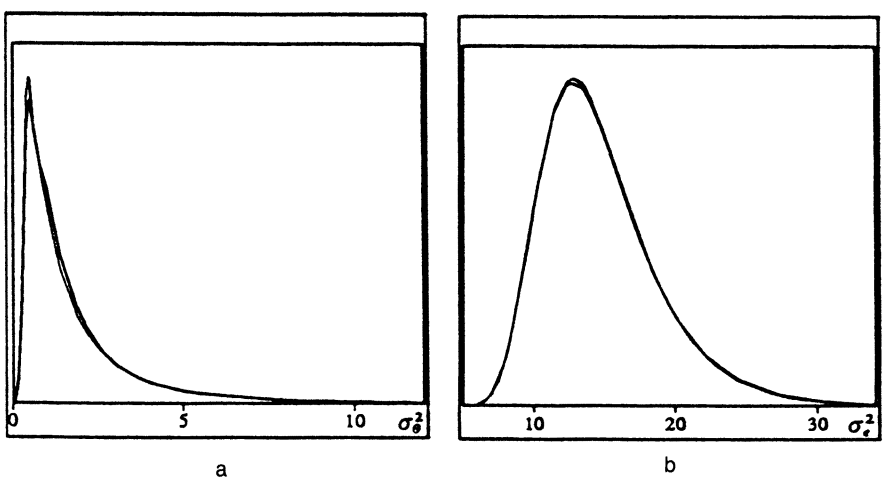

Figure 2. Convergence of Estimated Densities of Variance Components Under Prior Specification II.

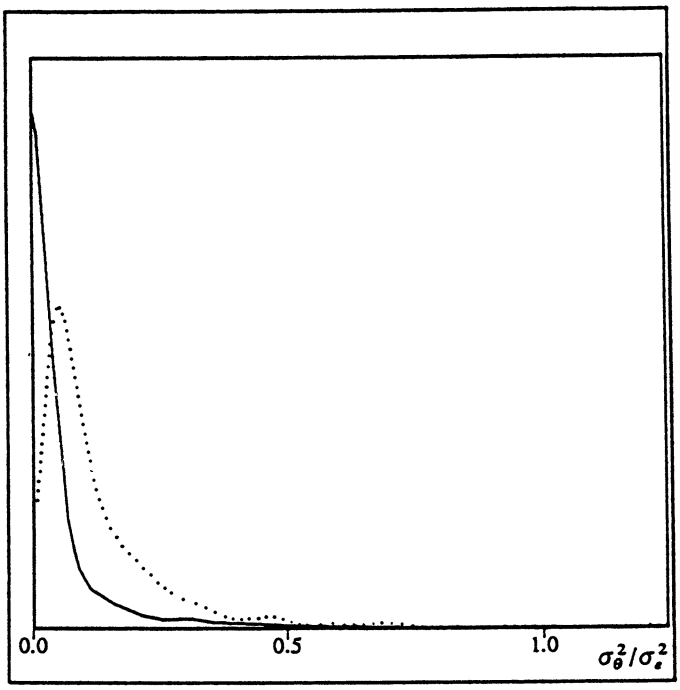

Figure 3. Estimated Density of the Variance Ratio $\sigma_\theta^2/\sigma_e^2$ for the Variance Components Problem: ——, Prior Specification I; ···, Prior Specification II.

and Tiao 1973) or tailored "one-off" numerical tricks (Skene 1983) or sophisticated adaptive quadrature methodology (Smith et al. 1987), in conjunction with subtle sensitivity to parameterization issues (see also Achcar and Smith 1989). By contrast, the Gibbs sampler approach is trivially implemented, without requiring any special one-off effort or sophisticated analytic or numerical insight on the part of the user. In addition, in the case of most of the more sophisticated techniques, substantial fresh effort is required (including, in some cases, beginning the analysis anew) if the focus of inferential interest changes [e.g., from σ_θ^2, σ_e^2 to $\sigma_\theta^2/\sigma_e^2$ or $\sigma_\theta^2/(\sigma_\theta^2 + \sigma_e^2)$]. By contrast, the sample-based nature of the Gibbs sampler enables such-a shift of focus to be accomplished with, essentially, no further computational effort. Thus, even in cases in which the Gibbs sampler is not the only approach available, its twofold virtues of simplicity of implementation and flexibility of inference response may more than compensate for its relative inefficiency.

5. NORMAL MEANS PROBLEMS

The comparison of means presumed from normal populations is arguably the most ubiquitous model in statistical inference, but issues such as unbalanced sampling and heterogeneity of variances have typically forced compromises in frequentist and empirical Bayes approaches. Historically, this has also been somewhat true in the purely Bayesian setting. Frequently, with regard to variance parameters the proper Bayesian procedure of marginalization by integration has been replaced by point estimation to reduce the dimensionality of numerical integrations needed to obtain marginal posterior distributions for mean parameters. Gibbs sampling provides a means of performing such integrations without having to make approximations. The Gibbs sampler was introduced in the context of problems of very high dimension (such as image reconstruction, expert systems, neural networks) and has been successful in such contexts. Its encouraging performance in our investigations is therefore not surprising, since even a large multiparameter Bayesian problem is of small dimension compared to typical image-processing problems.

In this section, we consider the comparison of I population means which, in conjunction with distinct unknown population variances and an exchangeable prior, results in a $2I + 2$ parameter problem. We show that the implementation of the Gibbs sampler is straightforward. The more general case in which the population means are represented as linear functions of a set of explanatory variables can be handled similarly, using by-now-familiar distribution theory given by, for example, Lindley and Smith (1972). Such an example appears in Section 6.

Often there are implicit order restrictions on the means to be compared. For instance, it may be known that the means are increasing as we traverse the populations from $i = 1$ up to I. If we incorporate this information into our prior specification using order statistics, the integrations required for marginalization are typically beyond the capacity of current numerical and analytic approximation methodology. I shall show, however, the Gibbs sampler is still straightforwardly implemented, since normal full conditionals are simply replaced by truncated normals.

The requisite distribution theory, assuming no order restrictions on the means, is as follows. Assuming conditional independence throughout, assume that $[Y_{ij} \mid \theta_i, \sigma_i^2] = N(\theta_i, \sigma_i^2)$ $(i = 1, \ldots, I; j = 1, \ldots, n_i)$, $[\theta_i \mid \mu, \tau^2] = N(\mu, \tau^2)$, $[\sigma_i^2] = \mathrm{IG}(a_1, b_1)$, $[\mu] = N(\mu_0, \sigma_0^2)$, and $[\tau^2] = \mathrm{IG}(a_2, b_2)$, where IG denotes the inverse gamma distribution and a_1, a_2, b_1, b_2, μ_0, and σ_0^2 are assumed known (often chosen to represent conventional improper prior forms; see Sec. 4). By sufficiency, we confine attention to $\overline{Y}_i = \Sigma_j Y_{ij}/n_i$ and $S_i^2 = \Sigma(Y_{ij} - \overline{Y}_i)^2/(n_i - 1)$. Assuming that $\theta = (\theta_1, \ldots, \theta_I)$, $\sigma^2 = (\sigma_1^2, \ldots, \sigma_I^2)$, and $Y = (\overline{Y}_1, \ldots, \overline{Y}_I, S_1^2, \ldots, S_I^2)$, we have, for given data Y, the following full conditional distributions:

$$[\theta \mid Y, \sigma^2, \mu, \tau^2] = N(\theta^*, D^*), \tag{4}$$

where

$$\theta_i^* = \frac{n_i \overline{Y}_i \tau^2 + \mu \sigma_i^2}{n_i \tau^2 + \sigma_i^2},$$

$$D_{ii}^* = \frac{\sigma_i^2 \tau^2}{n_i \tau + \sigma_i^2}, \qquad D_{ij}^* = 0, \quad i \neq j,$$

and

$$[\sigma^2 \mid Y, \theta, \mu, \tau^2] = \prod_{i=1}^{I} [\sigma_i^2 \mid \overline{Y}_i, S_i^2, \theta_i],$$

where

$$[\sigma_i^2 \mid \overline{Y}_i, S_i^2, \theta_i] = \mathrm{IG}\left(a_1 + \tfrac{1}{2} n_i, b_1 + \tfrac{1}{2} \sum_{j=1}^{n_i} (Y_{ij} - \theta_i)^2\right),$$

$$[\mu \mid Y, \theta, \sigma^2, \tau^2] = N\left(\frac{\tau^2\mu_0 + \sigma_0^2 \Sigma\theta_i}{\tau^2 + I\sigma_0^2}, \frac{\tau^2\sigma_0^2}{\tau^2 + I\sigma_0^2}\right),$$

and

$$[\tau^2 \mid Y, \theta, \sigma^2, \mu] = \mathrm{IG}(a_2 + \tfrac{1}{2}I, b_2 + \tfrac{1}{2}\Sigma(\theta_i - \mu)^2).$$

Suppose now that the means are known to be ordered, say, $\theta_1 < \theta_2 < \cdots < \theta_I$; see Barlow, Bartholomew, Bremner, and Brunk (1972) and Robertson, Wright, and Dykstra (1988) for applications and extensive references to the problem. If we assume as our prior that the θ_i arise as order statistics from a sample of size I from $N(\mu, \tau^2)$, then it is straightforward to show that $[\theta_i \mid Y, \theta_j\ (j \neq i), \sigma^2, \mu, \tau^2]$ is now precisely the marginal normal distribution in (4), but restricted to the interval $[\theta_{i-1}, \theta_{i+1}]$ (in which we adopt the convention $\theta_0 = -\infty$, $\theta_{I+1} = +\infty$) and so again is straightforwardly available for sampling. The full conditional distributions for σ^2, τ^2, and μ remain exactly as previously.

In sampling from the truncated normal distribution, the rejection method (discarding ineligible observations sampled from the nontruncated distribution) will tend to be wasteful and slow, particularly if $\theta_{i+1} - \theta_{i-1}$ is small. To draw an observation from $N(c, d^2)$ restricted to (a, b) a convenient "one-for-one" sampling method is the following (Devroye 1986). Generate U, a random uniform (0, 1) variate and calculate $Y = c + d\Phi^{-1}(p(U; a, b, c, d))$, where

$$p(U; a, b, c, d) = \Phi\left(\frac{a - c}{d}\right) + U\left(\Phi\left(\frac{b - c}{d}\right) - \Phi\left(\frac{a - c}{d}\right)\right),$$

with Φ denoting the standard normal cdf. It is straightforward to show that Y has the desired distribution. These ideas are easily extended to give a general account of Bayesian analysis for order-restricted parameters.

To study the performance of Gibbs sampling in the preceding setting, we analyzed generated data so as to be able to calibrate the results against the known situation. For the purpose of illustration, we created a rather unbalanced, extremely nonhomogeneous data set by setting $I = 5$ and, for the ith population, $i = 1, \ldots, 5$ drawing $n_i = 2i + 4$ independent observations from $N(i, i^2)$. The simulated data are summarized in Table 2, and note, in particular, the inversion of order of the sample means, $\bar{Y}_4$ and $\bar{Y}_5$.

For illustration, we specified priors $[\mu] = N(0, 10^5)$, $[\sigma_i^2] = \mathrm{IG}(\frac{1}{2}, 1)$, and $[\tau^2] = \mathrm{IG}(\frac{1}{2}, 1)$. For the Gibbs sampler convergence was achieved within 10 iterations for the unordered case using $m = 100$. The ordered case required at most 20 iterations, again using $m = 100$, except for $[\theta_4 \mid Y]$ and $[\theta_5 \mid y]$, which over repeated experiments required values of m ranging from 200 to 1,000. Rather than graphically documenting the convergence in this case, we compare the unordered and ordered marginal posteriors. Let $[\theta_i \mid Y]_u$ and $[\theta_i \mid Y]_o$ denote, respectively, the density estimates for the unordered and ordered cases. In Figure 4a we consider, for example, θ_2 and see that $[\theta_2 \mid Y]_u$ and $[\theta_2 \mid Y]_o$ have roughly the same mode but that $[\theta_2 \mid Y]_o$ is less dispersed. Using the order information results in a sharper inference. In Figure 4b, we consider both θ_4 and

Table 2. Summary of Simulated Data for Normal Means Problem

Sample	1	2	3	4	5
n_i	6	8	10	12	14
$\bar{Y}_i$	.3191	2.034	3.539	6.398	4.811
S_i^2	.2356	2.471	5.761	8.758	19.670

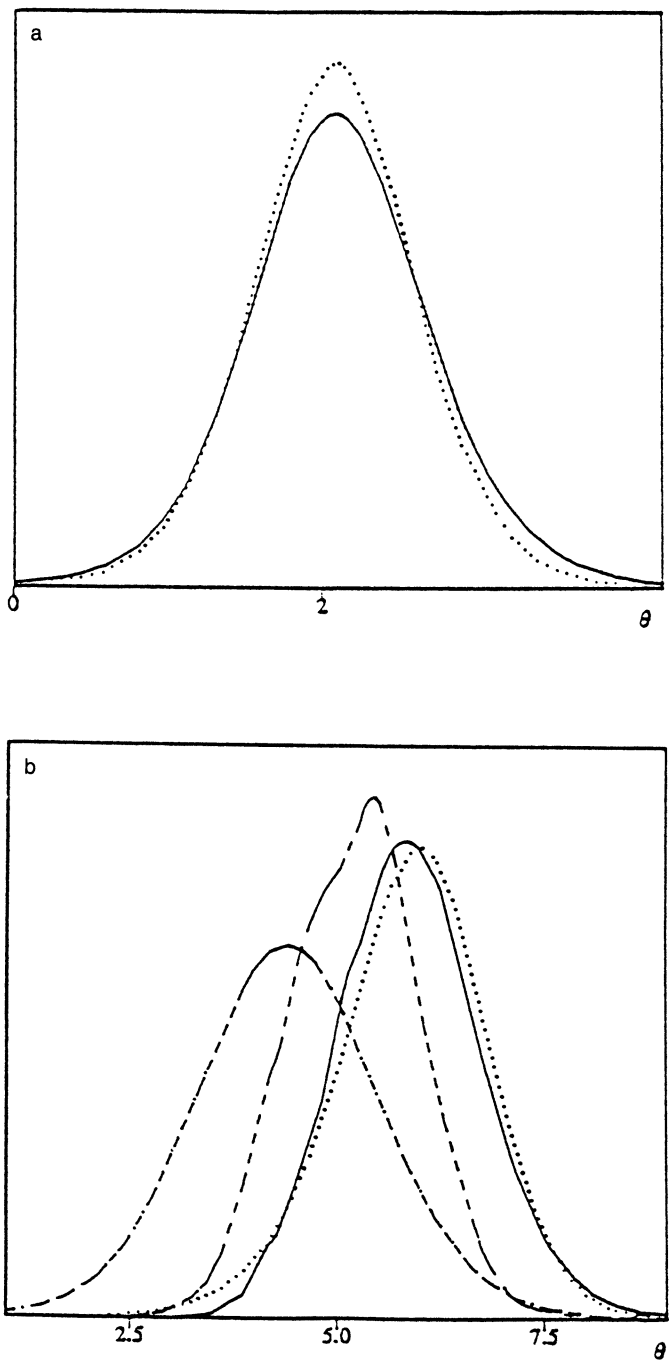

Figure 4. Comparison of Estimated Densities of Means: Unordered and Ordered Cases. (a) ——, $[\theta_2/Y]_u$; ···, $[\theta_2/Y]_o$. (b) ···, $[\theta_4/Y]_u$; –·–, $[\theta_5/Y]_u$; ---, $[\theta_4/Y]_o$; ——, $[\theta_5/Y]_o$.

θ_5. As would be expected, given the sufficient statistics, $[\theta_5 \uparrow Y]_u$ lies to the left of $[\theta_4 \uparrow Y]_u$ and is very dispersed. Using the order information places $[\theta_4 \mid Y]_o$ and $[\theta_5 \uparrow Y]_o$ in the proper stochastic order, pulls the modes in the correct direction, and reduces dispersion. The resultant Bayesian point and interval estimation are improved.

6. A HIERARCHICAL MODEL

Applications of hierarchical models of the kind introduced by Lindley and Smith (1972) abound in fields as diverse as educational testing (Rubin 1981), cancer studies (DuMouchel and Harris 1983), and biological growth curves (Strenio, Weisberg, and Bryk 1983). However, both Bayesian and empirical Bayesian methodologies for such models are typically forced to invoke a number of approximations, whose consequences are often unclear under the multiparameter likelihoods induced by the modeling. See, for example, Morris (1983), Racine-Poon (1985), and Racine-Poon and Smith (1990) for details of some approaches to implementing hierarchical model analysis. By contrast, a full implementation of the Bayesian approach is easily achieved using the Gibbs sampler, at least for the widely used normal linear hierarchical model structure.

For illustration, we focus on the following population growth problem. In a study conducted by the CIBA-GEIGY company, the weights of 30 young rats in a control group were measured weekly for five weeks. The data are given in Table 3, with weight measurements available for all five weeks. Later we discuss the substantive problem of comparison with data from a treatment group. Initially, however, we shall focus attention on the control group in order to illustrate the Gibbs sampling methodology.

For the time period considered, it is reasonable to assume individual straight-line growth curves, although nonlinear curves can be handled as well. We also assume homoscedastic normal measurement errors (nonhomogeneous variances can be accommodated as in the previous section), so that

$$Y_{ij} \sim N(\alpha_i + \beta_i x_{ij}, \sigma_c^2), \quad i = 1, \ldots, k; j = 1, \ldots, n_i,$$

provides the full measurement model (with $k = 30$, $n_i = 5$, and x_{ij} denoting the age in days of the ith rat when measurement j was taken). The population structure is modeled as

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \sim N\left\{ \begin{pmatrix} \alpha_c \\ \beta_c \end{pmatrix}, \Sigma_c \right\}, \qquad i = 1, \ldots, k,$$

assuming conditional independence throughout. A full Bayesian analysis now requires the specification of a prior for σ_c^2, $\mu_c = (\alpha_c, \beta_c)^T$, and Σ_c. Typical inferences of interest in such studies include marginal posteriors for the population parameters (α_c, β_c) and predictive intervals for individual future growth given the first-week measurement. We shall see that these are easily obtained using the Gibbs sampler.

For the prior specification, we assume independence, as is customary, taking

$$[\mu_c, \Sigma_c^{-1}, \sigma_c^2] = [\mu_c][\Sigma_c^{-1}][\sigma_c^2]$$

to have a normal–Wishart–inverse-gamma form:

$$[\mu_c] = N(\eta, C),$$

$$[\Sigma_c^{-1}] = W((\rho R)^{-1}, \rho),$$

$$[\sigma_c^2] = IG\left(\frac{\nu_o}{2}, \frac{\nu_o \tau_o^2}{2}\right).$$

Rewriting the measurement model for the ith individual as $Y_i \sim N(X_i\theta_i, \sigma_c^2 I_{n_i})$ where $\theta_i = (\alpha_i, \beta_i)^T$ and X_i denotes the appropriate design matrix, and defining

$$Y = (Y_1, \ldots, Y_k)^T, \qquad \bar{\theta} = k^{-1} \sum_{i=1}^{k} \theta_i, \qquad n = \sum_{i=1}^{k} n_i,$$

$$D_i = \sigma_c^{-2} X_i^T X_i + \Sigma_c^{-1},$$

$$V = (k\Sigma_c^{-1} + C^{-1})^{-1},$$

the Gibbs sampler for $\theta = (\theta_1, \ldots, \theta_k)$, Σ_c, and σ_c^2 (a total of 66 parameters in the above example) is straightforwardly seen to be specified by the conditional distributions

$$[\theta_i \mid Y, \mu_c, \Sigma_c^{-1}, \sigma_c^2] = N\{D_i(\sigma_c^{-2} X_i^T Y_i + \Sigma_c^{-1}\mu_c), D_i\} \qquad i = 1, \ldots, k,$$

Table 3. Rat Population Growth Data: Control Group

Rat	x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}	Rat	x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}
1	151	199	246	283	320	16	160	207	248	288	324
2	145	199	249	293	354	17	142	187	234	280	316
3	147	214	263	312	328	18	156	203	243	283	317
4	155	200	237	272	297	19	157	212	259	307	336
5	135	188	230	280	323	20	152	203	246	286	321
6	159	210	252	298	331	21	154	205	253	298	334
7	141	189	231	275	305	22	139	190	225	267	302
8	159	201	248	297	338	23	146	191	229	272	302
9	177	236	285	340	376	24	157	211	250	285	323
10	134	182	220	260	296	25	132	185	237	286	331
11	160	208	261	313	352	26	160	207	257	303	345
12	143	188	220	273	314	27	169	216	261	295	333
13	154	200	244	289	325	28	157	205	248	289	316
14	171	221	270	326	358	29	137	180	219	258	291
15	163	216	242	281	312	30	153	200	244	286	324

NOTE: $x_{i1} = 8$, $x_{i2} = 15$, $x_{i3} = 22$, $x_{i4} = 29$, $x_{i5} = 36$ days; $i = 1, \ldots, 30$.

$$[\mu_c \mid Y, \{\theta\}, \Sigma_c^{-1}, \sigma_c^2] = N\{V(k\Sigma_c^{-1}\bar{\theta} + C^{-1}\eta), V\},$$

$$[\Sigma_c^{-1} \mid Y, \{\theta\}, \mu_c, \sigma_c^2] = W\left\{\left[\sum_i (\theta_i - \mu_c)(\theta_i - \mu_c)^T + \rho R\right]^{-1}, k + \rho\right\},$$

$$[\sigma^2 \mid Y, \{\theta\}, \mu_c, \Sigma_c^{-1}] = IG\left(\frac{n + v_o}{2}, \frac{1}{2}\left[\sum_i (Y_i - X_i\theta_i)^T(Y_i - X_i\theta_i) + v_o\tau_o^2\right]\right). \quad (5)$$

For the analysis of the rat growth data given above, the hyperparameter prior specification was defined by

$$C^{-1} = 0, \quad v_o = 0, \quad \rho = 2, \quad R = \begin{pmatrix} 100 & 0 \\ 0 & 0.1 \end{pmatrix},$$

reflecting rather vague initial information relative to that to be provided by the data. Simulation from the Wishart distribution for the 2×2 matrix Σ_c^{-1} is easily accomplished using the algorithm of Odell and Feiveson (1966): with $G(\cdot,\cdot)$ denoting gamma distributions, draw independently from

$$[U_1] = G\left(\frac{v}{2}, \frac{1}{2}\right),$$

$$[U_2] = G\left(\frac{v - 1}{2}, \frac{1}{2}\right),$$

and

$$[N] = N(0, 1);$$

set

$$W = \begin{bmatrix} U_1 & N\sqrt{U_1} \\ N\sqrt{U_1} & U_2 + N^2 \end{bmatrix};$$

then if $S^{-1} = (H^{1/2})^T(H^{1/2})$,

$$\Sigma_c^{-1} = (H^{1/2})^T W(H^{1/2}) \sim W(S^{-1}, v).$$

The iterative process was monitored by observing empirical Q–Q plots for successive samples from α_c, β_c, σ_c^2, and the eigenvalues of Σ_c^{-1}. Though the α_i and β_i are of less interest, spot checking revealed satisfactory convergence, not surprising in view of (5), which suggests that convergence for the θ_i is comparable to that of μ_c. For the data set summarized in Table 3, convergence was achieved with about 35 cycles of $m = 50$ drawings.

As we remarked earlier, a full Bayesian analysis of structured hierarchical models involving convariates has hitherto presented difficulties and a number of Bayes/empirical Bayes approximation methods have been proposed. Racine-Poon and Smith (1990) reviewed a number of these and demonstrated, with a range of real and simulated data analyses, that the EM-type algorithm given by Racine-Poon (1985) is among the best of these proposed approximations. However, it can be seen from Figure 5, where we present the estimated posterior marginals for the population parameters, that, even with this fairly substantial data set of 30×5 observations, the EM-type approximation is not really an adequate substitute for the more refined numerical approximation provided by the Gibbs sampler. [Here, the EM-based "posterior density" is the normal conditional form given in (5) with the con-

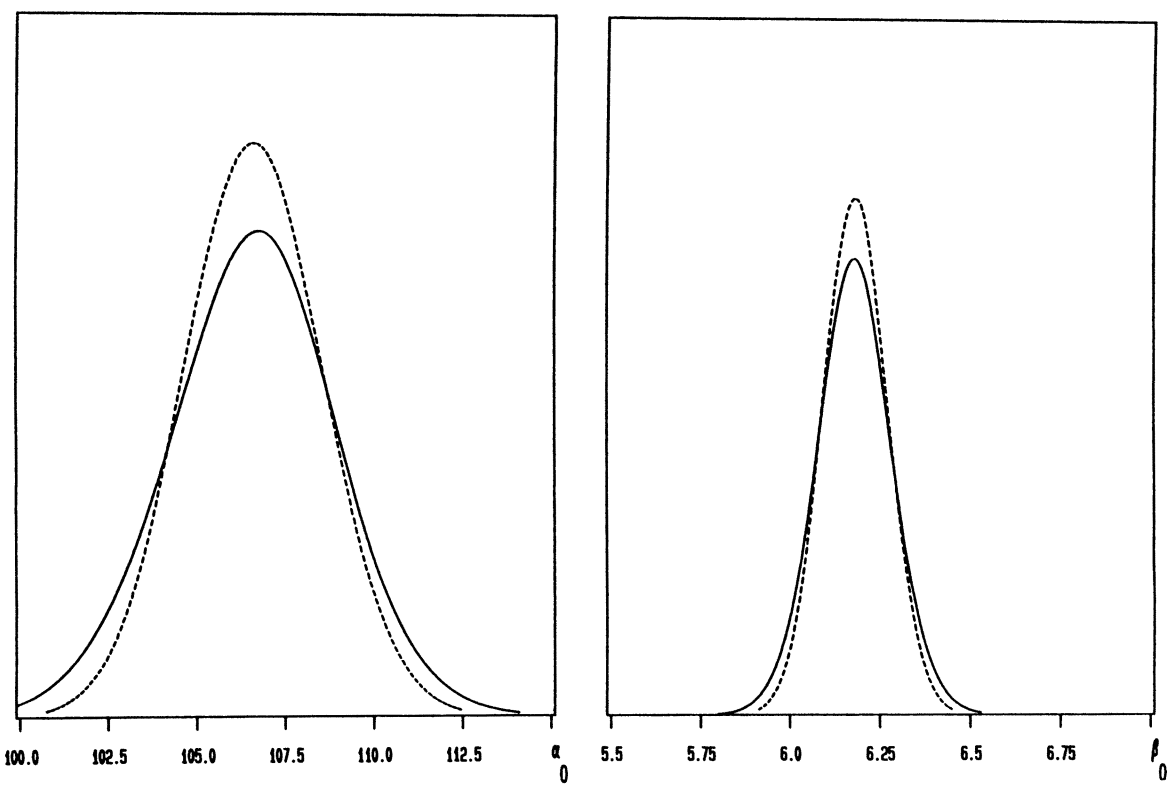

Figure 5. Estimated Densities for Population Initial Weight and Growth Rate for 150 Observation Case: ——, Gibbs Sampler; ---, EM.

verged estimates from the Racine–Poon algorithm substituted for the conditioning parameters.]

To underline further the effectiveness of the Gibbs sampler, and the danger of point estimation based on approximations in hierarchical models, we reanalyzed two subsets of the complete data set of 150 observations given in Table 3, chosen to present an increasing challenge to the algorithms. One subset consisted of 90 observations, obtained by omitting the final data point from rats 6–10, the final two data points from rats 11–20, the final three from rats 21–25, and the final four from rats 26–30. The other subset consisted of 75 observations, obtained from the 90 by retaining only one of the observations for each of the rats 16–30. Convergence for the first subset required about 50 iterations of $m = 50$; convergence for the second required about 65 iterations of $m = 50$.

Figure 6 summarizes the marginal posteriors for the growth rate parameter obtained for the two data subsets from the Gibbs and EM-type algorithms, respectively. It can be seen that while the EM approximation is perhaps tolerable for the full data set (Fig. 5), it is very poor for the smaller data sets.

Consider now the data set of 90 observations and suppose that the problem of interest is the prediction of the future growth pattern for one of the rats for which there is currently just the first observation available ($i = 26, \ldots, 30$). Specifically, suppose we consider predicting Y_{ij}, $j = 2, 3, 4, 5$, corresponding to $x_{i2} = 15$, $x_{i3} = 22$, $x_{i4} = 29$, $x_{i5} = 36$ days. Then, formally,

$$[Y_{ij} \mid Y] = \int [Y_{ij} \mid \theta_i, \sigma_c^2] * [\theta_i, \sigma_c^2 \mid Y],$$

where

$$[Y_{ij} \mid \theta_i, \sigma_c^2] = N(\alpha_i + \beta_i x_{ij}, \sigma_c^2). \tag{6}$$

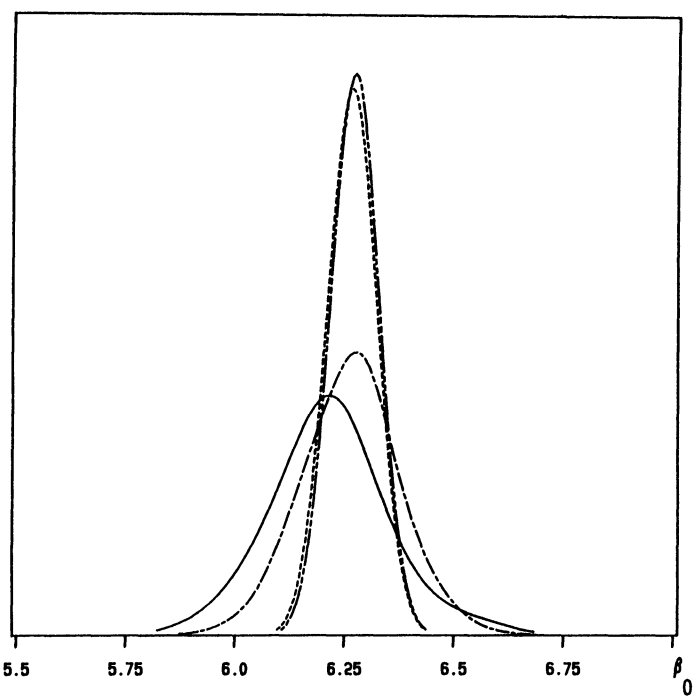

Figure 6. Estimated Densities for Population Growth Rate for 90 Observation Case (–·–, Gibbs sampler; ——, EM) and for 75 Observation Case (———, Gibbs sampler; ---, EM).

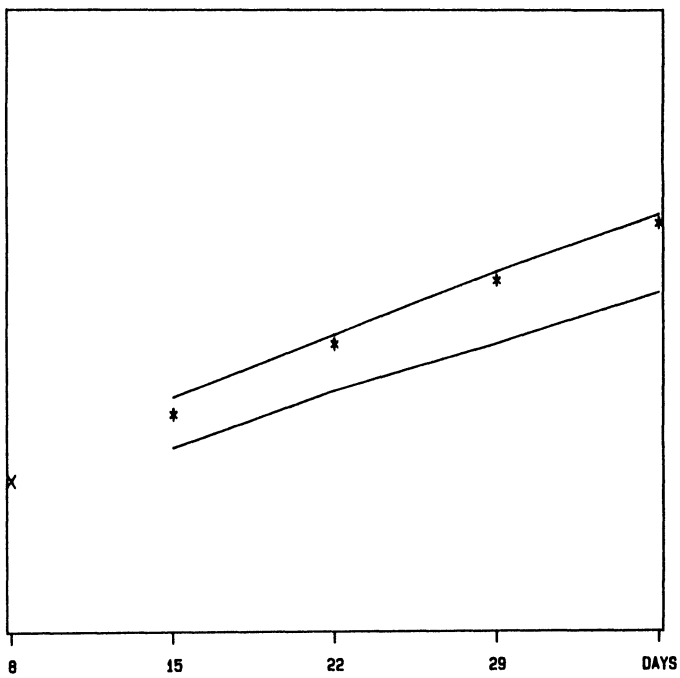

Figure 7. Estimated 95% Predictive Intervals for Future Observations (*) Given the First Observation (×) of Rat 26 (90 observation case).

An estimate of $[Y_{ij} \mid Y]$ of the form (1) is thus easily obtained by averaging $[Y_{ij} \mid \theta_i, \sigma_c^2]$ over pairs of (θ_i, σ_c^2) obtained at the final cycle of the Gibbs sampler. Figure 7 shows, for $i = 26$, bands drawn through the individual 95% predictive interval limits calculated at days 15, 22, 29, and 36, together with the subsequently observed values at those points. Alternatively, we could view the omitted or, in general, as yet unobserved data points as missing data. The Gibbs sampler could then be implemented treating such Y_{ij} as unobservable (in addition to the model parameters), since the required full conditional distributions have the form (6).

We have illustrated, using the control group data of Table 3, the simplicity and flexibility of the Gibbs sampler as a means of carrying out fully Bayesian inference and prediction for the normal linear hierarchical model. We turn now to the substantive applied problem, originally considered within the CIBA-GEIGY company, of comparing the control group with a treatment group, the data for which are given in Table 4.

The hierarchical model (together with hyperparameter prior specification) for the treatment group is assumed to have precisely the same form as that described above for the control group, except that, notationally, σ_c^2, α_c, β_c, μ_c, and Σ_c are replaced by σ_t^2, α_t, β_t, μ_t, Σ_t. In addition, prior assignments for the two groups are taken to be independent, so that inference within each group essentially proceeds separately.

The main parameters of practical interest are $\beta_t - \beta_c$, the difference in population growth rates, and $\sigma_{\beta t}^2/\sigma_{\beta c}^2$, the relative variation in growth rates in the two populations (given by the ratio of the second diagonal elements of Σ_t and Σ_c), and we focus our discussion on just these two parameters.

Table 4. Rat Population Growth Data: Treatment Group

Rat	x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}	Rat	x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}
1	114	151	188	214	253	16	129	170	209	251	289
2	140	189	229	258	303	17	122	154	191	226	254
3	133	176	220	252	282	18	134	178	212	247	276
4	132	168	208	234	270	19	131	169	206	242	278
5	119	155	186	207	233	20	119	155	190	225	264
6	155	190	220	243	262	21	140	178	220	249	285
7	117	146	188	214	245	22	136	167	204	232	271
8	129	160	199	228	261	23	148	184	225	250	284
9	148	192	227	247	278	24	145	178	219	252	290
10	144	183	220	245	278	25	146	175	212	233	266
11	117	145	177	210	245	26	118	154	178	203	226
12	156	208	242	278	319	27	147	190	221	259	286
13	108	137	164	189	221	28	143	182	210	240	264
14	140	178	210	235	265	29	148	197	235	279	309
15	139	180	221	256	289	30	136	177	223	256	287

NOTE: $x_{i1} = 8$, $x_{i2} = 15$, $x_{i3} = 22$, $x_{i4} = 29$, $x_{i5} = 36$ days; $i = 1, \ldots, 30$.

For both the control and treatment groups apparent convergence was achieved with less than 40 cycles of $m = 50$ drawings. In each case, several further cycles of $m = 100$ were run; first to check convergence, second to provide larger samples for the next stage. This consisted of forming the $10{,}000 = 100 \times 100$ pairs of sampled (β_c, β_t) and $(\sigma^2_{\beta c}, \sigma^2_{\beta t})$, from which either samples could be drawn, or the full 10,000 used, to form samples of $\beta_t - \beta_c$ and $\sigma^2_{\beta t}/\sigma^2_{\beta c}$, and thence to produce kernel density estimates of the required marginals. Once again, we note the utter simplicity of passing to whatever inference summary is required, following the convergence of the sampler. Figures 8 and 9 display the resulting marginal densities for $\beta_t - \beta_c$ and $\sigma^2_{\beta t}/\sigma^2_{\beta c}$. The clear messages are that the population growth rate is lower in the treatment group, and that the treatment group displays less variation around its population mean growth rate than does the control group.

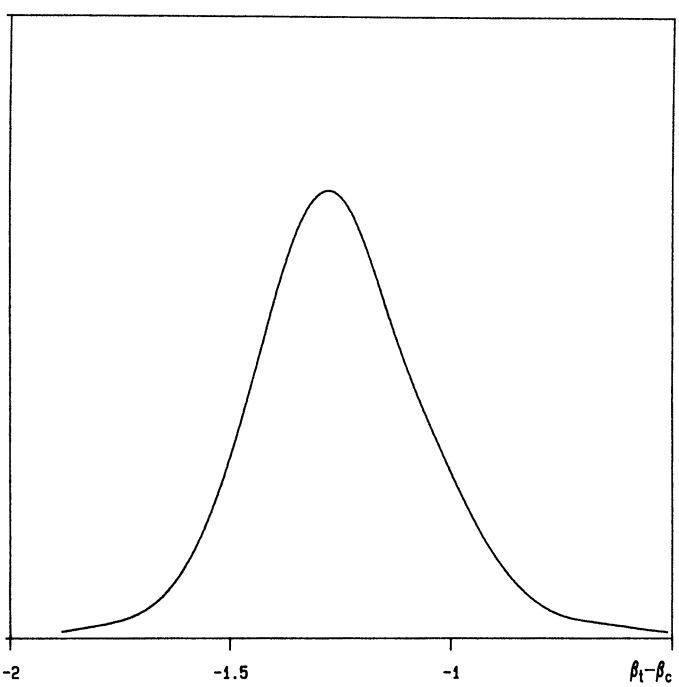

Figure 8. Estimated Density of Difference in Population Growth Rates.

We discussed earlier the inadequacy of other forms of approximate Bayes/empirical Bayes implementation of the normal linear hierarchical model. In particular, it is worth emphasizing again that, hitherto, no methods known to us have had the ability to calculate properly posterior uncertainty about population covariances (such as Σ_c, Σ_t), let alone deal with functions of parameters of such matrices (such as $\sigma^2_{\beta t}/\sigma^2_{\beta c}$).

7. MISSING DATA IN A CROSSOVER TRIAL

The balanced two-period crossover design is widely used; for example, in the pharmaceutical industry for bioequivalence studies involving a standard and a new drug formulation, A and B, say (Racine, Grieve, Flühler, and Smith 1986; Racine-Poon, Grieve, Flühler, and Smith 1987). Assuming n subjects, the standard random effects model for a two-period crossover is given by

$$Y_{i(jk)} = \mu + (-1)^{(j-1)}\left(\frac{\phi}{2}\right) + (-1)^{(k-1)}\left(\frac{\pi}{2}\right) + \delta_i + \varepsilon_{i(jk)},$$

where $Y_{i(jk)}$ = response to the ith subject ($i = 1, \ldots, n$) receiving the jth formulation ($j = 1, 2$) in the kth period ($k = 1, 2$); μ = overall mean level of response; ϕ = difference in formulation effects; π = difference in period effects; δ_i = random effect of ith subject; $\varepsilon_{i(jk)}$ = measurement error. The δ_i and the $\varepsilon_{i(jk)}$ are assumed independent for all i, j, k, with $\varepsilon_{i(jk)} \sim N(0, \sigma^2_1)$ and $\delta_i \sim N(0, \sigma^2_2)$. The essential problem of interest in the Bayesian approach to bioequivalence studies (see Racine et al. 1986) is to establish whether or not the parameter $\theta = \exp(\phi)$ lies in the interval (.8, 1.2) with high posterior probability. We note first the minor complication that the parameter of interest is a nonlinear function of a linear model parameter; and second, the more serious complication that crossover trials very often result in missing data for one or other of the periods, thereby spoiling the intended balanced design structure and subsequent standard form analysis. In this latter context, the question arises whether or not to omit subjects with missing data. On the one hand, omission may simplify the analysis; on the other hand, information is being discarded. It is therefore of consid-

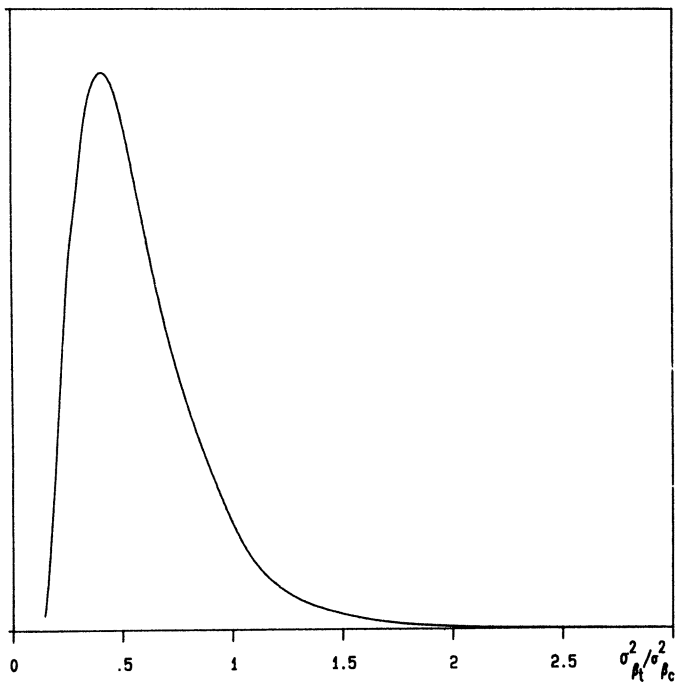

Figure 9. Estimated Density of the Ratio of Population Growth Rate Variances.

erable practical interest to be able to analyze numerically the missing data case, and to ascertain the relative loss of information in omitting subjects in order to revert to a more straightforward (analytically available) form of analysis. We illustrate the relative simplicity with which the Gibbs sampler approach achieves these goals.

To illustrate the general problem, suppose that subjects $i = 1, \ldots, M$ have data missing at random (Rubin 1976) from one of the two periods, and that subjects $i = M + 1, \ldots, n$ have complete data. We write

$$Y_i = \begin{pmatrix} U_i \\ V_i \end{pmatrix}, \qquad X_i = \begin{pmatrix} X_{iu} \\ X_{iv} \end{pmatrix}, \qquad i = 1, \ldots, M,$$

where the "observations" within subject i are labeled such that V_i is the observed data, U_i is missing, and X_i defines the corresponding design matrix. For subjects $i = M + 1, \ldots, n$, $Y_i = V_i$ simply denotes all the observed data. We write $U = (U_1, \ldots, U_M)^T$ and $V = (V_1, \ldots, V_n)^T$. Then conditional on

$$\psi = (\mu, \phi, \pi)^T, \quad \Sigma = \begin{bmatrix} \sigma^2_{12} & \sigma^2_2 \\ \sigma^2_2 & \sigma^2_{12} \end{bmatrix},$$

where $\sigma^2_{12} = \sigma^2_1 + \sigma^2_2$, we have

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \sim N\left\{ \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \psi, S \right\} \equiv N(X\theta, S),$$

where

$$S = \begin{bmatrix} \Sigma & & 0 \\ & \Sigma & \\ & & \ddots \\ 0 & & \Sigma \end{bmatrix}.$$

Here, Y is $2n \times 1$, X is $2n \times 2$, and S is $2n \times 2n$.

It is convenient to work with ψ, σ^2_1, σ^2_3, where $\sigma^2_3 = \sigma^2_1 + 2\sigma^2_2$, and to note that, if we define

$$Y^+_{i.} = \text{average response of the } i\text{th subject} = \begin{cases} \tfrac{1}{2}(Y_{i(11)} + Y_{i(22)}) \\ \tfrac{1}{2}(Y_{i(21)} + Y_{i(12)}) \end{cases} \text{for } \begin{matrix} AB \\ BA \end{matrix} \text{ sequence,}$$

$$Y^-_{i.} = \text{difference of the two responses for the } i\text{th subject} = \begin{cases} \tfrac{1}{2}(Y_{i(11)} - Y_{i(22)}) \\ \tfrac{1}{2}(Y_{i(21)} - Y_{i(12)}) \end{cases} \text{for } \begin{matrix} AB \\ BA \end{matrix} \text{ sequence,}$$

then for the AB sequence,

$$\begin{pmatrix} Y^-_{i.} \\ Y^+_{i.} \end{pmatrix} \sim N\left\{ \begin{pmatrix} \tfrac{1}{2}(\phi + \pi) \\ \mu \end{pmatrix}, \tfrac{1}{2} \begin{pmatrix} \sigma^2_1 & 0 \\ 0 & \sigma^2_3 \end{pmatrix} \right\};$$

and for the BA sequence,

$$\begin{pmatrix} Y^-_{i.} \\ Y^+_{i.} \end{pmatrix} \sim N\left\{ \begin{pmatrix} \tfrac{1}{2}(\pi - \phi) \\ \mu \end{pmatrix}, \tfrac{1}{2} \begin{pmatrix} \sigma^2_1 & 0 \\ 0 & \sigma^2_3 \end{pmatrix} \right\}.$$

If we now make the prior specification

$$[\psi, \sigma^2_1, \sigma^2_3] = N(\eta, C) IG\left(\frac{\nu_1}{2}, \frac{\nu_1\tau_1}{2}\right) \times IG\left(\frac{\nu_3}{2}, \frac{\nu_3\tau_3}{2}\right) I_{(\sigma^2_1 \le \sigma^2_3)},$$

it is apparent that even modulo normalization, the form of the joint posterior for σ^2_1, σ^2_3, θ, and U, is very messy indeed. However, continuing to treat U as an additional unknown vector, it can be seen that

$$[\sigma^2_1, [\sigma^2_3 \mid U, V, \psi] \propto (\sigma^2_1)^{-n2} \exp\left(-\frac{1}{2\sigma^2_1} SS_1\right) \times (\sigma^2_1)^{-(\nu_1/2+1)} \exp\left(-\frac{\nu_1\tau_1}{2\sigma^2_1}\right) \times (\sigma^2_3)^{-n/2} \exp\left(-\frac{1}{2\sigma^2_3} SS_3\right) \times (\sigma^2_3)^{-(\nu_3/2+1)} \exp\left(-\frac{\nu_3\tau_3}{2\sigma^2_3}\right) I_{(\sigma^2_1 \le \sigma^2_3)},$$

where

$$SS_1 = 2 \sum_{AB \text{ seq.}} \left[Y^-_{i.} - \left(\frac{\phi + \pi}{2}\right) \right]^2 + 2 \sum_{BA \text{ seq.}} \left[Y^-_{i.} - \left(\frac{\pi - \phi}{2}\right) \right]^2,$$

$$SS_3 = 2 \sum_{i=1}^{n} (Y^+_{i.} - \mu)^2,$$

from which it follows that a Gibbs sampler for σ_1^2, σ_3^2, ψ, and U is specified by

$$[\sigma_1^2 \mid U, V, \psi, \sigma_3^2] = \mathrm{IG}\left(\frac{n + \nu_1}{2}, \frac{SS_1 + \nu_1\tau_1}{2}\right) I_{(\sigma_1^2 \le \sigma_3^2)},$$

$$[\sigma_3^2 \mid U, V, \psi, \sigma_1^2] = \mathrm{IG}\left(\frac{n + \nu_3}{2}, \frac{SS_3 + \nu_3\tau_3}{2}\right) I_{(\sigma_1^2 \le \sigma_3^2)},$$

$$[\psi \mid U, V, \sigma_1^2, \sigma_3^2] = N\{D(X^TS^{-1}Y + C^{-1}\eta), D\},$$

where

$$X^TS^{-1}Y = \sum_{i=1}^{n} X_i^T\Sigma^{-1}Y_i,$$

$$X^TS^{-1}X = \sum_{i=1}^{n} X_i^T\Sigma^{-1}X_i,$$

$$D = \mathrm{X}^T\mathrm{S}^{-1}X + C^{-1},$$

$$[U \mid V, \psi, \sigma_1^2, \sigma_3^2] = N\left\{X_u\psi + \frac{\sigma_2^2}{\sigma_{12}^2}(V - X_w\psi), \sigma_{12}^2\left[1 - \left(\frac{\sigma_2^2}{\sigma_{12}^2}\right)^2\right] I_M\right\},$$

with

$$X_u = \begin{pmatrix} X_{1u} \\ \vdots \\ X_{Mu} \end{pmatrix}, \quad U = \begin{pmatrix} U_1 \\ \vdots \\ U_M \end{pmatrix},$$

$$X_w = \begin{pmatrix} X_{1v} \\ \vdots \\ X_{Mv} \end{pmatrix}, \quad W = \begin{pmatrix} V_1 \\ \vdots \\ V_M \end{pmatrix}.$$

Table 5 summarizes data from a (complete data) trial conducted with $n = 10$ subjects, in which, in order to illustrate the incomplete data problems referred to above, responses are treated as missing from subject 1 in period 1, subject 3 in period 2, and subject 6 in period 2.

Convergence was achieved within 30 iterations of $m = 50$. As we have remarked previously, given a sample of ϕ we can pass immediately to a sample of a function of ϕ; in this case, $\theta = \exp(\phi)$. Using a kernel density estimate based on a subsequent sample of $m = 100$, Figure 10 shows the marginal posterior densities for $\theta = \exp(\phi)$, the difference in the underlying mean concentration-time curve maxima, calculated from three different data sets, and illustrating the ease with which the Gibbs sampler permits analysis of this crossover model and enables informative sensitivity studies to be performed. The three posterior densities are based first on the complete data; second, on the data omitting the assumed missing values; and third, on the data for just the seven subjects for whom full data were assumed. The resulting posteriors reveal a typical finding in such trials: Namely, that if there is missing (completely at random) data from a subject we might just as well ignore the subject altogether; also, that the loss of 30% of subjects in a small trial results in substantially increased inferential uncertainty. However, none of the posteriors assigns high probability to the interval (.8, 1.2), and the substantive conclusion here is that the two formulations are not bioequivalent. The Gibbs sampler also automatically provides "predictive" densities for the missing responses; these are shown in Figure 11 and their locations may be compared with the actual missing values.

Table 5. Data From a Two-Period Crossover Trial

Subject	Sequence	Period 1	Period 2
1	AB	1.40	1.65
2	AB	1.64	1.57
3	BA	1.44	1.58
4	BA	1.36	1.68
5	BA	1.65	1.69
6	AB	1.08	1.31
7	AB	1.09	1.43
8	AB	1.25	1.44
9	BA	1.25	1.39
10	BA	1.30	1.52

NOTE: A = new tablet, B = standard tablet; formulations of Carbamazepine. The data are observations of the logarithms of maxima of concentration–time curves; see Maas, Garnett, Pollock, and Carnstock (1987) for background and further details.

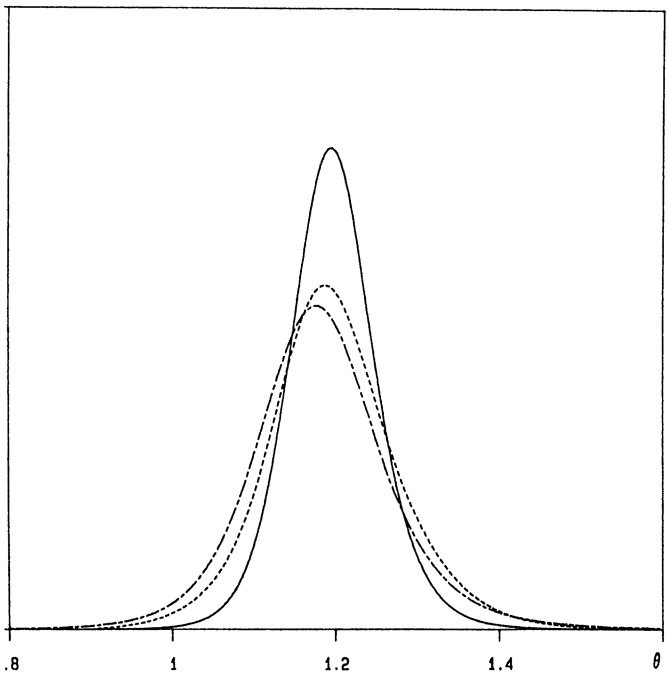

Figure 10. Estimated Densities of $\theta = \exp(\phi)$ in the Crossover Trial: ——, Complete Data; ---, Omitting Missing Values; –·–·, 7 Subjects.

We know of no "routinely" available alternative method for carrying out the kind of practical analysis we have outlined above. In the missing data case, there are eight unknown parameters: σ_1^2, σ_3^2, μ, π, U_1, U_2, and U_3. With this number of parameters, use of the adaptive numerical integration techniques as described in Smith et al. (1987) requires considerable familiarity with the procedures involved, as well as the need for subtle parameter transformation owing to the inequality constraints $0 \le \sigma_1^2 \le \sigma_3^2$. Given the complicated form of the likelihood, use of the Laplace approximation techniques of Tierney and Kadane seems unrealistic. Doubtless the skilled analyst could find a hybrid combination of profiling and integrating that would

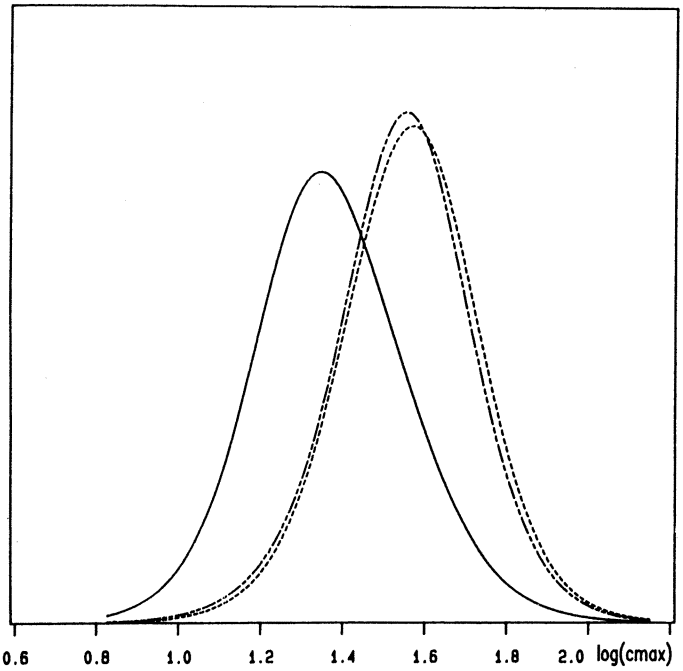

Figure 11. Estimated Predictive Densities for the Missing Data Values in the Crossover Trial: ——— = Subject 1, Actual Value = 1.40; --- = Subject 3, Actual Value = 1.44; —-— = Subject 6, Actual Value = 1.31.

avoid full numerical analysis in eight dimensions, but such an analysis is likely to be "one off" and, in any case, not routinely implementable by most applied statisticians. Once again, the virtue of the Gibbs sampler is its ease of implementation, despite the seeming complexity of the "missing data" likelihood.

8. SUMMARY DISCUSSION

The range of normal data problems considered above as illustrations of the ease with which numerical Bayesian inferences can be obtained via Gibbs sampling, include the following aspects:

- awkward posterior distributions, otherwise requiring subtle and sophisticated numerical or analytic approximation techniques (Secs. 4 and 5)
- further distributional complexity introduced by order constraints on model parameters (Secs. 5 and 7)
- dimensionality problems, typically putting out of reach the implementation of other sophisticated approximation techniques (Sec. 6)
- messy and intractable distribution theory arising from missing data in designed experiments (Sec. 7)
- general functions of model parameters (Secs. 4 and 6)
- awkward predictive inference (Sec. 6)

In all these situations, we saw that the Gibbs sampler approach is straightforward to specify distributionally, is trivial to implement computationally, and yields output readily translated into required inference summaries. In some of the examples we considered, alternative computational or approximation procedures are available, some of which can certainly prove more efficient than the Gibbs sampler in specific applications, although often requiring sophisticated numerical or analytic expertise. In several of the examples, however, there appears to be no currently available means of implementing a fully Bayesian analysis other than with the Gibbs sampler.

The potential of the methodology is enormous, rendering straightforward the analysis of a number of problems hitherto regarded as intractable from a Bayesian perspective. Work is in progress in extending the range of implementation: first, by developing, where necessary, purpose-built efficient random variate generators for conditional distribution forms arising in particular classes of applications; and second, by facilitating the reporting of bivariate and conditional inference summaries, in addition to univariate marginal curves.

[*Received February 1989. Revised March 1990.*]

REFERENCES

Achcar, J. A., and Smith, A. F. M. (1990), "Aspects of Reparametrization in Approximate Bayesian Inference," in *Essays in Honor of George A. Barnard,* ed. J. Hodges, Amsterdam: North-Holland.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972), *Statistical Inference Under Order Restrictions,* New York: John Wiley.

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

Besag, J. (1974), "Spatial Interaction and the Statistical Analysis of Lattice Systems" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 36, 192–326.

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

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

DuMouchel, W. H., and Harris, J. E. (1983), "Bayes Methods for Combining the Results of Cancer Studies in Humans and Other Species" (with discussion), *Journal of the American Statistical Association,* 78, 293–315.

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.

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, 721–741.

Geweke, J. (1988), "Antihetic Acceleration of Monte Carlo Integration in Bayesian Inference," *Journal of Econometrics,* 38, 73–90.

Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chain and Their Applications," *Biometrika,* 87, 97–109.

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

Lindley, D. V., and Smith, A. F. M. (1972), "Bayes Estimates for the Linear Model" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 34, 1–41.

Maas, B., Garnett, W. R., Pollock, I. M., and Carnstock, T. J. (1987), "A Comparative Bioavailability Study of Carbamazepine Tablets and the Chewable Formulation," *Therapeutic Drug Monitoring,* 9, 28–33.

Morris, C. (1983), "Parametric Empirical Bayes Inference: Theory and Applications," *Journal of the American Statistical Association,* 78, 47–59.

Naylor, J. C., and Smith, A. F. M. (1982), "Applications of a Method for the Efficient Computation of Posterior Distributions," *Applied Statistics,* 31, 214–225.

——— (1988), "Econometric Illustrations of Novel Numerical Integration Strategies for Bayesian Inference," *Journal of Econometrics,* 38, 103–126.

Odell, P. L., and Feiveson, A. H. (1966), "A Numerical Procedure to Generate a Sample Covariance Matrix," *Journal of the American Statistical Association,* 61, 198–203.

Racine, A., Grieve, A. P., Flühler, H., and Smith, A. F. M. (1986), "Bayesian Methods in Practice: Experiences in the Pharmaceutical Industry" (with discussion), *Applied Statistics,* 35, 93–150.

Racine-Poon, A. (1985), "A Bayesian Approach to Non-linear Random Effects Models," *Biometrics,* 41, 1015–1024.

Racine-Poon, A., Grieve, A. P., Flühler, H., and Smith, A. F. M. (1987), "A Two-Stage Procedure for Bioequivalence Studies," *Biometrics,* 43, 842–856.

Racine-Poon, A., and Smith, A. F. M. (1990), "Population Models," in *Statistical Methodology in the Pharmaceutical Sciences,* ed. D. Berry, New York: Marcel Dekker.

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988), *Order Restricted Statistical Inference,* New York: John Wiley.

Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika,* 63, 581–592.

——— (1981), "Estimation in Parallel Randomized Experiments," *Journal of Educational Statistics,* 6, 377–401.

Shaw, J. E. H. (1988), "A Quasirandom Approach to Integration in Bayesian Statistics," *The Annals of Statistics,* 16, 895–914.

Silverman, B. W., (1986), *Density Estimation for Statistics and Data Analysis,* London: Chapman & Hall.

Skene, A. M. (1983), "Computing Marginal Distributions for the Dispersion Parameters of Analysis of Variance Models," *The Statistician,* 32, 99–108.

Smith, A. F. M., Skene, A. M., Shaw, J. E. H., Naylor, J. C., and Dransfield, M. (1985), "The Implementation of the Bayesian Paradigm," *Communications in Statistics, Part A—Theory and Methods,* 14, 1079–1102.

Smith, A. F. M., Skene, A. M., Shaw, J. E. H., and Naylor, J. C. (1987), "Progress With Numerical and Graphical Methods for Bayesian Statistics," *The Statistician,* 36, 75–82.

Strenio, J. F., Weisberg, H. I., and Bryk, A. S. (1983), "Empirical Bayes Estimation of Individual Growth-Curve Parameters and Their Relationship to Covariates," *Biometrics,* 39, 71–86.

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

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