Sampling-Based Approaches to Calculating Marginal Densities

Alan E. Gelfand; Adrian F. M. Smith

Journal of the American Statistical Association, Vol. 85, No. 410. (Jun., 1990), pp. 398-409.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28199006%2985%3A410%3C398%3ASATCMD%3E2.0.CO%3B2-3

Journal of the American Statistical Association is currently published by American Statistical Association.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/astata.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

The JSTOR Archive is a trusted digital repository providing for long-term preservation and access to leading academic journals and scholarly literature from around the world. The Archive is supported by libraries, scholarly societies, publishers, and foundations. It is an initiative of JSTOR, a not-for-profit organization with a mission to help the scholarly community take advantage of advances in technology. For more information regarding JSTOR, please contact support@jstor.org.

http://www.jstor.org
Sat Nov 17 07:08:33 2007

Sampling-Based Approaches to Calculating Marginal Densities

ALAN E. GELFAND AND ADRIAN F. M. SMITH*

Stochastic substitution, the Gibbs sampler, and the sampling–importance–resampling algorithm can be viewed as three alternative sampling- (or Monte Carlo-) based approaches to the calculation of numerical estimates of marginal probability distributions. The three approaches will be reviewed, compared, and contrasted in relation to various joint probability structures frequently encountered in applications. In particular, the relevance of the approaches to calculating Bayesian posterior densities for a variety of structured models will be discussed and illustrated.

KEY WORDS: Conditional probability structure; Data augmentation; Gibbs sampler; Hierarchical models; Importance sampling; Missing data; Monte Carlo sampling; Posterior distributions; Stochastic substitution; Variance components.

1. INTRODUCTION

In relation to a collection of random variables, $U_1, U_2, \ldots, U_k$, suppose that either (a) for $i = 1, \ldots, k$, the conditional distributions $U_i \mid U_j\ (j \neq i)$ are available, perhaps having for some i reduced forms $U_i \mid U_j\ (j \in S_i \subset \{1, \ldots, k\})$, or (b) the functional form of the joint density of $U_1, U_2, \ldots, U_k$ is known, perhaps modulo the normalizing constant, and at least one $U_i \mid U_j\ (j \neq i)$ is available, where *available* means that samples of U_i can be straightforwardly and efficiently generated, given specified values of the appropriate conditioning variables.

The problem addressed in this article is the exploitation of the kind of structural information given by either (a) or (b), to obtain numerical estimates of nonanalytically available marginal densities of some or all of the U_i (when possible) simply by means of simulated samples from available conditional distributions, and without recourse to sophisticated numerical analytic methods. We do not claim that the sampling methods to be described are necessarily computationally efficient compared with expert use of the latter. Instead, the attraction of the sampling-based methods is their conceptual simplicity and ease of implementation for users with available computing resources but without numerical analytic expertise. All that the user requires is insight into the relevant conditional probability structure and techniques for the efficient generation of appropriate random variates (e.g., as described by Devroye 1986 and Ripley 1987).

In Section 2, we discuss and extend three alternative approaches put forward in the literature for calculating marginal densities via sampling algorithms. These are (variants of) the data-augmentation algorithm described by Tanner and Wong (1987), the Gibbs sampler algorithm introduced by Geman and Geman (1984), and the form of importance-sampling algorithm proposed by Rubin (1987, 1988). We note that the Gibbs sampler has been widely taken up in the image-processing literature and in other large-scale models such as neural networks and expert systems, but that its general potential for more conventional statistical problems seems to have been overlooked. As we show, there is a close relationship between the Gibbs sampler and the substitution or data-augmentation algorithm proposed by Tanner and Wong (1987). We generalize the latter and show that it is as least as efficient as the Gibbs sampler, and potentially more efficient, given the availability of distinct conditional distributions in addition to those in (a). We note that as a consequence of the relationship between the two algorithms, the convergence results established by Geman and Geman (1984) are applicable to the generalized substitution algorithm. The stronger convergence results established by Tanner and Wong (1987) require the availability of a particular set of conditional distributions, including those in (a). Both the substitution and Gibbs sampler algorithms are iterative Monte Carlo procedures, applicable when the kind of structural information given by (a) is available. When the structural information is of the kind described by (b), we see that an importance-sampling algorithm based on that of Rubin (1987, 1988) provides a noniterative Monte Carlo integration approach to calculating marginal densities.

In Section 3, we illustrate various model structures occurring frequently in applications, where one or more of these three approaches offers an easily implemented solution. In particular, we consider the calculation of Bayesian posterior distributions in incomplete-data problems, conjugate hierarchical models, and normal data models. In Section 4, we briefly summarize the results of some preliminary computational experience in two simple cases. (Detailed applications to complex, real-data problems will be presented in a subsequent paper.) Finally, in Section 5 we provide a summary discussion.

* Alan E. Gelfand is Professor, Department of Statistics, University of Connecticut, Storrs, CT 06269-3120. Adrian F. M. Smith is Professor, Department of Mathematics, Imperial College of Science, Technology, and Medicine, University of London, London SW7 2AZ, England. This work was carried out under the auspices of the U.K. Science and Engineering Research Council Complex Stochastic Systems Initiative, which provided travel support to the first author and supported Susan Hills, to whom the authors are indebted for carrying out extensive computer experimentation. The second author's collaboration with Amy Racine provided the initial stimulation for this work, and the authors have benefited from access to related independent work by David Clayton.

© 1990 American Statistical Association
Journal of the American Statistical Association
June 1990, Vol. 85, No. 410, Theory and Methods

2. SAMPLING APPROACHES

In the sequel, we assume that we are dealing with real, possibly vector-valued random variables having a joint distribution whose density function is strictly positive over the (product) sample space. This ensures that knowledge of all full conditional specifications [such as in (a) of Sec. 1] uniquely defines the full joint density (e.g., see Besag 1974). Throughout, we assume the existence of densities with respect to either Lebesgue or counting measure, as appropriate, for all marginal and conditional distributions. The terms *distribution* and *density* are therefore used interchangeably.

Densities are denoted generically by brackets, so joint, conditional, and marginal forms, for example, appear as $[X, Y]$, $[X \mid Y]$, and $[X]$. Multiplication of densities is denoted by $*$; for example, $[X, Y] = [X \mid Y] * [Y]$. The process of marginalization (i.e., integration) is denoted by forms such as $[X \mid Y] = \int [X \mid Y, Z, W] * [Z \mid W, Y] * [W \mid Y]$, with the convention that all variables appearing in the integrand but not in the resulting density have been integrated out. Thus the integration is with respect to Z and W. More generally, we use notation such as $\int h(Z, W) * [W]$ to denote, for given Z, the expectation of the function $h(Z, W)$ with respect to the marginal distribution for W.

2.1 Substitution or Data-Augmentation Algorithm

The substitution algorithm for finding fixed-point solutions to certain classes of integral equations is a standard mathematical tool that has received considerable attention in the literature (e.g., see Rall 1969). Its potential utility in statistical problems of the kind we are concerned with was observed by Tanner and Wong (1987) (who called it a data-augmentation algorithm) and the associated discussion. Briefly reviewing the essence of their development using the notation introduced previously, we have

$$[X] = \int [X \mid Y] * [Y] \tag{1}$$

and

$$[Y] = \int [Y \mid X] * [X], \tag{2}$$

so substituting (2) into (1) gives

$$\begin{aligned}[X] &= \int [X \mid Y] * \int [Y \mid X'] * [X'] \\ &= \int h(X, X') * [X'], \end{aligned} \tag{3}$$

where $h(X, X') = \int [X \mid Y] * [Y \mid X']$, with X' appearing as a dummy argument in (3), and of course $[X] \equiv [X']$. Now, suppose that on the right side of (3), $[X']$ were replaced by $[X]_i$, to be thought of as an estimate of $[X] \equiv [X']$ arising at the ith stage of an iterative process. Then, (3) implies that for some $[X]_{i+1}$, $[X]_{i+1} = \int h(X, X') * [X']_i = I_h[X]_i$, in a notation making explicit that I_h is the integral operator associated with h. Exploiting standard theory of such integral operators, Tanner and Wong (1987) showed that under mild regularity conditions this iterative process has the following properties (with obviously analogous results for $[Y]$).

TW1 (*uniqueness*). The true marginal density, $[X]$, is the unique solution to (3).

TW2 (*convergence*). For almost any $[X]_0$, the sequence $[X]_1, [X]_2, \ldots$ defined by $[X]_{i+1} = I_h[X]_i$ ($i = 0, 1, \ldots$) converges monotonically in L_1 to $[X]$.

TW3 (*rate*). $\int |[X]_i - [X]| \to 0$ geometrically in i.

Extending the substitution algorithm to three random variables X, Y, and Z, we may write [analogous to (1) and (2)]

$$[X] = \int [X, Z \mid Y] * [Y], \tag{4}$$

$$[Y] = \int [Y, X \mid Z] * [Z], \tag{5}$$

and

$$[Z] = \int [Z, Y \mid X] * [X]. \tag{6}$$

Substitution of (6) into (5) and then (5) into (4) produces a fixed-point equation analogous to (3). A new h function arises with associated integral operator I_h, and hence TW1, TW2, and TW3 continue to hold in this extended setting. Extension to k variables is straightforward. A noteworthy by-product, using TW1, is a simple proof that under weak conditions specification of the conditional distributions $[U_{r, r \neq s} \mid U_s]$ ($s = 1, 2, \ldots, k$) uniquely determines the joint density.

2.2 Substitution Sampling

Returning to (1) and (2), suppose that $[X \mid Y]$ and $[Y \mid X]$ are available in the sense defined at the beginning of Section 1. For an arbitrary (possibly degenerate) initial density $[X]_0$ draw a single $X^{(0)}$ from $[X]_0$. Given $X^{(0)}$, since $[Y \mid X]$ is available draw $Y^{(1)} \sim [Y \mid X^{(0)}]$, and hence from (2) the marginal distribution of $Y^{(1)}$ is $[Y]_1 = \int [Y \mid X] * [X]_0$. Now, complete a cycle by drawing $X^{(1)} \sim [X \mid Y^{(1)}]$. Using (1), we then have $X^{(1)} \sim [X]_1 = \int [X \mid Y] * [Y]_1 = \int h(X, X') * [X']_0 = I_h[X]_0$. Repetition of this cycle produces $Y^{(2)}$ and $X^{(2)}$, and eventually, after i iterations, the pair $(X^{(i)}, Y^{(i)})$ such that $X^{(i)} \xrightarrow{d} X \sim [X]$, and $Y^{(i)} \xrightarrow{d} Y \sim [Y]$, by virtue of TW2. Repetition of this sequence m times each to the ith iteration generates m iid pairs $(X_j^{(i)}, Y_j^{(i)})$ ($j = 1, \ldots, m$). We call this generation scheme *substitution sampling*. Note that though we have independence across j, we have dependence within a given j.

If we terminate all repetitions at the ith iteration, the proposed density estimate of $[X]$ (with an analogous

expression for $[Y]$) is the Monte Carlo integration

$$[\hat{X}]_i = \frac{1}{m}\sum_{j=1}^{m}[X \mid Y_j^{(i)}]. \tag{7}$$

Note that the $X_j^{(i)}$ are not used in (7) (see Sec. 2.6).

We note that this version of the substitution-sampling algorithm differs slightly from the imputation–posterior algorithm of Tanner and Wong (1987). At each iteration l ($l = 1, 2, \ldots, i$), they proposed creation of the mixture density estimate, $[\hat{X}]_l$, of the form in (7), with subsequent sampling from $[\hat{X}]_l$ to begin the next iteration. This mechanism introduces the additional randomness of equally likely selection from the $Y_j^{(l)}$ before obtaining an $X^{(l)}$. We suspect this sampling with replacement of the $Y^{(l)}$ was introduced to allow m to vary across iterations, which may be useful in reducing computational effort.

The L_1 convergence of $[\hat{X}]_i$ to $[X]$ is most easily studied by writing $\int |[\hat{X}]_i - [X]| \leq \int |[\hat{X}]_i - [X]_i| + \int |[X]_i - [X]|$. The second term on the right side can be made arbitrarily small as $i \to \infty$, as a consequence of TW2. The first term on the right can be made arbitrarily small as $m \to \infty$, since $[\hat{X}]_i \xrightarrow{P} [X]_i$ for almost all X (Glick 1974).

Extension of the substitution-sampling algorithm to more than two random variables is straightforward. We illustrate using the three-variable case, assuming the three conditional distributions in (4)–(6) are available. Taking an arbitrary starting marginal density for X, say $[X]_0$, we draw $X^{(0)} \sim [X]_0$, $(Z^{(0)\prime}, Y^{(0)\prime}) \sim [Z, Y \mid X^{(0)}]$, $(Y^{(1)}, X^{(0)\prime}) \sim [Y, X \mid Z^{(0)\prime}]$, and finally $(X^{(1)}, Z^{(1)}) \sim [X, Z \mid Y^{(1)}]$. A full cycle of the algorithm (i.e., to generate $X^{(1)}$ starting from $X^{(0)}$) thus requires six generated variates, rather than the two we saw earlier. Repeating such a cycle i times produces $(X^{(i)}, Y^{(i)}, Z^{(i)})$. The aforementioned theory ensures that $X^{(i)} \xrightarrow{d} X \sim [X]$, $Y^{(i)} \xrightarrow{d} Y \sim [Y]$, and $Z^{(i)} \xrightarrow{d} Z \sim [Z]$. If we repeat the entire process m times we obtain iid $(X_j^{(i)}, Y_j^{(i)}, Z_j^{(i)})$ ($j = 1, \ldots, m$) (independent between, but not within, j's). Note that implementation of the substitution-sampling algorithm does not require specification of the full joint distribution. Rather, what is needed is the availability of $[X, Z \mid Y]$, $[Y, X \mid Z]$, and $[Z, Y \mid Z]$. Of course, in many cases sampling from, say, $[X, Z \mid Y]$ requires, for example, $[X \mid Y, Z]$ and $[Y \mid Z]$, that is, the availability of a full conditional and a reduced conditional distribution. Paralleling (7), the density estimator of $[X]$ becomes

$$[\hat{X}]_i = \frac{1}{m}\sum_{j=1}^{m}[X \mid Y_j^{(i)}, Z_j^{(i)}], \tag{8}$$

with analogous expressions for estimating $[Y]$ and $[Z]$. L_1 convergence of (8) to $[X]$ again follows.

For k variables, $U_1, \ldots, U_k$, the substitution-sampling algorithm requires $k(k-1)$ random variate generations to complete a cycle. If we run m sequences out to the ith iteration [$mik(k-1)$ random generations] we obtain m iid k tuples $(U_{1j}^{(i)}, \ldots, U_{kj}^{(i)})$ ($j = 1, \ldots, m$), with the density estimator for $[U_s]$ ($s = 1, \ldots, k$) being

$$[\hat{U}_s]_i = \frac{1}{m}\sum_{j=1}^{m}[U_s \mid U_t = U_{tj}^{(i)}; t \neq s]. \tag{9}$$

2.3 Gibbs Sampling

Suppose that we write (4)–(6) in the form

$$[X] = \int [X \mid Z, Y] * [Z \mid Y] * [Y]$$

$$[Y] = \int [Y \mid X, Z] * [X \mid Z] * [Z]$$

$$[Z] = \int [Z \mid Y, X] * [Y \mid X] * [X]. \tag{10}$$

Implementation of substitution sampling requires the availability of all six conditional distributions on the right side of (10), rarely the case in our applications. As noted at the beginning of Section 2, the full conditional distributions alone, $[X \mid Y, Z]$, $[Y \mid Z, X]$, and $[Z \mid X, Y]$, uniquely determine the joint distribution (and hence the marginal distributions) in the situations under study. An algorithm for extracting the marginal distributions from these full conditional distributions was formally introduced by Geman and Geman (1984) and is known as the Gibbs sampler. An earlier article by Hastings (1970) developed essentially the same idea and suggested its potential for numerical problems arising in statistics.

The Gibbs sampler was developed and has mainly been applied in the context of complex stochastic models involving very large numbers of variables, such as image reconstruction, neural networks, and expert systems. In these cases, direct specification of a joint distribution is typically not feasible. Instead, the set of full conditionals is specified, usually by assuming that an individual full conditional distribution only depends on some "neighborhood" subset of the variables [a reduced form, in the terminology of (a) in Sec. 1]. More precisely, for the set of variables $U_1, U_2, \ldots, U_k$,

$$[U_i \mid U_j; \neq i] \equiv [U_i \mid U_j; j \in S_i], \qquad i = 1, \ldots, k, \tag{11}$$

where S_i is a small neighborhood subset of $\{1, 2, \ldots, k\}$. A crucial question is under what circumstances the specification (11) uniquely determines the joint distribution. The answer is taken up in great detail by Geman and Geman (1984), involving concepts such as graphs, neighborhood systems, cliques, Markov random fields, and Gibbs distributions. In all of the examples we consider, the joint distribution is uniquely defined. Our k's will be small to moderate, and the available set of full conditional distributions will, in fact, be calculated from specification of the joint density.

Gibbs sampling is a Markovian updating scheme that proceeds as follows. Given an arbitrary starting set of values $U_1^{(0)}, U_2^{(0)}, \ldots, U_k^{(0)}$, we draw $U_1^{(1)} \sim [U_1 \mid U_2^{(0)}, \ldots, U_k^{(0)}]$, $U_2^{(1)} \sim [U_2 \mid U_1^{(1)}, U_3^{(0)}, \ldots, U_k^{(0)}]$, $U_3^{(1)} \sim [U_3 \mid U_1^{(1)} U_2^{(1)}, U_4^{(0)}, \ldots, U_k^{(0)}]$, and so on, up to $U_k^{(1)} \sim [U_k \mid U_1^{(1)}, \ldots, U_{k-1}^{(1)}]$. Thus each variable is visited in the natural order and a cycle in this scheme requires K random variate generations. After i such iterations we would arrive at $(U_1^{(i)}, \ldots, U_k^{(i)})$. Under mild conditions, Geman and Geman showed that the following results hold.

GG1 (*convergence*). $(U_1^{(i)}, \ldots, U_k^{(i)}) \xrightarrow{d} [U_1, \ldots, U_k]$ and hence for each s, $U_s^{(i)} \xrightarrow{d} U_s \sim [U_s]$ as $i \to \infty$. In fact, a slightly stronger result is proven. Rather than requiring that each variable be visited in repetitions of the natural order, convergence still follows under any visiting scheme, provided that each variable is visited infinitely often (io).

GG2 (*rate*). Using the sup norm, rather than the L_1 norm, the joint density of $(U_1^{(i)}, \ldots, U_k^{(i)})$ converges to the true joint density at a geometric rate in i, under visiting in the natural order. A minor adjustment to the rate is required for an arbitrary io visiting scheme.

GGS (*ergodic theorem*). For any measurable function T of $U_1, \ldots, U_k$ whose expectation exists,

$$\lim_{i\to\infty} \frac{1}{i} \sum_{l=1}^{i} T(U_1^{(l)}, \ldots, U_k^{(l)}) \xrightarrow{\text{a.s.}} E(T(U_1, \ldots, U_k)).$$

As in Section 2.3, Gibbs sampling through m replications of the aforementioned i iterations (mik random variate generations) produces m iid k tuples $(U_{1j}^{(i)}, \ldots, U_{kj}^{(i)})$ $(j = 1, \ldots, m)$, with the proposed density estimate for $[U_s]$ having form (9).

2.4 Relationship Between Gibbs Sampling and Substitution Sampling

It is apparent that in the case of two random variables Gibbs sampling and substitution sampling are identical. For more than two variables, using (10) and its obvious generalization to k variables, we see that Gibbs sampling assumes the availability of the set of k full conditional distributions (the minimal set needed to determine the joint density uniquely). The substitution-sampling algorithm requires the availability of $k(k - 1)$ conditional distributions, including all of the full conditionals.

Gibbs sampling is known to converge slowly in applications with k very large. Regardless, fair comparison with substitution sampling, in the sense of the total amount of random variate generation, requires that we allow the Gibbs sampling algorithm $i(k - 1)$ iterations if the substitution-sampling algorithm is allowed i. Even so, there is clearly scope for accelerated convergence from the substitution-sampling algorithm, since it samples from the correct distribution each time, whereas Gibbs sampling only samples from the full conditional distributions. To amplify, we describe how the substitution-sampling algorithm might be carried out under availability of just the set of full conditional distributions. We see that it can be viewed as the Gibbs sampler, but under an io visiting scheme different from the natural one. We present the argument in the three-variable case for simplicity. Returning to (10), if $[Y \mid X]$ is unavailable we can create a sub-substitution loop to obtain it by means of

$$[Y \mid X] = \int [Y \mid X, Z] * [Z \mid X]$$

$$[Z \mid X] = \int [Z \mid X, Y] * [Y \mid X]. \qquad (12)$$

Similar subloops are clearly available to create $[X \mid Z]$ and $[Z \mid Y]$. In fact, for k variables this idea can be straightforwardly extended to the estimation of an arbitrary reduced conditional distribution, given the full conditionals. We omit the details.

The previous analysis suggests that we could view the reduced conditional densities such as $[Y \mid X]$ as available, and that we could thus carry out the substitution algorithm as if all needed conditional distributions were available; however, $[Y|X]$ is not available in our earlier sense. Under the subloop in (12), we can always obtain a density estimate for $[Y \mid X]$, given any specified X, say $X^{(0)}$. At the next cycle of the iteration, however, we would need a brand-new density estimate for $[Y \mid X]$ at $X = X^{(1)}$. Nonetheless, suppose we persevered in this manner, making our way through one cycle of (10). The reader may verify that the only distributions actually sampled from are, of course, the available full conditionals, that at the end of the cycle each full conditional will have been sampled from at least once, and thus that under repeated iterations each variable will be visited io. Therefore, this version of the substitution-sampling algorithm is merely Gibbs sampling with a different but still io visiting order. As a result, GG1, GG2, and GG3 still hold (TW1, TW2, and TW3 apply directly only when all required conditional distributions are available). Moreover, there is no gain in implementing the Gibbs sampler in this complicated order; the natural order is simpler and equally good.

This discussion may be readily extended to the case of k variables. As a result, we conclude that when only the set of k full conditionals is available the substitution-sampling algorithm and the Gibbs sampler are equivalent. Furthermore, we can now see when substitution sampling offers the possibility of acceleration relative to Gibbs sampling. This occurs when some reduced conditional distributions, distinct from the full conditional distributions, are available. Suppose that we write the substitution algorithm with appropriate conditioning to capture these available reduced conditionals. As we traverse a cycle, we would sample from these distributions as we come to them, otherwise sampling from the full conditional distributions.

An example will help clarify this idea. One way to carry out the Gibbs sampler in (10) is to follow the substitution order rather than the natural order. That is, given an initial $X^{(0)}$, $Y^{(0)}$, and $Z^{(0)}$, we start at the bottom line of (10), for example, drawing (a) $Y^{(0)'}$ from $[Y \mid X^{(0)}, Z^{(0)}]$, (b) $Z^{(0)'}$ from $[Z \mid Y^{(0)'}, X^{(0)}]$, (c) $X^{(0)'}$ from $[X \mid Z^{(0)'}, Y^{(0)'}]$, (d) $Y^{(1)}$ from $[Y \mid X^{(0)'}, Z^{(0)'}]$, (e) $Z^{(1)}$ from $[Z \mid Y^{(1)}, X^{(0)'}]$, and (f) $X^{(1)}$ from $[Z \mid Y^{(1)}, Z^{(1)}]$. Thus, in this case, one cycle using the substitution order corresponds to two cycles using the natural order. Suppose, however, that in addition to the full conditional distributions, $[Z \mid Y]$, say, is available and distinct from $[Z \mid X, Y]$. Following the substitution order, at step (e) we would instead draw $Z^{(1)}$ from the correct distribution, $[Z \mid Y^{(1)}]$.

In Section 3, we provide classes of examples where distinct reduced conditional distributions are available and classes where they generally are not. Our computational experience shows that the acceleration in convergence that arises from having available distributions in addition to the full conditionals is inconsequential (see Sec. 4).

2.5 The Rubin Importance-Sampling Algorithm

Rubin (1987) suggested a noniterative Monte Carlo method for generating marginal distributions using importance-sampling ideas. We first present the basic idea in the two-variable case. Suppose that we seek the marginal distribution of X, given only the functional form (modulo the normalizing constant) of the joint density $[X, Y]$ and the availability of the conditional distribution $[X \mid Y]$ [a special case of the conditions described in (b) of Sec. 1].

Suppose further (as is typically the case in applications) that the marginal distribution of Y is not known. Choose an importance-sampling distribution for Y that has positive support wherever $[Y]$ does and that has density $[Y]_s$, say. Then, $[X \mid Y] * [Y]_s$ provides an importance-sampling distribution for (X, Y). Suppose that we draw iid pairs (X_l, Y_l) $(l = 1, \ldots, N)$ from this joint distribution, for example, by drawing Y_l from $[Y]_s$ and X_l from $[X \mid Y_l]$. Rubin's idea is to calculate $r_l = [X_l, Y_l]/[X_l \mid Y_l] * [Y_l]_s$ $(l = 1, \ldots, N)$ and then estimate the marginal density for $[X]$ by

$$[\hat{X}] = \sum_{l=1}^{N} [X \mid Y_l] r_l \Big/ \sum_{i=1}^{N} r_l. \tag{13}$$

Note the important fact that $[X, Y]$ need only be specified up to a constant, since the latter cancels in (13). In other words, we do not need to evaluate the normalizing constant for $[X, Y]$. This feature is exploited in the examples of Section 3. By dividing the numerator and denominator of (13) by N and using the law of large numbers, we immediately have the following.

R1 (convergence). $[\hat{X}] \to [X]$ with probability 1 as $N \to \infty$ for almost every X.

In addition, if $[Y \mid X]$ is available we immediately have an estimate for the marginal distribution of Y: $[\hat{Y}] = \sum_{l=1}^{N} [Y \mid X_l] r_l / \sum_{l=1}^{N} r_l$.

The successful performance of (13) typically depends strongly on the choice of $[Y]_s$ and its closeness to $[Y]$. Thus the suggestion of Tanner and Wong (1987) in their rejoinder to Rubin, to perhaps use for $[Y]_s$ the density estimate created after i iterations of the substitution algorithm, merits further investigation. In fact, the whole problem of general strategies for synthesizing both the iterative and noniterative approaches under a fixed-budget (total number of random generations) criterion needs considerable further study.

The extension of the Rubin importance-sampling idea to the case of k variables is clear. For instance, when $k = 3$, suppose that we seek the marginal distribution of X, given the functional form of $[X, Y, Z]$ up to a constant and the availability of the full conditional $[X \mid Y, Z]$. In this case, the pair (Y, Z) plays the role of Y in the two-variable case discussed before, and in general we need to specify an importance-sampling distribution $[Y, Z]_s$. Nevertheless, if $[Y \mid Z]$ is available, for example, we only need to specify $[Z]_s$. In any case, we draw iid triples (X_l, Y_l, Z_l) $(l = 1, \ldots, N)$ and calculate $r_l = [X_l, Y_l, Z_l]/([X_l \mid Y_l, Z_l] * [Y_l, Z_l]_s)$. The marginal density estimate for $[X]$ then becomes [analogous to (13)]

$$[\hat{X}] = \sum_{l=1}^{N} [X \mid Y_l, Z_l] r_l \Big/ \sum_{l=1}^{N} r_l. \tag{14}$$

We note that in the k-variable case the Rubin importance-sampling algorithm requires Nk random variate generations, whereas Gibbs sampling stopped at iteration i requires mik generations. For fair comparison of the two algorithms, we should therefore set $N = mi$. The relationship between the estimators (7) and (13) may be clarified if we resample $Y_1^*, Y_2^*, \ldots, Y_m^*$ from the distribution that places mass $r_l/\Sigma r_l$ at Y_l $(l = 1, \ldots, N)$. We could then replace (13) with

$$[\hat{X}] = \frac{1}{m} \sum_{j=1}^{m} [X \mid Y_j^*], \tag{15}$$

so (7) and (15) are of the same form. Relative performance on average depends on whether the distribution of $Y^{(i)}$ or Y^* is closer to $[Y]$. Empirical work described in Section 4 suggests that under fair comparison (7) performs better than (14) or (15). It seems preferable to iterate through a learning process with small samples rather than to draw a one-off large sample at the beginning [an idea that underlies much modern work in adaptive Monte Carlo; e.g., see Smith, Skene, Shaw, and Naylor (1987)].

2.6 Density Estimation

In this section, we consider the problem of calculating a final form of marginal density from the final sample produced by either the substitution- or Gibbs sampling algorithms. Since for any estimated marginal the corresponding full conditional has been assumed available, efficient inference about the marginal should clearly be based on using this full conditional distribution. In the simplest case of two variables, this implies that $[X \mid Y]$ and the $Y_j^{(i)}$ $(j = 1, \ldots, m)$ should be used to make inferences about $[X]$, rather than imputing $X_j^{(i)}$ $(j = 1, \ldots, m)$ and basing inference on these $X_j^{(i)}$'s. Intuitively, this follows, because to estimate $[X]$ using the $X_j^{(i)}$ requires a kernel density estimate. Such an estimate ignores the known form $[X \mid Y]$ that is mixed to obtain $[X]$. The formal argument is essentially based on the Rao–Blackwell theorem. We sketch a proof in the context of the density estimator itself. If X is a continuous p-dimensional random variable, consider any kernel density estimator of $[X]$ based on the $X_j^{(i)}$ (e.g., see Devroye and Györfi 1985) evaluated at X_0: $\Delta_{X_0}^{(i)} = (1/mh_m^p) \sum_{j=1}^{m} K[(X_0 - X_j^{(i)})/h_m]$, say, where K is a bounded density on R^p and the sequence $\{h_m\}$ is such that as $m \to \infty$, $h_m \to 0$, whereas $mh_m^p \to \infty$. To simplify notation, set $Q_{m,X_0}(X) = (1/h_m^p)K[(X_0 - X)/h_m]$ so that $\Delta_{X_0}^{(i)} = (1/m) \sum_{j=1}^{m} Q_{m,x_0}(X_j^{(i)})$. Define $\gamma_{X_0}^{(i)} = (1/m) \sum_{j=1}^{m} E(Q_{m,X_0}(X) \mid Y_j^{(l)})$. By our earlier theory, both $\Delta_{X_0}^{(i)}$ and $\gamma_{X_0}^{(i)}$ have the same expectation. By the Rao–Blackwell theorem, $\operatorname{var} E(Q_{m,X_0}(X \mid Y)) \leq \operatorname{var} Q_{m,X_0}(X)$, and hence $\mathrm{MSE}(\gamma_{X_0}^{(i)}) \leq \mathrm{MSE}(\Delta_{X_0}^{(i)})$, where MSE denotes the mean squared error of the estimate of $[X_0]$.

Now, for fixed Y, as $m \to \infty$, $E(Q_{m,X_0}(X \mid Y)) \to [X_0 \mid Y]$ for almost every X_0, by the Lebesgue density theorem (see Devroye and Györfi 1985, p. 3). Thus in terms of

random variables we have $E(Q_{m,X_0}(X \mid Y)) \xrightarrow{d} [X_0 \mid Y]$, so for large m, $\gamma_{X_0}^{(i)} \dot{\sim} [\hat{X}_0]_i$ and $\mathrm{MSE}(\gamma_{X_0}^{(i)}) \approx \mathrm{MSE}([\hat{X}_0]_i)$, and hence $[\hat{X}_0]_i$ is preferred to $\Delta_{X_0}^{(i)}$.

The argument is simpler for estimation of $\eta = E(T(X)) = \int T(X) * [X]$, say. Here, $\hat{\eta}_1 = (1/m) \sum_{j=1}^m T(X_j^{(i)})$ is immediately seen to be dominated by $\hat{\eta}_2 = (1/m) \sum_{j=1}^m E(T(X) \mid Y_j^{(i)})$.

3. EXAMPLES

A major area of potential application of the methodology we have been discussing is in the calculation of marginal posterior densities within a Bayesian inference framework. In recent years, there have been many advances in numerical and analytic approximation techniques for such calculations (e.g., see Geweke 1988; Naylor and Smith 1982, 1988; Shaw 1988; Smith et al. 1987; Smith, Skene, Shaw, Naylor, and Dransfield 1985; Tierney and Kadane 1986), but implementation of these approaches typically requires sophisticated numerical analytic expertise, and possibly specialist software. By contrast, the sampling approaches we have discussed are straightforward to implement. For many practitioners, this feature will more than compensate for any relative computational inefficiency. To provide a flavor of the kinds of areas of application for which the methodology is suited, we present six illustrative examples.

3.1 A Class of Multinomial Models

We extend the one-parameter genetic-linkage example described by Tanner and Wong (1987, p. 530), which in its most general form involves multinomial sampling, where some observations are not assigned to individual cells but to aggregates of cells (see Dempster, Laird, and Rubin 1977; Hartley 1958). We give the model and distribution theory in detail for a two-parameter version, from which the extension to k parameters should be clear. Let the vector $Y = (Y_1, \ldots, Y_5)$ have a multinomial distribution $\mathrm{mult}(n, a_1\theta + b_1, a_2\theta + b_2, a_3\eta + b_3, a_4\eta + b_4, c(1 - \theta - \eta))$, where $a_i, b_i \geq 0$ are known and $0 < c = 1 - \sum_{i=1}^4 b_i = a_1 + a_2 = a_3 + a_4 < 1$. Thus θ and η range over $\theta \geq 0$, $\eta \geq 0$, and $\theta + \eta \leq 1$, so a three-parameter Dirichlet distribution, $\mathrm{Dirichlet}(\alpha_1, \alpha_2, \alpha_3)$, may be a natural choice of prior density for (θ, η). From the form of $[Y \mid \theta, \eta] * [\theta, \eta]$, note that obtaining the exact marginals $[\theta \mid Y]$ and $[\eta \mid Y]$ is somewhat messy (involving a two-dimensional numerical integral). Nevertheless, all three sampling approaches we have described are readily applicable here by considering the unobservable nine-cell multinomial model for $X = (X_1, X_2, \ldots, X_9)$, given by $\mathrm{mult}(n, a_1\theta, b_1, a_2\theta, b_2, a_3\eta, b_3, a_4\eta, b_4, c(1 - \theta - \eta))$. From the form of $[X \mid \theta, \eta] * [\theta, \eta]$ we see that $[\theta, \eta \mid X] \sim \mathrm{Dirichlet}(X_1 + X_3 + \alpha_1, X_5 + X_7 + \alpha_2, X_9 + \alpha_3)$, and hence $[\theta \mid X]$ and $[\eta \mid X]$ are available as beta distributions for sampling. Furthermore, $[\theta \mid X, \eta]$ and $[\eta \mid X, \theta]$ are available as scaled beta distributions, scaled to the intervals $[0, 1 - \eta]$ and $[0, 1 - \theta]$, respectively. If we let $Y_1 = X_1 + X_2$, $Y_2 = X_3 + X_4$, $Y_3 = X_5 + X_6$, $Y_4 = X_7 + X_8$, and $Y_5 = X_9$ and define $Z = (X_1, X_3, X_5, X_7)$, we see that specification of X is equivalent to specification of (Y, Z). In addition, $[Z \mid Y, \theta, \eta]$ is the product of four independent binomials for X_1, X_3, X_5, and X_7, given by $[X_i \mid Y, \theta, \eta] = \mathrm{binomial}(Y_i, a_i\theta/(a_i\theta + b_i))$ $(i = 1, 3, 5, 7)$, which are therefore readily available for sampling.

In the context of Section 2, we have a three-variable case, (θ, η, Z), with interest in the marginal distributions $[\theta \mid Y]$, $[\eta \mid Y]$, and $[Z \mid Y]$. Gibbs sampling requires $[\theta \mid Y, Z, \eta]$, $[\eta \mid Y, Z, \theta]$, and $[Z \mid Y, \theta, \eta]$, all of which are available. But in this case the reduced distributions $[\theta \mid Y, Z]$ and $[\eta \mid Y, Z]$ are available as well, enabling study of accelerated convergence. These reduced distributions substantially simplify the Rubin importance-sampling algorithm in obtaining $[\theta \mid Y]$ and $[\eta \mid Y]$; only an importance-sampling distribution $[Z \mid Y]_s$ need be specified (e.g., a default choice might be binomials with chance equal to $\frac{1}{2}$). Detailed comparison of the performance of the three algorithms for a specific case of this multinomial class is given in Section 4.

3.2 Hierarchical Models Under Conjugacy

Consider a general Bayesian hierarchical model having k stages. In an obvious notation, we write the joint distribution of the data and parameters as

$$[Y \mid \theta_1] * [\theta_1 \mid \theta_2] * [\theta_2 \mid \theta_3] * \cdots * [\theta_{k-1} \mid \theta_k] * [\theta_k], \quad (16)$$

where we assume all components of prior specification to be available for sampling. Primary interest is usually in the marginal posterior $[\theta_1 \mid Y]$. The hierarchical structure implies that

$$\begin{aligned} [\theta_i \mid Y, \theta_{jj} \neq i] &= [\theta_1 \mid Y, \theta_2], & i &= 1 \\ &= [\theta_i \mid \theta_{i-1}, \theta_{i+1}], & 1 < i &< k - 1 \\ &= [\theta_k \mid \theta_{k-1}], & i &= k. \end{aligned} \quad (17)$$

Suppose that we assume proper conjugate distributions at each stage. This is common practice in the formulation of such models, except perhaps for $[\theta_k]$, which is often assumed vague. Nevertheless, conjugate priors can generally be made arbitrarily diffuse by appropriate choices of hyperparameters, so this case is implicitly subsumed within the conjugate framework. In fact, $[\theta_k]$ can be vague, provided $[\theta_k \mid \theta_{k-1}]$ is still proper and available (see Secs. 3.4 and 3.5). Conjugacy implies that the densities in (17) will be available as updated versions of the respective priors (e.g., see Morris 1983a). Typically, no distinct reduced conditional distributions are available, and Gibbs sampling would be used to estimate the desired marginal posterior densities. To clarify this latter point, consider the case $k = 3$. The six conditional distributions in (10) would be $[\theta_1 \mid y, \theta_2, \theta_3]$, $[\theta_2 \mid y, \theta_1, \theta_3]$, $[\theta_3 \mid y, \theta_1, \theta_2]$, $[\theta_3 \mid y, \theta_2]$, $[\theta_1 \mid y, \theta_3]$, and $[\theta_2 \mid y, \theta_1]$. The first three are available as in (17), the fourth is available but not distinct from the third, and the last two are usually unavailable.

As a concrete illustration, consider an exchangeable Poisson model, which is illustrated further in Section 4 with the reanalysis of a published data set. Suppose that we observe independent counts, s_i, over differing lengths of time, t_i (with resultant rate $\rho_i = s_i/t_i$) $(i = 1, \ldots, p)$. Assume $[s_i \mid \lambda_i] = P_0(\lambda_i t_i)$ and that the λ_i are iid from $G(\alpha,$

β), with density $\lambda_i^{\alpha-1}e^{-\lambda_i/\beta}/\beta^\alpha\Gamma(\alpha)$. The parameter α is assumed known (in practice, we might treat α as a tuning parameter, or perhaps, in an empirical Bayes spirit, estimate it from the marginal distribution of the s_i's), and β is assumed to arise from an inverse gamma distribution IG(γ, δ) with density $\delta^\gamma e^{-\delta/\beta}/\beta^{\gamma+1}\Gamma(\gamma)$. (A diffuse version of this final-stage distribution is obtained by taking δ and γ to be very small, perhaps 0.)

Letting $Y = (s_1, \ldots, s_p)$, the conditional distributions $[\lambda_j \mid Y]$ are sought. The full conditional distribution of λ_j is given by

$$[\lambda_j \mid Y, \beta, \lambda_{i,i\neq j}] = G(\alpha + s_j, (t_j + 1/\beta)^{-1}), \quad j = 1, \ldots, p, \quad (18)$$

whereas the full conditional distribution for β is given by

$$[\beta \mid Y, \lambda_1, \ldots, \lambda_p] = \mathrm{IG}(\gamma + p\alpha, \Sigma\lambda_i + \delta). \quad (19)$$

No distinct reduced conditional distributions are available. The conditional distribution of λ_j, given Y and β, is (18), regardless of which or how many λ_i ($i \neq j$) are given. The conditional distribution of β, given Y and any subset of the λ_j's, is unavailable. Given $(\lambda_1^{(0)}, \lambda_2^{(0)}, \ldots, \lambda_p^{(0)}, \beta^{(0)})$, the Gibbs sampler draws $\lambda_j^{(1)} \sim G(\alpha + s_j, (t_j + 1/\beta^{(0)})^{-1})$ ($j = 1, \ldots, p$) and $\beta^{(1)} \sim \mathrm{IG}(\gamma + \alpha p, \Sigma_{j=1}^p \lambda_j^{(1)} + \delta)$ to complete one cycle. If we carry out m repetitions each of i iterations, generating $(\lambda_{1l}^{(i)}, \ldots, \lambda_{rl}^{(i)}, \beta_l^{(i)})$ ($l = 1, \ldots, m$), the marginal density estimate for λ_j is

$$[\lambda_j \uparrow Y] = \frac{1}{m}\sum_{l=1}^{m} G\left(\alpha + s_j, \left(t_j + \frac{1}{\beta_l^{(i)}}\right)^{-1}\right), \quad j = 1, \ldots, p, \quad (20)$$

whereas

$$[\beta \uparrow Y] = \frac{1}{m}\sum_{l=1}^{m} \mathrm{IG}(\gamma + \alpha p, \Sigma\lambda_{jl}^{(i)} + \delta). \quad (21)$$

Rubin's importance-sampling algorithm is applicable in the setting (16) as well, taking a particularly simple form in the cases $k = 2, 3$. For $k = 3$, suppose that we seek $[\theta_1 \mid y]$. The joint density $[\theta_1, \theta_2, \theta_3 \mid Y] = [Y, \theta_1, \theta_2, \theta_3]/[Y]$, where the functional form of the numerator is given in (16). An importance-sampling density for $[\theta_1, \theta_2, \theta_3 \mid Y]$ could be sampled as $[\theta_1 \mid Y, \theta_2] * [\theta_3 \mid \theta_2] * [\theta_2 \mid Y]_s$ for some $[\theta_2 \mid Y]_s$. As remarked in Section 2.5, a good choice for $[\theta_2 \mid Y]_s$ might be obtained through a few iterations of the substitution-sampling algorithm. In any case, for $l = 1, \ldots, N$ we would generate θ_{2l} from $[\theta_2 \mid Y]_s$, θ_{3l} from $[\theta_3 \mid \theta_{2l}]$, and θ_{1l} from $[\theta_1 \mid Y, \theta_{2l}]$. Calculating

$$r_l = \frac{[Y, \theta_{1l}, \theta_{2l}, \theta_{3l}]}{[\theta_{1l} \mid Y, \theta_{2l}] * [\theta_{3l} \mid \theta_{2l}] * [\theta_{2l} \mid Y]_s},$$

we obtain the density estimator $[\theta_1 \uparrow Y] = \Sigma[\theta_1 \mid Y, \theta_{2l}]r_l/\Sigma r_l$. Note that (in the terminology of Rubin) the algorithm in this case can be streamlined by writing the joint density in the numerator of r_l as $[\theta_{1l} \mid Y, \theta_{2l}] * [Y \mid \theta_{2l}] * [\theta_{2l} \mid \theta_{3l}] * [\theta_{3l}]$ and noting that r_l does not involve θ_{1l}, so we need not actually generate the θ_{1l}.

Returning to the exchangeable Poisson model, the estimator of the marginal density of λ_j under Rubin's importance-sampling algorithm is

$$[\lambda_j \uparrow Y] = \sum_{l=1}^{N} G\left(\alpha + s_j, \left(t_j + \frac{1}{\beta_l}\right)^{-1}\right) r_l \Big/ \sum_{l=1}^{N} r_l.$$

Here $r_l = [Y \mid \beta_l] * [\beta_l]/[\beta_l \mid Y]_s$, where $[Y \mid \beta_l]$ is the product of negative binomial densities; that is,

$$[Y \mid \beta_l] = \prod_{j=1}^{p}\left(\frac{\Gamma(s_j + \alpha)t_j^{s_j}\beta_l^\alpha}{s_j!\Gamma(\alpha)(t_j + \beta_l)^{s_j+\alpha}}\right),$$

and $[\beta_l]$ is the IG prior evaluated at β_l. If $[\beta \mid Y]_s$ is not obtained from the substitution-sampling algorithm, as in (21), an alternative choice is IG($\gamma + \alpha p$, $\Sigma_{i=1}^p \rho_i + 1$). This arises because $[\beta \mid Y] = E_{[\lambda_1,\ldots,\lambda_p|Y]}[\beta \mid Y, \lambda_1, \ldots, \lambda_p] \approx [\beta \mid Y, \hat{\lambda}_1, \ldots, \hat{\lambda}_p]$, using $\hat{\lambda}_j = \rho_j$ in (19).

3.3 Multivariate Normal Sampling

A commonly occurring problem in combining continuous multivariate data is that often not all variables are observed for each experimental unit (e.g., see Dempster et al. 1977). If the data are sampled from multivariate normal populations with conjugate priors for the mean and covariance matrix, we have a general class of models where all full conditional distributions and at least some reduced conditional distributions are available. We illustrate in the simplest case, where we assume that $\binom{U_{1i}}{U_{2i}}$ ($i = 1, \ldots, n_1$), $\binom{V_{1j}}{V_{2j}}$ ($j = 1, \ldots, n_2$), and $\binom{W_{1k}}{W_{2k}}$ ($k = 1, \ldots, n_3$) are all iid $N(\theta, \Delta)$ with $\theta \sim N(\mu, \Sigma)$, where $\theta = \binom{\theta_1}{\theta_2}$ is not observable but $\mu = \binom{\mu_1}{\mu_2}$, Δ, and Σ are assumed known. Let $U = \binom{U_1}{U_2} = \binom{U_{11}\cdots U_{1n_1}}{U_{21}\cdots U_{2n_1}}$, with similar notation for V and W. Finally, let $X = (U, V, W)$ and $2 \times N$, with $\overline{X} = N^{-1}X1$, where 1 is a column vector of N 1s and $N = n_1 + n_2 + n_3$. Standard calculations show that $[\theta \mid X]$ is $N(\eta, \Omega)$, where $\eta = (N\Delta^{-1} + \Sigma^{-1})^{-1}(N\Delta^{-1}\overline{X} + \Sigma^{-1}\mu)$ and $\Omega = (N\Delta^{-1} + \Sigma^{-1})^{-1}$. With the obvious partitioning, $\eta = \binom{\eta_1}{\eta_2}$, $\Omega = \begin{pmatrix}\Omega_{11} & \Omega_{12}\\ \Omega_{21} & \Omega_{22}\end{pmatrix}$, the marginals $[\theta_1 \mid X] = N(\eta_1, \Omega_{11})$, and $[\theta_2 \mid X] = N(\eta_2, \Omega_{22})$ are available. Suppose, however, that V_2 and W_1, say, are unobserved. Let $Y = (U, V_1, W_2)$ and $Z = (V_2, W_1)$ so that $X \equiv (Y, Z)$. As in Section 3.1, we have a three-variable problem, here involving θ_1, θ_2, and Z. The full conditional distributions are all normal and hence available. For θ_1 and θ_2, $[\theta_1 \mid Y, Z, \theta_2] = N(\eta_1 + \Omega_{12}\Omega_{22}^{-1}(\theta_2 - \eta_2), \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21})$ and $[\theta_2 \mid Y, Z, \theta_1] = N(\eta_2 + \Omega_{21}\Omega_{11}^{-1}(\theta_1 - \eta_1), \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12})$. Letting $\overline{U}_1 = n_1^{-1}U_1 1$, with similar notation for $\overline{U}_2$, $\overline{V}_1$, $\overline{V}_2$, $\overline{W}_1$, and $\overline{W}_2$, we note by sufficiency that with regard to Z we only need the full posterior $[\overline{Z} \mid \overline{Y}, \theta_1, \theta_2]$, where $\overline{Z}^T = (\overline{V}_2, \overline{W}_1)$ and $\overline{Y}^T = (\overline{U}_1, \overline{U}_2, \overline{V}_1, \overline{W}_2)$. Since

$$\overline{X}^T \equiv (\overline{U}_1, \overline{U}_2, \overline{V}_1, \overline{V}_2, \overline{W}_1, \overline{W}_2) \mid (\theta_1, \theta_2) \sim N\left(\begin{pmatrix}\theta\\ \theta\\ \theta\end{pmatrix}, \begin{pmatrix} n_1^{-1}\Delta & 0 & 0\\ 0 & n_2^{-1}\Delta & 0\\ 0 & 0 & n_3^{-1}\Delta\end{pmatrix}\right),$$

the conditional distribution $[\overline{Z} \mid \overline{Y}, \theta_1, \theta_2]$ is clearly normal. With the full conditionals and the reduced conditionals $[\theta_1 \mid Y, Z]$ and $[\theta_2 \mid Y, Z]$ available, the accelerated substitution algorithm can be used to obtain $[\theta_1 \mid Y]$ and $[\theta_2 \mid Y]$.

The Rubin importance-sampling algorithm is straightforward in this case. Simplifying notation by working with the sufficient statistic $(\overline{Y}, \overline{Z})$, suppose that we seek the density estimator of $[\theta_1 \mid Y]$ for instance. We have $[\theta_1 \uparrow Y] = \Sigma[\theta_1 \mid \overline{Y}, \overline{Z}_l, \theta_{2l}]r_l/\Sigma r_l$, where

$$r_l = \frac{[\overline{X}_l \mid \theta_{1l}, \theta_{2l}] * [\theta_{1l}, \theta_{2l}]}{[\theta_{1l} \mid \overline{Y}, \overline{Z}_l, \theta_{2l}] * [\theta_{2l} \mid \overline{Y}, \overline{Z}_l] * [\overline{Z}_l \mid \overline{Y}]_s},$$

with $\overline{X}_l \equiv (\overline{Y}, \overline{Z}_l)$ and $[\overline{Z} \mid \overline{Y}]_s$ a specified importance-sampling density. Thus for $l = 1, \ldots, N$ we generate $\overline{Z}_l \sim [\overline{Z} \mid \overline{Y}]$, $\theta_{2l} \sim [\theta_2 \mid \overline{Y}, \overline{Z}_l]$, and $\theta_{1l} \sim [\theta_1 \mid \overline{Y}, \overline{Z}_l, \theta_{2l}]$. Again, the choice of $[\overline{Z} \mid \overline{Y}]_s$ could be made using a few iterations of substitution sampling, or perhaps based on the intuitively appealing estimated conditional form, $[\overline{Z} \mid \overline{Y}, \hat{\theta}_1, \hat{\theta}_2]$, where $\hat{\theta}_1 = (n_1\overline{U}_1 + n_2\overline{V}_1)/(n_1 + n_2)$ and $\hat{\theta}_2 = (n_1\overline{U}_2 + n_3\overline{W}_3)/(n_1 + n_3)$.

3.4 Variance Component Models

Bayesian inference for variance components has typically required subtle numerical analysis or intricate analytic approximation (e.g., as evidenced by Box and Tiao 1973, chaps. 5 and 6). In marked contrast to such sophistication, marginal posterior densities for variance components are readily obtained through simple Gibbs sampling.

We illustrate this for the simplest variance components model defined by $Y_{ij} = \theta_i + \varepsilon_{ij}$ $(i = 1, \ldots, K, j = 1, \ldots, J)$, where, assuming conditional independence throughout, $[\theta_i \mid \mu, \sigma_\theta^2] = N(\mu, \sigma_\theta^2)$ and $[\varepsilon_{ij} \mid \sigma_e^2] = N(0, \sigma_e^2)$, so $[Y_{ij} \mid \theta_i, \sigma_e^2] = N(\theta_i, \sigma_e^2)$.

Let $\theta = (\theta_1, \ldots, \theta_K)$ and $Y = (Y_{11}, \ldots, Y_{KJ})$ and assume that μ, σ_θ^2, and σ_e^2 are independent, with priors specified by $[\mu] \sim N(\mu_0, \sigma_0^2)$, $[\sigma_\theta^2] \sim \mathrm{IG}(a_1, b_1)$, and $[\sigma_e^2] \sim \mathrm{IG}[a_2, b_2]$, where μ_0, σ_0^2, a_1, b_1, a_2, and b_2 are assumed known (possibly chosen to correspond to diffuse priors).

The joint distribution $[Y, \theta, \mu, \sigma_\theta^2, \sigma_e^2]$ can be written as

$$[Y \mid \theta, \sigma_e^2] * [\theta \mid \mu, \sigma_\theta^2] * [\mu] * [\sigma_\theta^2] * [\sigma_e^2], \qquad (22)$$

and we follow Box and Tiao (1973, chap. 5) in focusing interest on $[\sigma_\theta^2 \mid Y]$ and $[\sigma_e^2 \mid Y]$.

From the Gibbs sampling perspective, we have a four-variable system, $(\theta, \mu, \sigma_\theta^2, \sigma_e^2)$, with the following full conditional distributions:

$$\begin{aligned}
[\sigma_\theta^2 \mid Y, \mu, \theta, \sigma_e^2] &= [\sigma_\theta^2 \mid \mu, \theta] \\
&= \mathrm{IG}(a_1 + \tfrac{1}{2}K, b_1 + \tfrac{1}{2}\Sigma(\theta_i - \mu)^2), \\
[\sigma_e^2 \mid Y, \mu, \theta, \sigma_\theta^2] &= [\sigma_e^2 \mid Y, \theta] \\
&= \mathrm{IG}(a_2 + \tfrac{1}{2}KJ, b_2 + \tfrac{1}{2}\Sigma\Sigma(Y_{ij} - \mu_i)^2), \\
[\mu \mid Y, \theta, \sigma_\theta^2, \sigma_e^2] &= [\mu \mid \sigma_\theta^2, \theta] \\
&= N\left(\frac{\sigma_\theta^2\mu_0 + \sigma_0^2\Sigma\theta_i}{\sigma_\theta^2 + K\sigma_0^2}, \frac{\sigma_\theta^2\sigma_0^2}{\sigma_\theta^2 + K\sigma_0^2}\right),
\end{aligned}$$

and

$$\begin{aligned}
&[\theta \mid Y, \mu, \sigma_\theta^2, \sigma_e^2] \\
&= N\left(\frac{J\sigma_\theta^2}{J\sigma_\theta^2 + \sigma_e^2}\overline{Y} + \frac{\sigma_e^2}{J\sigma_\theta^2 + \sigma_e^2}\mu 1, \frac{\sigma_\theta^2\sigma_e^2}{J\sigma_\theta^2 + \sigma_e^2} I\right),
\end{aligned}$$

where $\overline{Y}^T = (\overline{Y}_1, \ldots, \overline{Y}_I)$, $\overline{Y}_{i.} = (1/J)\sum_{j=1}^J Y_{ij}$, 1 is a $K \times 1$ column vector of 1s, and I is a $K \times K$ identity matrix.

Since all of these full conditionals are available, implementation of the Gibbs sampler is straightforward. Moreover, extensions to more elaborate variance component models follow precisely the same pattern, since the full conditional distributions for μ and θ continue to be normal, and those for the variance components continue to be IG.

3.5 Normal Means Model

The exchangeable k-group normal means model with different, unknown measurement variances in each group provides a simple example of an unbalanced class of models that has proved difficult to handle using empirical Bayes approaches to estimating posterior distributions (e.g., see Morris 1983b, 1987). Such models are straightforwardly handled by iterative sampling approaches, as we saw with the Poisson example of Section 3.2 and further illustrate here for this classical normal means example.

Suppose, then, assuming conditional independence throughout, that $Y_{ij} \sim N(\theta_i, \sigma_i^2)$, $\theta_i \sim N(\mu, \tau^2)$, $\sigma_i^2 \sim \mathrm{IG}(a_1, b_1)$ $(i = 1, \ldots, I, j = 1, \ldots, J_i)$, $\mu \sim N(\mu_0, \sigma_0^2)$, and $\tau^2 \sim \mathrm{IG}(a_2, b_2)$, where μ_0, σ_0^2, a_1, b_1, a_2, and b_2 are assumed known (possibly chosen to reflect diffuse prior information). By sufficiency, we can confine attention to $Y = \{(\overline{Y}_{i.}, S_i^2); i = 1, \ldots, I\}$, where $\overline{Y}_{i.} = (1/J_i)\Sigma Y_{ij}$ and $S_i^2 = (1/J_i)\Sigma(Y_{ij} - \overline{Y}_{i.})^2$. Then, if we write $\theta = (\theta_1, \ldots, \theta_I)$ and $\sigma^2 = (\sigma_1^2, \ldots, \sigma_I^2)$, the joint distribution of Y, θ, σ^2, μ, and τ^2 takes the form

$$[Y \mid \theta, \sigma^2] * [\theta \mid \mu, \tau^2] * [\sigma^2] * [\mu] * [\tau^2], \qquad (23)$$

where

$$\begin{aligned}
&[Y \mid \theta, \sigma^2] * [\theta \mid \mu, \tau^2] * [\sigma^2] \\
&\quad = \prod_{i=1}^{I} [Y_i \mid \theta_i, \sigma_i^2] * [S_i^2 \mid \sigma_i^2] * [\theta_i \mid \mu, \tau^2] * [\sigma_i^2].
\end{aligned}$$

Of course, there is an obvious similarity between (22) and (23), but here we focus on $[\theta_i \mid Y]$ $(i = 1, \ldots, I)$. From the Gibbs sampling perspective, this is a $(2I + 2)$-variable problem: (θ_i, σ_i^2) $(i = 1, \ldots, I)$, together with μ and τ^2. To identify the forms of the full conditionals, we first note that

$$[\theta \mid Y, \sigma^2, \mu, \tau^2] = N(\theta^*, D^*), \qquad (24)$$

where $\theta_i^* = (J_i\overline{Y}_{i.}\tau^2 + \mu\sigma_i^2)/(J_i\tau^2 + \sigma_i^2)$, $D_{ii}^* = \sigma_i^2\tau^2/(J_i\tau^2 + \sigma_i^2)$, and $D_{ij}^* = 0$ $(i \neq j)$. Thus the full conditional distributions $[\theta_i \mid Y, \theta_i j \neq i, \sigma^2, \mu, \tau^2]$ $(i = 1, \ldots, I)$ are just the normal marginals of (24) and therefore available for sampling. From (23), we easily see that $[\sigma^2 \mid Y, \theta, \mu, \tau^2] = [\sigma^2 \mid Y, \theta] = \prod_{i=1}^I [\sigma_i^2 \mid \overline{Y}_{i.}, S_i^2, \theta_i]$, where $[\sigma_i^2 \mid \overline{Y}_{i.}, S_i^2, \theta_i] = \mathrm{IG}(a_1 + \tfrac{1}{2}J_i, b_1 + \tfrac{1}{2}\Sigma_j(Y_{ij} - \theta_i)^2)$. Finally, and closely resembling the forms obtained in Section 3.4,

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

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

3.6 An Errors-in-Variables Model

Again, we consider a simple illustrative special case. Consider Y to be a vector of responses assumed related to levels X of a covariate according to the straight-line model

$$Y \sim N\left((1X)\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix}, \sigma^2 I\right).$$

Responses are obtained at specified levels X_0 of the covariate, but suppose that these are not the actual levels X_a. Rather, given the former, beliefs about the latter are represented by $X_a \sim N(X_0, \tau^2 I)$. Interest centers on $\theta = (\theta_1, \theta_2)$, and to complete the distributional specification, suppose that we place independent conjugate priors on θ, σ^2, and τ^2. The joint distribution on $(Y, X_a, \theta, \sigma^2, \tau^2)$ then has the form

$$[Y \mid X_a, \theta, \sigma^2] * [X_a \mid \tau^2] * [\tau^2] * [\theta] * [\sigma^2], \quad (25)$$

where again there is obvious similarity to (22) and (23). The Gibbs sampler requires $[\theta \mid Y, X_a, \sigma^2, \tau^2] = [\theta \mid Y, X_a, \sigma^2]$, $[\sigma^2 \mid Y, X_a, \theta, \tau^2] = [\sigma^2 \mid Y, X_a, \theta]$, $[\tau^2 \mid Y, X_a, \theta, \sigma^2] = [\tau^2 \mid X_a]$, and $[X_a \mid Y, \theta, \sigma^2, \tau^2]$. If we assume a normal prior for θ and IG priors for σ^2 and τ^2, we obtain normal full conditionals for θ and X_a and IG full conditionals for σ^2 and τ^2. We omit the details, which are somewhat similar to those in Sections 3.4 and 3.5.

4. NUMERICAL ILLUSTRATIONS

4.1 A Multinomial Model

We provide some preliminary insights into the relative performance and properties of the substitution-, Gibbs, and Rubin importance-sampling approaches by considering an artificial problem based on the class of multinomial models discussed in Section 3.1.

We suppose that data $Y = (Y_1, Y_2, Y_3, Y_4, Y_5) = (14, 1, 1, 1, 5)$ are available as a sample from the multinomial distribution $\mathrm{mult}(22, \frac{1}{4}\theta + \frac{1}{8}, \frac{1}{4}\eta, \frac{1}{4}\eta + \frac{3}{8}, \frac{1}{2}(1 - \theta - \eta))$, and that the prior for (θ, η) is taken to be a Dirichlet(1, 1, 1) distribution. In the general notation of Section 3.1, we therefore have $a_1 = \frac{1}{4}$, $b_1 = \frac{1}{8}$, $a_2 = \frac{1}{4}$, $b_2 = 0$, $a_3 = \frac{1}{4}$, $b_3 = 0$, $a_4 = \frac{1}{4}$, $b_4 = \frac{3}{8}$, and $\alpha_1 = \alpha_2 = \alpha_3 = 1$, with interest centering on the calculation of the marginal posterior densities $[\theta \mid Y]$ and $[\eta \mid Y]$.

By considering instead a split-cell multinomial, which in this case takes the form

$$X = (X_1, X_2, \ldots, X_7) \sim \mathrm{mult}(22, \tfrac{1}{4}\theta, \tfrac{1}{8}, \tfrac{1}{4}\theta, \tfrac{1}{4}\eta, \tfrac{1}{4}\eta, \tfrac{3}{8}, \tfrac{1}{2}(1 - \theta - \eta)),$$

we can use the analysis of Section 3.1 for this special case of a seven-cell multinomial to construct substitution- and Gibbs sampling algorithms involving θ, η, and $Z = (X_1, X_5)$.

As noted in Section 3.1, we can compare the two forms of iterative sampling. To do so, we first obtained very accurate numerical estimates of $[\theta \mid Y]$ and $[\eta \mid Y]$ using techniques described by Smith et al. (1985, 1987), and from these obtained the true 5, 25, 50, 75, and 95 posterior percentile points for each parameter. Iterative cycles of the two samplers were then run, calibrated so that the total number of random variates generated was the same in both cases (as described in Sec. 2.4). The initialization was defined (for an arbitrary generating seed) in each case by taking independent samples from $\theta \sim U(0, 1)$ and $\eta \sim U(0, 1)$, subject to $0 \leq \theta + \eta \leq 1$. At each cycle, $m = 10$ drawings of the parameters were then made, and from estimates of the form (9) estimates of the cumulative posterior probabilities corresponding to each of the five true percentile points for each parameter were obtained. This process was replicated 5,000 times, enabling us to study the mean estimates of the cumulative probabilities, together with their standard errors, as well as the percentage of occasions on which each sampler was closest to the true value. A summary of the results following each of the first four cycles is given in Table 1.

Table 1. Comparison of Substitution (S) and Gibbs (G) Samplers

		Estimate (SE)				S closer than G	
		θ		η			
Cycle	cdf value	G	S	G	S	θ	η
1	.05	.231 (.08)	.217 (.08)	.033 (.01)	.044 (.01)	56%	75%
	.25	.504 (.10)	.492 (.09)	.177 (.04)	.225 (.04)	55%	78%
	.50	.713 (.08)	.706 (.08)	.380 (.06)	.459 (.06)	54%	80%
	.75	.873 (.05)	.871 (.05)	.620 (.06)	.706 (.06)	51%	80%
	.95	.978 (.01)	.978 (.01)	.878 (.04)	.926 (.03)	49%	80%
2	.05	.067 (.04)	.055 (.03)	.047 (.01)	.048 (.01)	56%	51%
	.25	.286 (.07)	.266 (.07)	.236 (.04)	.241 (.04)	56%	52%
	.50	.535 (.08)	.522 (.07)	.478 (.06)	.487 (.06)	53%	52%
	.75	.773 (.06)	.768 (.05)	.728 (.05)	.737 (.05)	51%	52%
	.95	.956 (.02)	.956 (.02)	.940 (.02)	.944 (.02)	51%	53%
3	.05	.052 (.03)	.049 (.03)	.049 (.01)	.049 (.01)	51%	50%
	.25	.254 (.06)	.252 (.06)	.247 (.04)	.247 (.04)	51%	50%
	.50	.505 (.07)	.508 (.07)	.496 (.06)	.496 (.06)	51%	49%
	.75	.754 (.06)	.760 (.05)	.746 (.05)	.747 (.05)	51%	50%
	.95	.951 (.02)	.954 (.02)	.949 (.02)	.949 (.02)	51%	50%
4	.05	.050 (.03)	.047 (.03)	.050 (.01)	.050 (.01)	51%	51%
	.25	.250 (.06)	.249 (.06)	.250 (.04)	.249 (.04)	50%	51%
	.50	.500 (.07)	.505 (.07)	.499 (.06)	.499 (.06)	51%	51%
	.75	.751 (.06)	.757 (.05)	.750 (.05)	.751 (.05)	51%	51%
	.95	.950 (.02)	.953 (.02)	.950 (.02)	.951 (.02)	51%	49%

NOTE: Standard errors (SE's) are in parentheses

We note from Table 1 that initially (cycles 1 and 2) the substitution sampler adapts more quickly than the Gibbs sampler, particularly for η. By the time we reach the third and fourth cycles, however, the two approaches are performing indistinguishably. What is astonishing, perhaps, is how remarkably good their performance is. By the fourth cycle, using only $m = 10$ drawings and starting from a default noninformative baseline, the marginal posterior density estimators based on (8) are providing on average extremely accurate estimates of cumulative probabilities. Our experiences with this and other examples (see Sec. 4.2) suggest that satisfactory convergence with iterative sampling requires only a small fraction of the levels of random variate generation reported by Tanner and Wong (1987).

The noniterative Rubin importance-sampling algorithm (Sec. 2.5) requires us to choose a sampling density, $[Z \mid Y]_s$, and then to proceed as follows, for $l = 1, \ldots, m$: Draw Z_l from $[Z \mid Y]_s$, η_l from $[\eta \mid Z, Y]$, and θ_l from $[\theta \mid \eta, Z, Y]$, with the latter two distributions as detailed previously, thus creating a triple (θ_l, η_l, Z_l). Then, calculate

$$r_l = \frac{[Y, Z_l \mid \theta_l, \eta_l] * [\theta_l, \eta_l]}{[\theta_l \mid \eta_l, Z_l, Y] * [\eta_l \mid Z_l, Y] * [Z_l \mid Y]_s},$$

and form estimates $[\theta \uparrow Y] = \sum_{l=1}^{m} [\theta \mid \eta_l, Z_l, Y] r_l / \sum_{l=1}^{m} r_l$ and $[\eta \uparrow Y] = \sum_{l=1}^{m} [\eta \mid \theta_l, Z_l, Y] r_l / \sum_{l=1}^{m} r_l$.

Table 2 shows the average cumulative posterior probability estimates from this approach, based on 2,500 replicates of $m = 40$ and $m = 200$ and taking $[Z \mid Y]_s$ to be the product of $X_1 \sim \text{binomial}(Y_1, \frac{1}{2})$ and $X_5 \sim \text{binomial}(Y_4, \frac{1}{2})$. Despite the much larger number of drawings compared with the iterative samplers, the estimation is rather poor. In general, experience suggests that the algorithm is highly sensitive to the choice of $[Z \mid Y]_s$ and that the larger one-off simulation is no match for iterative adaptation via small simulations.

4.2 A Conjugate Hierarchical Model

We apply the exchangeable Poisson model discussed in Section 3.2 to data on pump failures previously analyzed by Gaver and O'Muircheartaigh (1987) (reproduced here in Table 3), where s_i is the number of failures and t_i is the length of time in thousands of hours.

Recalling the model structure of Section 3.2 and the forms of conditional distribution given by (18) and (19), we illustrate the use of the Gibbs sampler for this data set, with $p = 10$, $\delta = 1$, $\gamma = 0.1$, and, for the purposes of illustration, $\alpha = \bar{p}^2/(S_p^2 - p^{-1}\bar{p}\sum_{i=1}^{p} t_i^{-1})$, with the latter derived by a method-of-moments empirical Bayes argument based on $E(\rho_i) = EE(\rho_i \mid \lambda_i) = \alpha/\beta \approx \bar{p}$:

$$V(\rho_i) = VE(\rho_i \mid \lambda) + EV(\rho_i \mid \lambda_i)$$
$$= (\alpha/\beta^2) + (\alpha/\beta t_i) \approx S_p^2 = p^{-1}\Sigma(\rho_i - \bar{p})^2.$$

Figure 1 shows a selection of four marginal posterior densities (for $\lambda_2, \lambda_4, \lambda_8, \lambda_9$) calculated from (20) following a run of 10 cycles of the algorithm. In fact, three densities are superposed: One corresponds to $m = 10$, one to $m = 100$, and the third is the exact density calculated using techniques described by Smith et al. (1985, 1987). Even in the cases of λ_8 and λ_9 (chosen as worst cases from $\lambda_1, \ldots, \lambda_{10}$), the densities are hardly distinguishable—a remarkable convergence from such a small number of drawings.

Table 2. Estimates From the Rubin Importance-Sampling Algorithm

	Estimates: m = 40 (200)	
cdf value	θ	η
.05	.105 (.150)	.049 (.049)
.25	.311 (.351)	.244 (.241)
.50	.521 (.537)	.485 (.477)
.75	.739 (.734)	.729 (.714)
.95	.939 (.932)	.934 (.921)

Table 3. Pump-Failure Data

Pump system	s_i	t_i	ρ_i ($\times 10^2$)
1	5	94.320	5.3
2	1	15.720	6.4
3	5	62.880	8.0
4	14	125.760	11.1
5	3	5.240	57.3
6	19	31.440	60.4
7	1	1.048	95.4
8	1	1.048	95.4
9	4	2.096	191.0
10	22	10.480	209.9

5. DISCUSSION

We have emphasized providing a comparative review and explication of three possible sampling approaches to the calculation of intractable marginal densities. The substitution-, Gibbs, and importance-sampling algorithms are all straightforward to implement in several frequently occurring practical situations, thus avoiding complicated numerical or analytic approximation exercises (often necessitating intricate attention to reparameterization and other subtleties requiring case-by-case consideration). For this latter reason if for no other the techniques deserve to be better known and experimented with for a wide range of problems. We hope that the unified exposition attempted here will provide a general, clarifying perspective within which to view the work of Geman and Geman (1984), Rubin (1987, 1988), and Tanner and Wong (1987), and to evaluate its potential for other structured problems. For example, in addition to the model structures given in Section 3, the methods find immediate and powerful application to problems involving ordered parameters or change points. In future work we shall provide detailed and extensive numerical illustration of many such problems.

The preliminary computational experience reported here illustrates the following points. Iterative, adaptive sampling (substitution or Gibbs) invariably provides better value, in terms of efficient use of generated variates, than an equivalent sample-size, noniterative, one-off approach (Rubin), provided a suitable structure for iterative sampling exists. In problems where certain reduced conditionals are available, there is scope for accelerating the substitution algorithm so that it becomes more efficient (particularly in early cycles) than the Gibbs algorithm; however, the gain in efficiency is only likely to be of consequence when the number of reduced conditionals is a relatively large fraction of the total number of conditionals involved in a cycle. There are important practical problems in tuning monitoring and stopping-rules procedures for iterative sampling in large-scale complex problems; we shall report on these in future work as well. Finally, we note that even in cases where ultimate convergence of the iterative sampling procedures proves slow, moment or other information provided by a few initial cycles can be used to provide highly effective starting values for more

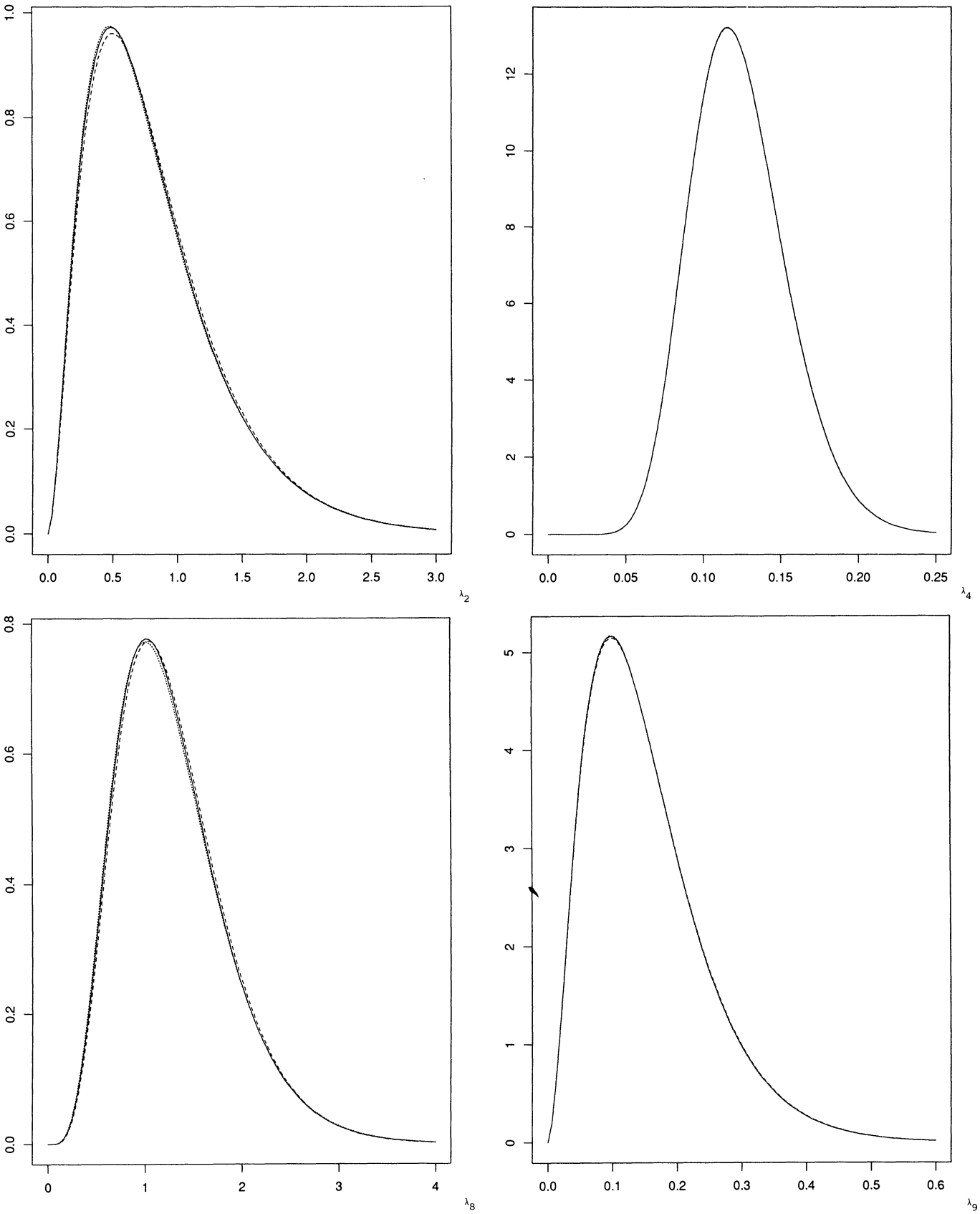

Figure 1. Density Estimates for Pump-Failure Data: ···, m = 10; - - -, m = 100; ——, Exact.

sophisticated numerical or analytic approximation techniques.

[*Received November 1988. Revised October 1989.*]

REFERENCES

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.

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

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

Devroye, L., and Györfi, L. (1985), *Non-parametric Density Estimation: The L_1 View*, New York: John Wiley.

Gaver, D., and O'Muircheartaigh, I. (1987), "Robust Empirical Bayes Analysis of Event Rates," *Technometrics*, 29, 1–15.

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), "Antithetic Acceleration of Monte Carlo Integration in Bayesian Inference," *Journal of Econometrics*, 38, 73–90.

Glick, N. (1974), "Consistency Conditions for Probability Estimators and Integrals of Density Estimators," *Utilitas Mathematica*, 6, 61–74.

Hartley, H. (1958), "Maximum Likelihood Estimation From Incomplete Data," *Biometrics*, 14, 174–194.

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

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

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

——— (1987), "Determining the Accuracy of Bayesian Empirical Bayes Estimates in Familiar Exponential Families," in *Statistical Decision Theory and Related Topics* (Vol. 4), eds. S. S. Gupta and J. O. Berger, New York: Springer-Verlag, pp. 251–264.

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 Inferences," *Journal of Econometrics*, 38, 103–126.

Rall, L. (1969), *Computational Solution of Non-linear Operator Equations*, New York: John Wiley.

Ripley, B. (1987), *Stochastic Simulation*, New York: John Wiley.

Rubin, D. B. (1987), Comment on "The Calculation of Posterior Distributions by Data Augmentation," by M. A. Tanner and W. H. Wong, *Journal of the American Statistical Association*, 82, 543–546.

——— (1988), "Using the SIR Algorithm to Simulate Posterior Distributions," in *Bayesian Statistics 3*, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, pp. 395–402.

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

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.

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—Theory and Methods*, 14, 1079–1102.

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