SAMPLING BASED APPROACHES TO CALCULATING MARGINAL DENSITIES

BY

ALAN E. GELFAND and ADRIAN F. M. SMITH

TECHNICAL REPORT NO. 415
APRIL 11, 1989

Prepared Under Contract
N00014-86-K-0156 (NR-042-267)
For the Office of Naval Research

Herbert Solomon, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

SAMPLING BASED APPROACHES TO CALCULATING MARGINAL DENSITIES

Alan E Gelfand

and

Adrian F M Smith

SUMMARY

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, and 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.

Keywords: marginal density; Monte Carlo sampling; stochastic substitution; Gibbs sampler; importance sampling; conditional probability structure; posterior distributions; data augmentation; hierarchical models; missing data; variance components.

1. Introduction

In relation to a collection of random variables, $U_1, U_2, \ldots, U_k$, suppose that either;

(i) for $i = 1, \ldots, k$, the conditional distributions $U_i | U_j, j \neq i$ are available, perhaps having, for some i, reduced forms $U_i | U_j, j \in S_i \subset \{1, \ldots, k\}$, or

(ii) 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 $U_i | U_j, j \neq i$ is available,

where *available* is here taken to mean that samples of U_i can be straightforwardly and efficiently generated, given specified values of the appropriate conditioning variables.

The problem addressed in this paper is the exploitation of the kind of structural information given by either (i) or (ii) in order to obtain numerical estimates of non-analytically 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. No claim will be made that the sampling methods to be described are necessarily computationally efficient compared with expert use of the latter. The attraction of the sampling based methods is instead their conceptual simplicity and ease of implementation for users with available computing resource but without numerical analytic expertise. All that the user requires is insight into the relevant conditional probability structure, together with techniques for the efficient generation of appropriate random variates as described, for example, 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 stochastic substitution 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 shall show (and as has also been observed by Clayton, 1988), there is a close relationship between the Gibbs sampler and the substitution algorithm proposed by Tanner and Wong (1987). We shall 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 (i) above. 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 generalised 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 (i).

Both the substitution and Gibbs sampler algorithms are iterative Monte Carlo procedures, applicable when the kind of structural information given by (i) above is available. When the structural information is of the kind described by (ii), we shall see that an importance sampling algorithm based on Rubin (1987, 1988) provides a non-iterative, 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.

2. Sampling approaches

In the sequel, we shall 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 (i) of Section 1) uniquely defines the full joint density: see, for example, Besag (1974). Throughout, we shall 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 will therefore be used interchangeably.

Densities will be denoted, generically, by square brackets, so that joint, conditional and marginal forms appear, for example, as $[X,Y]$, $[X|Y]$ and $[X]$. Multiplication of densities will be denoted by $*$, so that, for example,

$$[X,Y] = [X|Y] * [Y].$$

The process of marginalisation (i.e. integration) will be denoted by forms such as

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

with the convention that all variables appearing in the integrand but not in the resulting density have been integrated out. Thus, in the above, the integration is with respect to Z and W. More generally, we shall 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 algorithm

The substitution algorithm for finding fixed point solutions to certain classes of integral equations is a standard mathematical tool, which has received considerable attention in the literature: see, for example, Rall (1969). Its potential utility in statistical problems of the kind we are concerned with in this paper was recently observed by Tanner and Wong (1987) and associated discussion. Briefly reviewing the essence of their development using the notation introduced above, we have

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

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

so that, substituting (2) into (1), we obtain

$$[X] = \int [X|Y] * \int [Y|X'] * [X'] = \int h(X,X') * [X'], \tag{3}$$

where

$$h(X,X') = \int [X|Y] * [Y|X'],$$

with X' appearing as a "dummy argument" in (3) and, of course, $[X] \equiv [X']$. Now suppose that, on the right-hand 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 $[X]_{i+1}$,

$$\begin{aligned} [X]_{i+1} &= \int h(X,X') * [X']_i \\ &= I_h[X]_i, \end{aligned}$$

in a notation making explicit the fact that I_h is the integral operator associated with h. Exploiting standard theory of such integral operators, Tanner and Wong (1987) show 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 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, Z we may write, analogous to (1) and (2),

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

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

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

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, whence *TW*1, *TW*2 and *TW*3 will continue to hold in this extended setting. Extension to k variables is straightforward. A noteworthy by-product, using *TW*1, 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 $[X|Y]$ and $[Y|X]$ 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 $[Y|X]$ is available draw $Y^{(1)} \sim [Y|X^{(0)}]$, whence from (2) the marginal distribution of $Y^{(1)}$ is $[Y]_1 = \int [Y|X] * [X]_0$. Now complete a cycle by drawing $X^{(1)} \sim [X|Y^{(1)}]$. Using (1), we then have

$$X^{(1)} \sim [X]_1 = \int [X|Y] * [Y]_1 = \int h(x, x') * [x']_0 = I_h[X]_0.$$

Repetition of this cycle will produce $Y^{(2)}$ and $X^{(2)}$ and eventually, after i iterations, the pair $(X^{(i)}, Y^{(i)})$ such that

$$X^{(i)} \xrightarrow{d} X \sim [X], \qquad Y^{(i)} \xrightarrow{d} Y \sim [Y],$$

by virtue of *TW*2. Repetition of this sequence m times each to the ith iteration will generate m i.i.d. 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. Some practical experience with regard to the autocorrelation and cross correlation in very special cases of the sequence $\{(X_j^{(i)}, Y_j^{(i)}\}$, $i = 1, 2, \ldots$ is described in Clayton (1988).

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 Section 2.6) and in fact need not be generated unless $[Y]$ is to be estimated as well.

We note that this version of the substitution sampling algorithm differs slightly from the Imputation-Posterior (IP) algorithm in Tanner and Wong (1987). They propose, at each iteration l, $l = 1, 2, \ldots, i$, 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 simple random sampling of the $Y^{(l)}$ was introduced to allow m to vary across iterations but it seems unnecessary. Systematic sampling of the $X_j^{(l)}$ (as we propose for m constant) is simpler and, as the distribution theory above shows, gives a convergent procedure. Empirical investigation reveals little difference between the two modes of sampling the $Y_j^{(l)}$ with regard to the goodness of the resultant estimated marginal density at the ith iteration. The unnecessary resampling

implicit in Tanner and Wong has also been noted by Clayton (1988), and systematic selection of the $Y_j^{(i)}$ was also proposed by Morris (1987b) in his discussion of the Tanner and Wong paper.

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-hand side can be made arbitrarily small as $i \to \infty$ as a consequence of *TW2* above. The first term in the r.h.s. 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 $X^{(0)} \sim [X]_0$ then $(Z^{(0)'}, Y^{(0)'}) \sim [Z, Y \mid X^{(0)}]$, then $(Y^{(1)}, X^{(0)'}) \sim [Y, X \mid Z^{(0)'}]$ 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 will produce $(X^{(i)}, Y^{(i)}, Z^{(i)})$. The above 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 i.i.d. $(X_j^{(i)}, Y_j^{(i)}, Z_j^{(i)})$, $j = 1, \ldots, m$ (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 $[X, Z|Y]$, $[Y, X|Z]$ and $[Z, Y|Z]$. Of course, in many cases sampling from, say, $[X, Z|Y]$ requires, for example, $[X|Y, Z]$ and $[Y|Z]$, i.e. the availability of a full conditional and also a reduced conditional distribution. Paralleling (7), the density estimator of $[X]$ becomes

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

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 will require $k(k-1)$ random variate generations to complete a cycle. If we run m sequences out to the ith iteration (a total of $mik(k-1)$ random generations) we obtain m i.i.d. 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 we write (4)–(6) in the form

$$\begin{aligned} [X] &= \int [X|Z, Y] * [Z|Y] * [Y], \\ [Y] &= \int [Y|X, Z] * [X|Z] * [Z], \\ [Z] &= \int [Z|Y, X] * [Y|X] * [X]. \end{aligned} \tag{10}$$

Implementation of substitution sampling requires availability of all six conditional distributions on the r.h.s. of (10), rarely the case in practice. As noted at the beginning of Section 2, the full conditional distributions alone, $[X|Y, Z]$, $[Y|Z, X]$, $[Z|X, Y]$, will 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 in Geman and Geman (1984) and is known as the Gibbs sampler.

The Gibbs sampler was developed and has been mainly 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 upon some 'neighbourhood' subset of the variables (a reduced form, in the terminology of (i) in Section 1). More precisely, for the set of variables $U_1, U_2, \ldots, U_k$

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

where S_i is a 'small neighbourhood' subset of $\{1,2,\ldots,k\}$. A crucial question to ask is under what circumstances the specification (11) uniquely determines the joint distribution. The answer is taken up in great detail in Geman and Geman (1984), involving concepts such as graphs, neighbourhood systems, cliques, Markov Random Fields and Gibbs distributions. We refer the reader to that reference for details. In all the examples we shall consider, the joint distribution will be 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 which proceeds as follows. Given an arbitrary starting set of values $U_1^{(0)}, U_2^{(0)}, \ldots, U_k^{(0)}$, we draw $U_1^{(1)} \sim [U_1 | U_2^{(0)}, \ldots, U_k^{(0)}]$ then $U_2^{(1)} \sim [U_2 | U_1^{(1)}, U_3^{(0)}, \ldots, U_k^{(0)}]$, $U_3^{(1)} \sim [U_3 | U_1^{(1)} U_2^{(1)}, U_4^{(0)}, \ldots, U_k^{(0)}]$, ... and so on, up to $U_k^{(1)} \sim [U_k | U_1^{(1)}, \ldots, U_{k-1}^{(1)}]$. 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 $(U_1^{(i)}, \ldots, U_k^{(i)})$. Geman and Geman show, under mild conditions, that the following results hold.

GG1 (convergence) $U_s^{(i)} \xrightarrow{d} U_s \sim [U_s]$ 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 (i.o.).

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 i.o. visiting scheme.

GG3 (ergodic theorem) For any measurable function T of $U_1, \ldots, U_k$ whose expectation exists,

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

As in the previous section, Gibbs sampling through m replications of the above i iterations (a total of mik random variate generations) produces m i.i.d. k-tuples $(U_{1j}^{(i)}, \ldots, U_{kj}^{(i)})$, $j = 1, \ldots, m$, with the proposed density estimate for $[U_s]$ having exactly the form (9).

It seems sensible that we might utilize the $U_{sj}^{(l)}$, $l < i$ to improve upon (9). More precisely, defining $W_{sij} = [U_s | U_t = U_{tj}^{(i)}, t \neq s]$, (9) is $\sum_{j=1}^{m} W_{sij}/m$. Suppose we replace W_{sij} by $V_{sij} = \sum_{l=1}^{i} W_{sij}/i$ and consider

$$[\hat{U_s}]_i' = \frac{1}{m} \sum_{j=1}^{m} V_{sij}. \tag{*}$$

By $GG3$, as $i \to \infty$, $V_{sij} \xrightarrow{a.s.} [U_s]$, a constant; the convergence of (*) depends only on i not on m. By contrast, as $i \to \infty$, W_{sij} converges in distribution to a random variable. Upon averaging such random variables over m we obtain convergence to $[U_s]$. In terms of variability $\text{var}[\hat{U_s}]_i' < \text{var}[\hat{U_s}]_i$ since $\text{var}(V_{sij}) < \text{var}(W_{sij})$ whence (*) is better than (9). In practice (9) has performed so well that we haven't needed to resort to (*). Also in practice we might delete from V_{sij} several of the early, "far from converged", W_{slj}.

2.4 Relationship Between Gibbs Sampling and Substitution Sampling

It is apparent, as also noted by Clayton (1988), 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 uniquely determine the joint density). The substitution sampling algorithm requires the availability of a total of $k(k-1)$ conditional distributions including all 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 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, while Gibbs sampling only samples from the full conditional distributions.

To amplify upon this point, we describe how the substitution sampling algorithm might be carried out under availability of just the set of full conditional distributions. We shall see that it can be viewed as the Gibbs sampler, but under an i.o. visiting scheme different from the natural one. We present the argument in the three variable case for simplicity. Returning to (10), if $[Y|X]$ is unavailable we can create a sub-substitution loop to obtain it by means of:

$$\begin{aligned} [Y|X] &= \int [Y|X,Z] * [Z|X], \\ [Z|X] &= \int [Z|X,Y] * [Y|X]. \end{aligned} \tag{12}$$

Similar subloops are clearly available to create $[X|Z]$ and $[Z|Y]$. 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 above analysis suggests that we could in fact view the reduced conditional densities such as $[Y|X]$ as 'available', and that we could thus carry out the substitution algorithm as if all needed conditional distributions were available. In fact, $[Y|X]$, for example, is not 'available' in our earlier sense. Under the subloop in (12), we can always obtain a density estimate for $[Y|X]$ given any specified X, say $X^{(0)}$. However, at the next cycle of the iteration we would need a brand new density estimate for $[Y|X]$ 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 i.o.. Therefore, this version of the substitution sampling algorithm is in fact merely Gibbs sampling with a different but still i.o. visiting order. As a result, $GG1$, $GG2$ and $GG3$ still hold ($TW1$, $TW2$, $TW3$ 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 will occur when some reduced conditional distributions, distinct from the full conditional distributions, are available. Suppose 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 to 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 $X^{(0)}, Y^{(0)}, Z^{(0)}$ we start, for example, at the bottom line of (10), drawing

(i) $Y^{(0)'}$ from $[Y|X^{(0)}, Z^{(0)}]$,

(ii) $Z^{(0)'}$ from $[Z|Y^{(0)'}, X^{(0)}]$,

(iii) $X^{(0)'}$ from $[X|Z^{(0)'}, Y^{(0)'}]$,

(iv) $Y^{(1)}$ from $[Y|X^{(0)'}, Z^{(0)'}]$,

(v) $Z^{(1)}$ from $[Z|Y^{(1)}, X^{(0)'}]$,

(vi) $X^{(1)}$ from $[Z|Y^{(1)}, Z^{(1)}]$.

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, $[Z|Y]$, say, is available and is distinct from $[Z|X,Y]$. Following the substitution order, at step (v) we would instead draw $Z^{(1)}$ from the 'correct' distribution, $[Z|Y^{(1)}]$.

In Section 3, we provide classes of examples where distinct reduced conditional distributions will be available and classes where they generally will not. In Section 4, we present some preliminary computations which attempt to quantify the acceleration in convergence which arises from having available distributions additional to the full conditionals.

2.5 The Rubin Importance Sampling Algorithm

Rubin's comments (1987) to Tanner and Wong include the suggestion of a non-iterative Monte Carlo method for generating marginal distributions utilizing importance sampling ideas. We present the basic idea first in the two variable case. Suppose 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 $[X|Y]$ (a special case of the conditions described in (ii) of Section 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 which has positive support wherever $[Y]$ does and which has density $[Y]_s$, say. Then $[X|Y]*[Y]_s$ provides an importance sampling distribution for (X,Y). Suppose we draw i.i.d. pairs (X_l, Y_l), $l = 1,\ldots,N$ from this joint distribution; for example, by drawing Y_l from $[Y]_s$ and then X_l from $[X|Y_l]$. Rubin's idea is to calculate $r_l = [X_l, Y_l]/[X_l|Y_l]*[Y_l]_s$, $l = 1,\ldots,N$ and then estimate the marginal density for $[X]$ by

$$[\hat{X}] = \frac{\sum_{l=1}^{N} [X|Y_l] r_l}{\sum_{l=1}^{N} r_l}. \tag{13}$$

Note the important fact that $[X,Y]$ need only be specified up to a constant since the latter will cancel 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 top and bottom of (13) by N and using the Law of Large Numbers, we immediately have:

RI (convergence) $[\hat{X}] \rightarrow [X]$ w.p.1 as $N \rightarrow \infty$ for almost every X.

If, additionally, $[Y|X]$ is available we immediately have an estimate for the marginal distribution of Y:

$$[\hat{Y}] = \frac{\sum_{l=1}^{N} [Y|X_l] r_l}{\sum_{l=1}^{N} r_l}.$$

The successful performance of (13) will typically depend strongly upon the choice of $[Y]_s$ and its closeness to $[Y]$. Thus the suggestion of Tanner and Wong in their rejoinder to Rubin's discussion of their paper, 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 non-iterative 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 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 $[X|Y,Z]$. In this case, the pair (Y,Z) plays the role of Y in the two variable case discussed above and, in general, we need to specify an importance sampling distribution $[Y,Z]_s$. However, if, for example, $[Y|Z]$ is available we will only need to specify $[Z]_s$. In any case, we draw i.i.d. triples (X_l, Y_l, Z_l), $l = 1,\ldots,N$ and calculate

$$r_l = \frac{[X_l, Y_l, Z_l]}{[X_l|Y_l, Z_l] * [Y_l, Z_l]_s}.$$

The marginal density estimate for $[X]$ then becomes, analogous to (13),

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

We note that in the k-variable case the Rubin importance sampling algorithm requires a total of Nk random variate generations, while Gibbs sampling stopped at iteration i will require 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 which places mass $r_l/\Sigma r_l$ at Y_l, $l = 1,\ldots,N$. We could then replace (13) by

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

so that (7) and (15) are of the same form. Relative performance on average depends upon whether the distribution of $Y^{(i)}$ or of 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 which underlies much modern work in adaptive Monte Carlo: see, for example, Smith *et al*, 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 utilizing this full conditional distribution. In the simplest case of two variables, this implies that $[X|Y]$ 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 upon these $X_j^{(i)}$'s. The formal argument is essentially that of Rao-Blackwellization, for which we shall 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 upon the $X_j^{(i)}$ (see, for example, Devroye and Györfi, 1985) evaluated at say X_0:

$$\Delta_{X_0}^{(i)} = \frac{1}{mh_m^p} \sum_{j=1}^{m} K\left(\frac{X_0 - X_j^{(i)}}{h_m}\right),$$

where K is a bounded density on R^p and the sequence $\{h_m\}$ is such that as $m \to \infty$, $h_m \to 0$ while $mh_m^p \to \infty$. To simplify notation, set

$$Q_{m,X_0}(X) = \frac{1}{h_m^p} K\left(\frac{X_0 - X}{h_m}\right)$$

so that $\Delta_{X_0}^{(i)} = \frac{1}{m} \sum_{j=1}^{m} Q_{m,X_0}(X_j^{(i)})$. Define

$$\gamma_{X_0}^{(i)} = \frac{1}{m} \sum_{j=1}^{m} E(Q_{m,X_0}(X) | Y_j^{(i)}).$$

By our earlier theory, both $\Delta_{X_0}^{(i)}$ and $\gamma_{X_0}^{(i)}$ have the same expectation. By the Rao-Blackwell theorem,

$$\operatorname{var} E(Q_{m,X_0}(X|Y)) \leq \operatorname{var} Q_{m,X_0}(X),$$

whence

$$MSE(\gamma_{X_0}^{(i)}) \leq MSE(\Delta_{X_0}^{(i)}),$$

where MSE denotes mean square error of the estimate of $[X_0]$.

Now for fixed Y, as $m \to \infty$, $E(Q_{m,X_0}(X|Y)) \to [X_0|Y]$ for almost every X_0 by the Lebesgue Density Theorem (see Devroye and Györfi, p.3). Thus in terms of random variables we have $E(Q_{m,X_0}(X|Y)) \xrightarrow{d} [X_0|Y]$, so that, for large m, $\gamma_{X_0}^{(i)} \doteq [\hat{X}_0]_i$, $MSE(\gamma_{X_0}^{(i)}) \doteq MSE([\hat{X}_0]_i)$, whence $[\hat{X}_0]_i$ is preferred to $\Delta_{X_0}^{(i)}$.

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

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 a number of advances in numerical and analytic approximation techniques for such calculations—see, for example, Naylor and Smith (1982, 1988), Smith *et al* (1985, 1987), Tierney and Kadane (1986), Shaw (1988), Geweke (1988)—but implementation of these approaches typically requires sophisticated numerical analytic expertise and possibly specialist software. In stark contrast, the three sampling approaches we have discussed are essentially trivial to implement and, for many practitioners, this feature will more than compensate for any relative computational inefficiency. To provide a flavour of the kinds of area of application for which the methodology is suited, we present six examples of typical probability structures that arise.

3.1 A Class of Multinomial Models

We extend the one parameter genetic linkage example described in 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 Hartley, 1958; Dempster, Laird and Rubin, 1977). 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

$$\text{Mult}(n, a_1\theta + b_1, a_2\theta + b_2, a_3\eta + b_3, a_4\eta + b_4, c(1 - \theta - \eta)),$$

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, η range over $\theta \geq 0$, $\eta \geq 0$ and $\theta + \eta \leq 1$, so that a three parameter Dirichlet distribution, Dirichlet$(\alpha_1, \alpha_2, \alpha_3)$ may be a natural choice of prior density for (θ, η). From the form of $[Y|\theta,\eta] * [\theta,\eta]$ the reader will note that obtaining the exact marginals $[\theta|Y], [\eta|Y]$ will be somewhat messy (involving a two-dimensional numerical integral). However, 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

$$\text{Mult}(n, a_1\theta, b_1, a_2\theta, b_2, a_3\eta, b_3, a_4\eta, b_4, c(1-\theta-\eta)).$$

From the form of $[X|\theta,\eta] * [\theta,\eta]$ we see that

$$[\theta,\eta|X] \sim \text{Dirichlet}(X_1+X_3+\alpha_1, X_5+X_7+\alpha_2, X_9+\alpha_3),$$

whence $[\theta|X], [\eta|X]$ are available as Beta distributions for sampling. Furthermore, $[\theta|X,\eta]$ and $[\eta|X,\theta]$ are available as scaled Beta distributions, scaled respectively to the intervals $[0, 1-\eta]$ and $[0, 1-\theta]$. If we let $Y_1 = X_1 + X_2$, $Y_2 = X_3 + X_4$, $Y_3 = X_5 + X_6$, $Y_4 = X_7 + X_8$, $Y_5 = X_9$ and define $Z = (X_1, X_3, X_5, X_7)$, we see that specification of X is equivalent to specification of (Y, Z). Also, $[Z|Y,\theta,\eta]$ is the product of four independent Binomials for X_1, X_3, X_5, X_7, given by

$$[X_i|Y,\theta,\eta] = Bi\left(Y_i, \frac{a_i\theta}{(a_i\theta + b_i)}\right), \quad i = 1,3,5,7,$$

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 $[\theta|Y], [\eta|Y], [Z|Y]$. Gibbs sampling requires $[\theta|Y,Z,\eta]$, $[\eta|Y,Z,\theta]$ and $[Z|Y,\theta,\eta]$, all of which are available. But in this case the reduced distributions $[\theta|Y,Z]$ and $[\eta|Y,Z]$ are also available and this allows us to accelerate the substitution sampling algorithm. These reduced distributions also substantially simplify the Rubin importance sampling algorithm in obtaining $[\theta|Y]$ and $[\eta|Y]$; only an importance sampling distribution $[Z|Y]$, need be specified (for example, a 'default' choice might be binomials with chance equal to one half). Detailed comparison of the performance of the three algorithm for a specific case of this multinomial class will be 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

$$[Y|\theta_1] * [\theta_1|\theta_2] * [\theta_2|\theta_3] \ldots * [\theta_{k-1}|\theta_k] * [\theta_k], \tag{16}$$

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

We note in passing that the collection of variables $Y, \theta_1, \ldots, \theta_k$ form a random Markov field with an 'adjacent' neighbourhood system and that the joint distribution of the variables is a Gibbs distribution. See Geman and Geman (1984) for definitions and other examples.

The hierarchical structure implies that

$$[\theta_i|Y, \theta_j, j \neq i] = \begin{cases} [\theta_1|Y,\theta_2] & i = 1, \\ [\theta_i|\theta_{i-1},\theta_{i+1}] & 1 < i < k-1, \\ [\theta_k|\theta_{k-1}] & i = k. \end{cases} \tag{17}$$

Suppose 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. However, conjugate priors can generally be made arbitrarily diffuse by appropriate choices of hyperparameters and so this case is also implicitly subsumed within the conjugate framework. $[\theta_k]$ can, in fact, be vague provided $[\theta_k|\theta_{k-1}]$ is still proper and available. (See examples 3.4, 3.5.) Conjugacy implies that the densities in (17) will be 'available' as 'updated' versions of the respective priors (see e.g. Morris, 1983a). Typically, no distinct

reduced conditional distributions will be 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 $[\theta_1|y,\theta_2,\theta_3]$, $[\theta_2|y,\theta_1,\theta_3]$, $[\theta_3|y,\theta_1,\theta_2]$, $[\theta_3|y,\theta_2]$, $[\theta_1|y,\theta_3]$ and $[\theta_2|y,\theta_1]$. The first three are available as in (17), the fourth is available but is not distinct from the third and the last two are usually unavailable.

As a concrete illustration, consider an exchangeable Poisson model, which will be further illustrated in Section 4 with the reanalysis of a published data set. Suppose 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 $[s_i|\lambda_i] = P_0(\lambda_i t_i)$ and that the λ_i are i.i.d. from $G(\alpha,\beta)$, with density $\lambda_i^{\alpha-1}e^{-\lambda_i/\beta}/\beta^{\alpha}\Gamma(\alpha)$. The parameter α will be 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, in turn, is assumed to arise from an Inverse Gamma distribution $IG(\gamma,\delta)$ 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 zero.)

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

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

while the full posterior for β is given by

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

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 $\lambda_j^{(1)} \sim G\left(\alpha+s_j,\left(t_j+\frac{1}{\beta^{(0)}}\right)^{-1}\right)$, $j = 1,\ldots,p$, and then $\beta^{(1)} \sim IG\left(\gamma+\alpha p,\sum_{j=1}^{p}\lambda_j^{(1)}+\delta\right)$ to complete one cycle. If we carry out m repetitions each of i iterations, generating $(\lambda_1^{(i)},\ldots,\lambda_p^{(i)},\beta_l^{(i)})$, $l = 1,\ldots,m$, the marginal density estimate for λ_j is

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

while

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

Rubin's importance sampling algorithm is also applicable in the setting (16), taking a particularly simple form in the cases $k = 2,3$. For $k = 3$, suppose we seek $[\theta_1|y]$. The joint density $[\theta_1,\theta_2,\theta_3|Y] = [Y,\theta_1,\theta_2,\theta_3]/[Y]$, where the functional form of the numerator is given in (16). An importance sampling density for $[\theta_1,\theta_2,\theta_3|Y]$ could be sampled as $[\theta_1|Y,\theta_2]*[\theta_3|\theta_2]*[\theta_2|Y]_s$, for some $[\theta_2|Y]_s$. As remarked in Section 2.5, a 'good' choice for $[\theta_2|Y]_s$ might possibly 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 $[\theta_2|Y]_s$, θ_{3l} from $[\theta_3|\theta_{2l}]$ and θ_{1l} from $[\theta_1|Y,\theta_{2l}]$. Calculating

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

we obtain the density estimator

$$\hat{[\theta_1|Y]} = \frac{\Sigma[\theta_1|Y,\theta_{2l}]\cdot r_l}{\Sigma r_l}.$$

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 $[\theta_{1l}|Y,\theta_{2l}]*[Y|\theta_{2l}]*[\theta_{2l}|\theta_{3l}]*[\theta_{3l}]$ and noting that r_l does not involve θ_{1l}, so that 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

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

Here

$$r_l = [Y|\beta_l] * [\beta_l]/[\beta_l|Y]_s,$$

where $[Y|\beta_l]$ is the product of Negative Binomial densities, i.e.

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

while $[\beta_l]$ is the inverse gamma prior evaluated at β_l. If $[\beta|Y]_s$ is not obtained from the substitution sampling algorithm, as in (21), an alternative choice is $IG\left(\gamma+\alpha p, \sum_{i=1}^{p} \rho_i + 1\right)$. This arises since,

$$[\beta|Y] = E_{[\lambda_1,\ldots,\lambda_p|Y]}[\beta|Y,\lambda_1,\ldots,\lambda_p] \approx [\beta|Y,\hat{\lambda}_1,\ldots,\hat{\lambda}_p]$$

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: see, for example, Dempster, Laird and Rubin (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 will be available. We illustrate in the simplest case, where we assume

$$\binom{U_{1i}}{U_{2i}}, i = 1,\ldots,n_1, \quad \binom{V_{1j}}{V_{2j}}, j = 1,\ldots,n_2, \quad \binom{W_{1k}}{W_{2k}}, k = 1,\ldots,n_3$$

are all i.i.d. $N(\theta,\Delta)$ 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

$$U = \binom{U_1}{U_2} = \begin{pmatrix} U_{11}\ldots U_{1n_1} \\ U_{21}\ldots U_{2n_1} \end{pmatrix}$$

with similar notation for V and W. Finally, let $X = (U,V,W)$, $2\times N$, with $\bar{X} = N^{-1}X1$, where 1 is a column vector of N 1's and $N = n_1+n_2+n_3$. Standard calculations show that $[\theta|X]$ is $N(\eta,\Omega)$, where

$$\eta = (N\Delta^{-1}+\Sigma^{-1})^{-1}(N\Delta^{-1}\bar{X}+\Sigma^{-1}\mu)$$

and

$$\Omega = (N\Delta^{-1}+\Sigma^{-1})^{-1}.$$

With the obvious partitioning,

$$\eta = \binom{\eta_1}{\eta_2} \qquad \Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix},$$

the marginals $[\theta_1|X] = N(\eta_1,\Omega_{11})$ and $[\theta_2|X] = N(\eta_2,\Omega_{22})$ are available. Suppose, however, that V_2, W_1, say, are unobserved. Let $Y = (U,V_1,W_2)$, $Z = (V_2,W_1)$ so that $X \equiv (Y,Z)$. As in Section 3.1, we have a 'three variable' problem, here involving θ_1,θ_2,Z. The full conditional distributions are all normal and hence available. For θ_1 and θ_2,

$$[\theta_1|Y,Z,\theta_2] = N(\eta_1+\Omega_{12}\Omega_{22}^{-1}(\theta_2-\eta_2), \Omega_{11}-\Omega_{12}\Omega_{22}^{-1}\Omega_{21})$$

$$[\theta_2|Y,Z,\theta_1] = N(\eta_2+\Omega_{21}\Omega_{11}^{-1}(\theta_1-\eta_1), \Omega_{22}-\Omega_{21}\Omega_{11}^{-1}\Omega_{12}).$$

Letting $\bar{U}_1 = n_1^{-1}U_1 1$, with similar notation for $\bar{U}_2,\bar{V}_1,\bar{V}_2,\bar{W}_1,\bar{W}_2$, we note, by sufficiency, that with regard

to Z we only need the full posterior $[\bar{Z}|\bar{Y},\theta_1,\theta_2]$, where $\bar{Z}^\tau = (\bar{V}_2,\bar{W}_1)$, $\bar{Y}^\tau = (\bar{U}_1,\bar{U}_2,\bar{V}_1,\bar{W}_2)$. Since

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

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

The Rubin importance sampling algorithm is also straightforward in this case. Simplifying notation by working with the sufficient statistic $(\bar{Y},\bar{Z})$, suppose, for instance, we seek the density estimator of $[\theta_1|Y]$. We have

$$\hat{[\theta_1|Y]} = \Sigma[\theta_1|\bar{Y},\bar{Z}_l,\theta_{2l}]r_l/\Sigma r_l,$$

where

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

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

3.4 Variance component models

Bayesian inference for variance components has typically required subtle numerical analysis or intricate analytic approximation, as evidenced, for example, in Box and Tiao (1973, Chapters 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

$$Y_{ij} = \theta_i + \varepsilon_{ij}, \quad i = 1,\ldots,K, \quad j = 1,\ldots,J,$$

where, assuming conditional independence throughout, $[\theta_i|\mu,\sigma_\theta^2] = N(\mu,\sigma_\theta^2)$ and $[\varepsilon_{ij}|\sigma_\varepsilon^2] = N(0,\sigma_\varepsilon^2)$, so that $[Y_{ij}|\theta_i,\sigma_\varepsilon^2] = N(\theta_i,\sigma_\varepsilon^2)$.

Let $\theta = (\theta_1,\ldots,\theta_K)$, $Y = (Y_{11},\ldots,Y_{KJ})$ and assume that $\mu,\sigma_\theta^2,\sigma_\varepsilon^2$ are independent, with priors specified by $[\mu] \sim N(\mu_0,\sigma_0^2)$, $[\sigma_\theta^2] \sim IG(a_1,b_1)$ and $[\sigma_\varepsilon^2] \sim IG[a_2,b_2]$, where IG denotes the inverse gamma distribution (as in Example 3.2) and $\mu_0,\sigma_0^2,a_1,b_1,a_2,b_2$ are assumed known (possibly chosen to correspond to diffuse priors).

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

$$[Y|\theta,\sigma_\varepsilon^2] * [\theta|\mu,\sigma_\theta^2] * [\mu] * [\sigma_\theta^2] * [\sigma_\varepsilon^2] \tag{22}$$

and we shall follow Box and Tiao (1973, Chapter 5) in focussing interest on $[\sigma_\theta^2|Y]$ and $[\sigma_\varepsilon^2|Y]$.

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

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

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

$$[\mu|Y,\theta,\sigma_\theta^2,\sigma_\epsilon^2] = [\mu|\sigma_\theta^2,\theta] = N\left(\frac{\sigma_\theta^2\mu_0+\sigma_0^2\Sigma\theta_i}{\sigma_\theta^2+K\sigma_0^2}, \frac{\sigma_\theta^2\sigma_0^2}{\sigma_\theta^2+K\sigma_0^2}\right)$$

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

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

Since all 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 will continue to be normal, and those for the variance components will continue to be inverse gamma.

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 which has proved difficult to handle using empirical Bayes approaches to 'estimating' posterior distributions (see, for example, Morris, 1983b, 1987a). Such models are straightforwardly handled by iterative sampling approaches, as we saw with the Poisson example of Section 3.2 and will further illustrate here for this classical normal means example.

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

$$[Y|\theta,\sigma^2] * [\theta|\mu,\tau^2] * [\sigma^2] * [\mu] * [\tau^2], \tag{23}$$

where

$$[Y|\theta,\sigma^2] * [\theta|\mu,\tau^2] * [\sigma^2] = \prod_{i=1}^{I} [Y_i|\theta_i,\sigma_i^2] * [S_i^2|\sigma_i^2] * [\theta_i|\mu,\tau^2] * [\sigma_i^2].$$

There is, of course, an obvious similarity between (22) and (23), but here interest is taken to focus on the $[\theta_i|Y]$, $i = 1,\ldots,I$. 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

$$[\theta|Y,\sigma^2,\mu,\tau^2] = N(\theta^*,D^*), \tag{24}$$

where

$$\theta_i^* = \frac{J_i\bar{Y}_{i\cdot}\tau^2+\mu\sigma_i^2}{J_i\tau^2+\sigma_i^2},$$

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

Thus the full conditional distributions $[\theta_i|Y,\theta_j,j\neq i,\sigma^2,\mu,\tau^2]$, $i = 1,\ldots,I$, are, in fact, just the normal marginals of (24) and are therefore available for sampling. From (23), we easily see that

$$[\sigma^2|Y,\theta,\mu,\tau^2] = [\sigma^2|Y,\theta] = \prod_{i=1}^{I} [\sigma_i^2|\bar{Y}_{i\cdot},S_i^2,\theta_i],$$

where

$$[\sigma_i^2 | \bar{Y}_{i.}, S_i^2, \theta_i] = IG(a_1 + \tfrac{1}{2}J_i, b_1 + \tfrac{1}{2}\Sigma_j(Y_{ij} - \theta_i)^2).$$

Finally, and closely resembling the forms obtained in Section 3.4,

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

and

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

3.6 An errors-in-variable model

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

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

Responses are obtained at specified levels, X_0, of the covariate but suppose, in fact, 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 centres on $\theta = (\theta_1, \theta_2)$ and to complete the distributional specification suppose 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

$$[Y | X_a, \theta, \sigma^2] * [X_a | \tau^2] * [\tau^2] * [\theta] * [\sigma^2], \tag{25}$$

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

4. Numerical Illustrations

4.1 A multinomial model

We shall 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

$$\text{Mult}(22, \tfrac{1}{4}\theta + \tfrac{1}{8}, \tfrac{1}{4}\theta, \tfrac{1}{4}\eta, \tfrac{1}{4}\eta + \tfrac{3}{8}, \tfrac{1}{2}(1 - \theta - \eta)),$$

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 $a_1 = \frac{1}{4}$, $b_1 = \frac{1}{8}$, $a_2 = \frac{1}{4}$, $b_2 = 0$, $a_3 = \frac{1}{4}$, $b_3 = 0$, $a_4 = \frac{1}{4}$, $b_4 = \frac{3}{8}$, $\alpha_1 = \alpha_2 = \alpha_3 = 1$, with interest centring on the calculation of the marginal posterior densities $[\theta | Y]$, $[\eta | Y]$.

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

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

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 $Z = (X_1, X_5)$.

The Gibbs sampler, based on the full conditional distributions, iterates around the cycle:

set arbitrary $\theta^{(0)}, \eta^{(0)}$;

draw $Z^{(1)}$ from $[Z|\theta^{(0)},\eta^{(0)},Y]$, which is given by the product of two independent binomial distributions, $X_1^{(1)} \sim \text{Bi}(Y_1, 2\theta^{(0)}(1+2\theta^{(0)})^{-1})$, $X_5^{(1)} \sim \text{Bi}(Y_4, 2\eta^{(0)}(3+2\eta^{(0)})^{-1})$;

draw $\theta^{(1)}$ from $[\theta|\eta^{(0)},Z^{(1)},Y]$, which is available since $\theta/(1-\eta^{(0)})$ has a Beta$(X_1^{(1)}+Y_2+1, Y_5+1)$ distribution;

draw $\eta^{(1)}$ from $[\eta|\theta^{(1)},Z^{(1)},Y]$, where $\eta/(1-\theta^{(1)})$ has a Beta$(Y_3+X_5^{(1)}+1, Y_5+1)$ distribution;

reinitialize the cycle with $\theta^{(1)}, \eta^{(1)}$ and iterate, replicating each cycle m times.

The substitution sampler makes use of the fact that $[\theta|Y,Z,\eta]$ and $[\eta|Y,Z,\theta]$ can be replaced in this case by the reduced forms $[\theta|Y,Z]$, leading to iteration around the cycle:

set arbitrary $\theta^{(0)}, \eta^{(0)}$;

draw $Z^{(1)}$ from $[Z|\theta^{(0)},\eta^{(0)},Y]$, a product of two independent binomial distributions, $X_1^{(1)} \sim \text{Bi}(Y_1, 2\theta^{(0)}(1+2\theta^{(0)})^{-1})$, $X_5^{(1)} \sim \text{Bi}(Y_4, 2\eta^{(0)}(3+2\eta^{(0)})^{-1})$;

draw $\eta^{(1)}$ from $[\eta|\theta^{(0)},Z^{(1)},Y]$, where $\eta/(1-\theta^{(0)})$ has a Beta$(Y_3+X_5^{(1)}+1, Y_5+1)$ distribution;

draw $\theta^{(1)}$ from $[\theta|\eta^{(1)},Z^{(1)},Y]$, where $\theta/(1-\eta^{(1)})$ has a Beta$(X_1^{(1)}+Y_2+1, Y_5+1)$ distribution;

draw $Z^{(2)}$ from $[Z|\theta^{(1)},\eta^{(1)},Y]$, a product of binomials analogous to the above;

draw $\eta^{(2)}$ from $[\eta|Z^{(2)},Y]$, where η has a Beta$(Y_3+X_5^{(2)}+1, X_1^{(2)}+Y_2+Y_5+2]$ distribution;

draw $\theta^{(2)}$ from $[\theta|\eta^{(2)},Z^{(2)},Y]$, where $\theta/(1-\eta^{(2)})$ has a Beta$(X_1^{(2)}+Y_2+1, Y_5+1)$ distribution;

reinitialize the cycle with $\theta^{(2)}, \eta^{(2)}$ and iterate, replicating each cycle m times.

To compare the two forms of iterative sampling, we first obtained very accurate numerical estimates of $[\theta|Y]$, $[\eta|Y]$ using techniques described in Smith *et al* (1985, 1987) and from these then obtained the 'true' 5, 25, 50, 75 and 95 posterior percentile points for each parameter. Iterative cycles of the two samples were then run, calibrated so that the total number of random variates generated was the same in both cases, as described in Section 2.4. The initialization was defined (for an arbitrary generating seed) in each case by taking independent samples from $\theta \sim U(0,1)$, $\eta \sim U(0,1)$, subject to $0 \leq \theta+\eta \leq 1$. At each cycle, $m = 10$ drawings of the parameters were then made and estimates of the cumulative posterior probabilities corresponding to each of the five true percentile points for each parameter were obtained. This process was replicated 5000 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

	cdf value	*estimate (sd)*				*S closer than G*	
		θ		η		θ	η
		G	*S*	*G*	*S*		
	.05	.231(.08)	.217(.08)	.033(.01)	.044(.01)	56%	75%
	.25	.504(.10)	.492(.09)	.177(.04)	.225(.04)	55%	78%
cycle 1	.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%
	.05	.067(.04)	.055(.03)	.047(.01)	.048(.01)	56%	51%
	.25	.286(.07)	.266(.07)	.236(.04)	.241(.04)	56%	52%
cycle 2	.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%
	.05	.052(.03)	.049(.03)	.049(.01)	.049(.01)	51%	50%
	.25	.254(.06)	.252(.06)	.247(.04)	.247(.04)	51%	50%
cycle 3	.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%
	.05	.050(.03)	.047(.03)	.050(.01)	.050(.01)	51%	51%
	.25	.250(.06)	.249(.06)	.250(.04)	.249(.04)	50%	51%
cycle 4	.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%

We note from Table 1 that, initially (cycles 1 and 2) the substitution sampler adapts more quickly than the Gibbs sampler, particularly for η. However, by the time we reach the 3rd and 4th cycles, the two approaches are performing indistinguishably. What is astonishing, perhaps, is just how remarkably good their performance is. By the 4th cycle, using only $m = 10$ drawings and starting from a default non-informative 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 Section 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 non-iterative Rubin importance sampling algorithm, Section 2.5, requires us to choose a sampling density, $[Z|Y]_s$, and then to proceed as follows, for $l = 1,\ldots,m$:

draw Z_l from $[Z|Y]_s$, η_l from $[\eta|Z,Y]$, θ_l from $[\theta|\eta,Z,Y]$, with the latter two distributions as detailed above, thus creating a triple (θ_l,η_l,Z_l);

calculate

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

form estimates,

$$[\theta|Y] = \frac{\sum_{l=1}^{m} [\theta|\eta_l, Z_l, Y] r_l}{\sum_{l=1}^{m} r_l},$$

$$[\eta|Y] = \frac{\sum_{l=1}^{m} [\eta|\theta_l, Z_l, Y] r_l}{\sum_{l=1}^{m} r_l}.$$

Table 2 shows the average cumulative posterior probability estimates from this approach based on 2500 replicates of $m = 40$ and $m = 200$, taking $[Z|Y]_s$ to be the product of $X_1 \sim \text{Bi}(Y_1, \frac{1}{2})$ and $X_5 \sim \text{Bi}(Y_4, \frac{1}{2})$.

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)

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 $[Z|Y]_s$ and that the larger one-off simulation is no match for iterative adaptation via small simulations.

4.2 A conjugate hierarchical model

We shall apply the exchangeable Poisson model, discussed in Section 3.2, to data on pump failures, previously analysed by Gaver and O'Muircheartaigh (1987), and 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.

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

Recalling the model structure of Section 3.2 and the forms of conditional distribution given by (18) and (19), we shall 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,

$$\alpha = \frac{\bar{\rho}^2}{\left(S_\rho^2 - p\bar{\rho}\sum_{i=1}^{p} t_i^{-1}\right)},$$

the latter derived by a method-of-moments empirical Bayes argument based on

$$E(\rho_i) = EE(\rho_i|\lambda_i) = \frac{\alpha}{\beta} = \bar{\rho},$$

$$V(\rho_i) = VE(\rho_i|\lambda) + EV(\rho_i|\lambda_i) = \frac{\alpha}{\beta^2} + \frac{\alpha}{\beta t_i} = S_\rho^2 = p^{-1}\Sigma(\rho_i - \bar{\rho})^2.$$

The cycle is defined as follows:

draw initial $\beta^{(0)}$ from $[\beta]$, where $\beta \sim IG(\gamma, \delta)$;

draw independent $\lambda_j^{(1)}$ from $[\lambda_j|Y, \beta^{(0)}, \lambda_j, j \neq i] = [\lambda_j|Y, \beta^{(0)}]$, which is a $G(\alpha + s_j, (t_j + 1/\beta^{(0)})^{-1})$ distribution, $j = 1,\ldots,p$;

draw $\beta^{(1)}$ from $[\beta|Y, \lambda_1^{(1)},\ldots,\lambda_p^{(1)}]$, which is an $IG(\gamma + \alpha p, \delta + \Sigma\lambda_i^{(1)})$ distribution;

reinitialize the cycle with $\beta^{(1)}$ and iterate, replicating each cycle m times.

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 rather remarkable convergence from such a small number of drawings.

Figure 1

5. Discussion

The emphasis in this paper has been on 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 all share the characteristic of being straightforward to implement in a number of frequently occurring practical situations, thus avoiding complicated numerical or analytic approximation exercises (often necessitating intricate attention to reparametrisation 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 4, the methods find immediate and powerful application to problems involving ordered random variables, or involving change-points. We shall provide detailed and extensive numerical illustration of a number of such problems in a subsequent applications paper.

The preliminary computational experience reported here serves to illustrate 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, non-iterative, 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 show markedly 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 rule procedures for iterative sampling in large-scale complex problems; we shall report on these in the applications paper referred to earlier.

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 sophisticated numerical or analytic approximation techniques.

Acknowledgements

This work was carried out by the first author with the partial support of the U.S. Office of Naval Research and under the auspices of the UK SERC Complex Stochastic Systems Initiative. The second author's collaboration with Dr Amy Racine, Mathematical Applications, CIBA-GEIGY, Basel, provided the initial stimulation for the work reported here.

References

Besag J (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society B*, **36**, 192–326.

Box G E P and Tiao G C (1973). *Bayesian inference in statistical analysis*. Addison-Wesley, Reading, MA.

Clayton D G (1988). Simulation in Hierarchical Models. Technical Report, University of Leicester.

Dempster A, Laird N and Rubin D B (1977). Maximum likelihood from incomplete data (with discussion). *Journal of the Royal Statistical Society B*, **39**, 1–38.

Devroye L (1986). *Non-uniform random variate generation*. Springer-Verlag, NY.

Devroye L and Györfi L (1985). *Non-parametric density estimation; the L_1 view*. J Wiley and Sons, NY.

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.

Morris C (1983a). Natural exponential families with quadratic vauance functions: statistical theory. *Annals of Statistics*, **11**, 515–529.

Morris C (1983b). Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association*, **78**, 47–59.

Morris C (1987a). Determining the accuracy of Bayesian empirical Bayes estimates in familiar exponential families, Technical Report 46, University of Texas, Austin.

Morris C (1987b). Comment 'Simulation in hierarchical models'. *Journal of the American Statistical Association*, **82**, 542–43.

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.

Naylor J C and Smith A F M (1988). Econometric illustrations of novel numerical integration strategies for Bayesian inferences. *J of Econometrics*, **38**, 103–126.

Rall L (1969). *Computational solution of non-linear operator equations*. John Wiley and Sons, NY.

Ripley B (1987). *Stochastic simulation*. John Wiley and Sons, NY.

Rubin (1987). Comment 'A non-iterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: the SIR algorithm'. *Journal of the American Statistical Association*, **82**, 543–546.

Rubin (1988). Using the SIR algorithm to simulate posterior distributions. *Bayesian Statistics 3*, J M Bernardo *et al*, editors (to appear).

Shaw J E H (1988). *Numerical integration and display methods for Bayesian inference*. PhD Thesis, Department of Mathematics, University of Nottingham.

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.

Smith A F M, Skene A M, Shaw J E H, Naylor J C (1987). Progress with numerical and graphical methods for Bayesian statistics. *Statistician*, **36**, 75–82.

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.

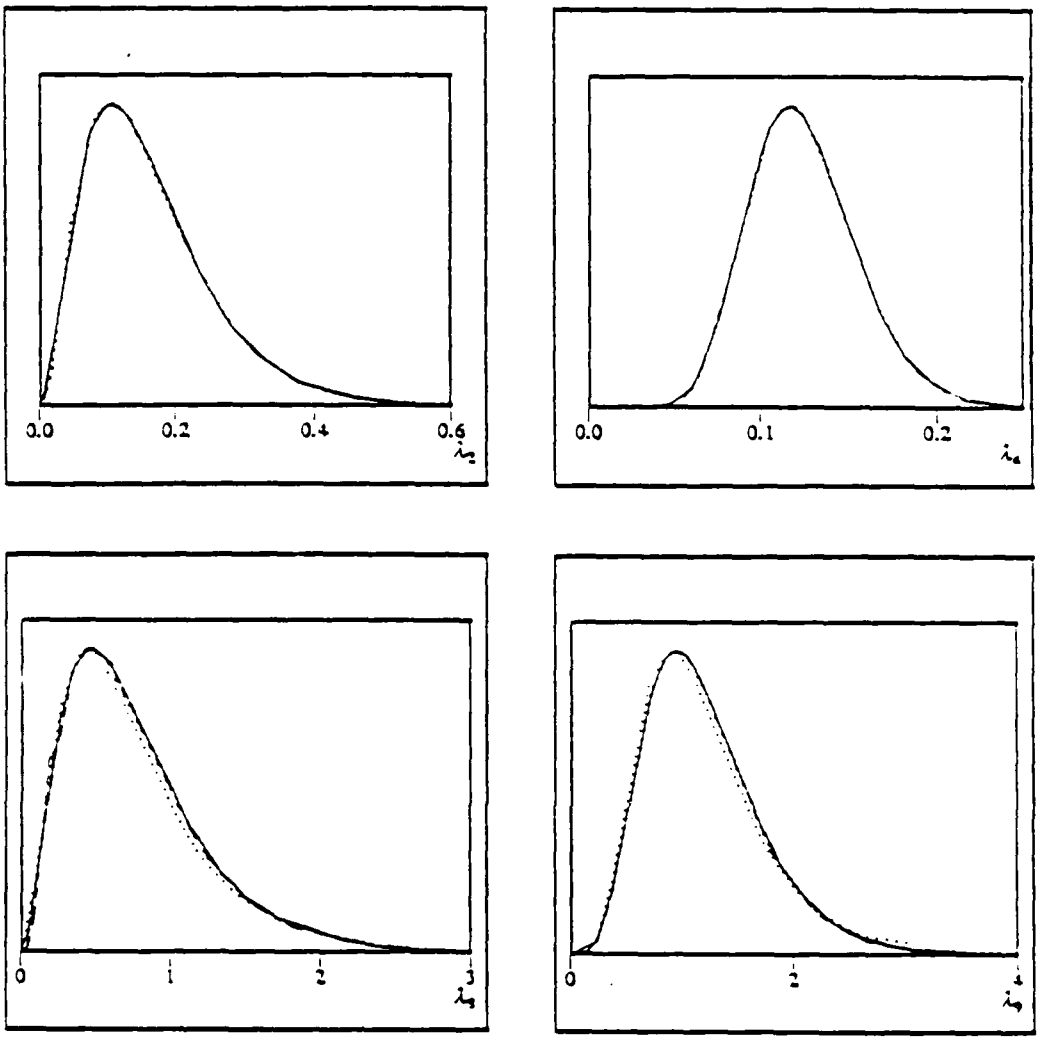

Figure 1: Density Estimates for Pump Failure Data, ... is m = 10, —— is m = 100, —— is 'exact'.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 415	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Sampling Based Approaches To Calculating Marginal Densities		5. TYPE OF REPORT & PERIOD COVERED TECHNICAL REPORT 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Alan E. Gelfand and Adrian F. M. Smith		8. CONTRACT OR GRANT NUMBER(s) N00014-86-K-0156
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Stanford University Stanford, CA 94305		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR-042-267
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Statistics & Probability Program Code 1111		12. REPORT DATE April 11, 1989 13. NUMBER OF PAGES 25
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

marginal density; Monte Carlo sampling; stochastic substitution; Gibbs sampler; importance sampling; conditional probability structure; posterior distributions; data augmentation; hierarchical models; missing data; variance components.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

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, and 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.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)