Statistical Science
1996, Vol. 11, No. 1, 54–64

Comments on "Methods for Approximating Integrals in Statistics with Special Emphasis on Bayesian Integration Problems" by Michael Evans and Tim Swartz

J. E. H. Shaw, Alan Genz, John Monahan, Mark J. Schervish, Larry Wasserman and Russ Wolfinger

Abstract. The original paper by Michael Evans and Tim Swartz appeared in the August 1995 issue of *Statistical Science* (**10** 254–272). Due to a publication error the following discussion of the paper was not included in that issue.

Comment

J. E. H. Shaw

I am very grateful to the authors for their clear summary of the current state of approximate integration in statistics and for the opportunity to contribute some further suggestions based on my own research and experience in practical Bayesian inference.

Suppose for the moment that the aim is to use a numerical integration method over R^k to integrate with respect to a fairly well-behaved unnormalized posterior density $f(\theta)$. The various recommended methods then have similar stages: (1) initial parameterization, (2) iterative search for an appropriate representation of $f(\theta)$ that uses at least first and second moments and (3) estimation of the required integrals, preferably with diagnostics allowing one to reassess stages 1 and 2.

For example, *asymptotic methods* might use conjugate gradient methods in stage 2 to find the mode $\hat{\theta}$ and Hessian $H(\hat{\theta})$, followed by Laplace approximation for stage 3. *Iterative importance sampling* typically aims to represent $f(\theta)$ approximately by an importance sampler $w_\alpha(\theta)$ with appropriate first and second moments and tail behavior. *Multiple quadrature* approaches might at stage 2 use the Naylor–Smith algorithm to arrive at approximate first and second moments to standardize $f(\theta)$ and use a high-degree integration rule at stage 3. Similarly, *Markov chain methods* might iteratively choose the transition function $r(\theta, \cdot)$ based partly on current estimates of posterior moments and increase the number of points for a given $r(\theta, \cdot)$ when the process appears to be converging.

As the authors say, it is important for the user to have the option of using any of these methods. This facilitates convergence at stage 2 as well as error estimation at stage 3. Before investigating such an ambitious unified approach, comments follow on some specific integration methods that I have found useful.

IMPORTANCE SAMPLING

An alternative to spherically symmetric importance sampling is given by forming the importance sampler w as a product of k univariate importance sampling functions. This approach was briefly described in Shaw (1988), and implemented in the BAYES FOUR package as described in Naylor and

J. E. H. Shaw is a Lecturer, Department of Statistics, University of Warwick, Coventry CV4 7AL, U.K.

Shaw (1991). Particularly useful univariate families for importance sampling are given implicitly by

$$X_A = A\phi(U) - (1 - A)\phi(1 - U), \tag{1}$$

where U is a random variable with a uniform $U(0, 1)$ distribution, $0 \le A \le 1$, and $\phi(\cdot)$ is a simple monotonic increasing function on $(0, 1)$, such that $\phi(u) \to -\infty$ as $u \to 0$. For example, if $\phi(u) = \log(u)$, then X_0 has an exponential distribution, $X_{0.5}$ has a logistic distribution and X_A for other values of $A \in (0, 1)$ has a skewed distribution with support $(-\infty, \infty)$ and with exponential tail behavior. Another example is $\phi(u) = \tan(0.5\pi(u - 1))$, for which $X_{0.5}$ has a Cauchy distribution and X_0 has a half-Cauchy distribution.

These families have several attractive properties apart from their ease of generation. For example, for a given $\phi(\cdot)$, the distance from the pth to the $(100 - p)$th percentile is independent of A, so that the interpercentile distances are convenient summaries of scale and tail behavior. Also, the median and the mean of X_A are both linear in A, and the variance is quadratic in A. Such properties help one choose the importance sampler iteratively to match its characteristics to those of the integrand in each dimension separately. Note by the way that the importance sampler in BAYES FOUR is scaled so as to integrate a multivariate normal density with minimum mean squared error, rather than using the inverse Hessian at the mode as described in Section 8 of the paper.

For a smooth unnormalized posterior $f(\cdot)$, the efficiency of the above importance sampling method is intimately related to the use of quasirandom sequences as in Shaw (1988). Provided the importance sampling function $w(\cdot)$ has heavier tails than $f(\cdot)$ (which is in any case usually desirable for stability), the above approach is equivalent to integrating over $(0, 1)^k$ a smooth function that tends to 0 at the boundary. Quasirandom sequences are widely recognized to be extremely efficient with periodic integrands like this. They can also be useful in many other integration methods, for example, by generating points efficiently on concentric ellipsoidal shells as in Shaw (1988a). Important additional references to quasirandom sequences and related methods are Fang and Wang (1993) and Sloan and Joe (1994).

MULTIPLE QUADRATURE

The development of certain novel monomial rules, with positive weights and real (rather than complex) nodes was reported briefly in Shaw (1993). These *designed rules* use error-correcting codes and related theory to attain high efficiency. For example, a designed Gauss–Hermite rule of monomial degree 7 in 5, 10 and 20 dimensions requires roughly 100, 1000 and 8500 function evaluations respectively, compared to roughly 1000, 10^6 and 10^{12} evaluations for the corresponding product Gauss–Hermite rules! This allows multiple quadrature methods to become competitive for reasonable integrands in much higher dimensions than the $k \le 6$ suggested in Section 8 of the paper. Note that designed rules do not suffer from the curse of dimensionality as does the subregion adaptive algorithm.

Cameron and van Lint (1991) and Conway and Sloane (1988) are excellent references for coding theory. I hope to use computer algebra (see, e.g., Cox, Little and O'Shea, 1992) to help find further efficient quadrature rules including ones with degree $d > 7$.

A UNIFIED APPROACH?

The above importance sampling and multiple quadrature approaches may fail if $f(\theta)$ is not as well behaved as initially supposed; for example, if it exhibits multimodality, extreme skewness or heavy tails, "banana-shaped contours" and so forth.

Automatic reparameterization based simply on the given ranges of the initial parameters may cure such behavior and can be made transparent to the user once the statistical problem has been specified. However, if no simple transformation can be found to transform the integrand into one with a single dominant peak and approximately elliptical contours, then it is often very dangerous to attempt to identify a complicated reparameterization: the transformed $f(\theta)$ can become more and more badly behaved, but in ways that become harder and harder to pick up.

An alternative approach is to decompose $f(\theta)$:

$$f(\theta) = f_1(\theta) + \cdots + f_n(\theta), \tag{2}$$

where each $f_i(\cdot)$ is a nonnegative function and attempts to capture a particular aspect of $f(\cdot)$. For example, an $f_i(\cdot)$ may be centered at each mode. One possibility is to define

$$f_i(\theta) = f(\theta) \times g_i(\theta) \Big/ \sum_j^n g_j(\theta), \tag{3}$$

where each $g_i(\cdot)$ has a multivariate normal shape whose mean, variance and weight are chosen iteratively. This *multikernel* approach works well with a range of specific statistical models, but has so far proved difficult to automate in full generality for the naive user.

I have been developing a general Bayesian analysis package called BINGO based on this unified approach of automatic initial reparameterization and

iterative multikernel representation with a wide choice of integration methods. The Warwick Statistics World-Wide Web page (currently *http://www.warwick.ac.uk/WWW/faculties/science/Statistics*) will contain further information and downloadable software, including a catalogue of efficient designed integration rules in up to 20 dimensions, and an updated version of Shaw (1993).

Comment

Alan Genz and John Monahan

The authors are to be congratulated on their excellent survey (including an extensive and current reference list) of numerical integration methods for statistical computation problems. They also demonstrate how a multivariate Student$_k(\zeta)$ density can be used effectively in an adaptive importance sampling algorithm for some problems.

PARAMETERIZATION

A key issue for statistical numerical integration methods is what the authors refer to as parameterization (or reparameterization). Most methods begin with some kind of standardizing transformation, and the most common choice is $\boldsymbol{\theta} = \boldsymbol{\mu} + C\boldsymbol{\omega}$, where $\boldsymbol{\mu}$ is the posterior mode and C is the Cholesky factor of the negative inverse Hessian at $\boldsymbol{\mu}$. This kind of transformation is very useful for problems where the posterior has a global maximum at μ, but after this transformation has been made, two features of the transformed posterior are sometimes present: skewness and thick tails. When these features are not present, a multivariate normal model usually accounts for most of the posterior behavior, so that asymptotic methods can be used to provide good estimates for simple posterior expectations, and a variety of other simple methods can be used if higher accuracy if desired. However, if either of these features is present, the use of the simple multivariate normal (MVN) model can lead to inaccurate and misleading results. Thick tails create the biggest problems and are the most difficult to recognize and remedy. In high dimensions, the tail behavior of the posterior function is often impossible to determine analytically, and even in one dimension, the logistic distribution resembles the normal in shape so that the two may be difficult to distinguish visually. The differences in tail behavior may have unrecognizable consequences.

After standardization, the tail of the normal extends as far as 3.1, where the tail area is 0.001, or perhaps to 3.7, where the tail area is 0.0001. For a logistic, these points correspond to 6.9 and 9.2, and, after the modal standardization, still extend to 5 and 6.5. Using normal-like tail cutoff points will miss substantial probability mass, and the effect on first and second moments will be substantial. Some methods, such as the subregion adaptive approach, which essentially do not rely on tail behavior, may still be affected. The inverse normal transformation is often used to map the unit interval to the real line. Here the problem arises in that there may be no points near 1 which correspond to any beyond 10 on the real line, even using double precision. Alternately, if in order to avoid a transformation to $[0, 1]^k$, a Gauss–Hermite product rule is applied directly to the standardized posterior, a rule with at least 32 points in each variable needs to be used to account for tails with significant content beyond $\omega_i = \pm 10$, and the use of such a rule is probably infeasible for $k > 3$ or 4.

The multivariate Student$_k(5)$ has a starkly contrasting very thick tail. Here our fear may instead be that so much effort is devoted to controlling the tail that most of the time might be spent evaluating the posterior where the mass may be small and insignificant. Clearly, a middle ground may be preferable, and a useful addition to the authors' method would be a heuristic for estimating a good value for the degrees of freedom parameter ζ.

The authors do suggest using the sum $\sum W_i^{*2}$ as a diagnostic for troublesome tail behavior in importance sampling. The second author of this comment (Monahan, 1993) investigated the use of this diagnostic for testing whether the weights have a finite variance, and found that estimates of tail rates work much better in this regard. The reader should note that when improper priors are used, these weights may even fail to have a finite mean.

Alan Genz is Professor, Department of Mathematics, Washington State University, Pullman, Washington 99164-3113. John Monahan is Professor, Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695-8203.

The authors combine the use of the multivariate Student model with an adaptive strategy for updating the location μ and scale C parameters. This strategy can handle skewness by shifting the center of the model. However, this strategy is still based on a centrally symmetric model and this approach may be an inefficient way to deal with skewness.

THE EXAMPLES

The results for the examples show that the adaptive importance sampling method with a multivariate Student model can be very effective. The actual problem for the first example is really only a two-dimensional problem when the θ_{10} variable is treated as an outer integration variable, because the integrand for this variable is then just a product of nine independent one-dimensional integrals. The authors could have exploited this feature to save time computing "exact" values for the integrals in this problem. However, the problem as posed does make a good 10-dimensional test problem for statistical numerical integration software. The integrands for each of the other nine θ_i variables behave asymptotically like θ_i^{-20} or a Student(19), and this suggests that the adaptive importance sampling method should be more efficient if a larger value of ζ is used, even though the authors did not observe this with the tests that they have reported. The comparisons with results from the subregion adaptive method (software ADBAYS) are perhaps a little unfair, for two reasons. One reason is that the authors use ADBAYS only with an MVN model and without adaptively adjusting the model parameters. The first author of this comment did obtain improved results with ADBAYS using a Student(5) model combined with adaptive adjustment of the model parameters. The second reason is that the error estimates produced by ADBAYS are known to be very conservative (Genz and Kass, 1993), and the long running times for ADBAYS could have been reduced if a less conservative error estimate (as discussed in Genz and Kass, 1993) had been used. Similar comments could be made about the use of ADBAYS for the second example. However even after taking these comments into account, the dimensionality in both of these problems is probably too high for currently available subregion adaptive methods to be competitive with adaptive importance sampling. Overall, the the authors have made a fair assessment of the usefulness of subregion adaptive methods for statistical numerical integration problems. The third example is probably not a very good example for the adaptive importance sampling method, because there are already available good methods for computing MVN probabilities (see Genz, 1993).

Comment

Mark J. Schervish, Larry Wasserman and Russ Wolfinger

1. INTRODUCTION

We congratulate the authors on a timely and well-written paper. Given the current attention that Markov chain Monte Carlo (MCMC) is receiving, it is important to be reminded that there are other methods available. The choice of method should certainly be problem dependent. We find ourselves in agreement with much of what the authors say. Our discussion focuses on several different issues. In Section 2 we discuss the importance of using algorithms that are simple to program. In Section 3 we discuss the examples and, in particular, we note that Example 2 may not be appropriate for comparing the methods. In Section 4 we touch upon the problem of computing normalizing constants from simulation output. Finally, in Section 5 we consider the authors' classes of methods in the context of the variance components model.

Mark Schervish is Professor, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania. Larry Wasserman is an Associate Professor, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania. Russ Wolfinger is a Senior Research Statistician, SAS Institute Inc., Cary, North Carolina.

2. MACHINE EFFICIENCY VERSUS HUMAN EFFICIENCY

In comparing the efficiency of various methods we must bear in mind that CPU time is only one dimension. Another is how difficult it is to write a

program to use the method. If a method is so simple that a user can write a program from scratch very quickly, then the method has a chance of becoming widely used. Consider Example 2. The authors did not consider a MCMC approach. By introducing a latent variable that indicates which of the two contingency tables each subject is from, it is quite simple to write a Gibbs sampling program; we did so in less than 30 min. It seems unlikely that the other methods can be programmed so quickly and easily. The actual running time of the program is often not crucial since we can do other things while the computer runs the program. In contrast, we are pretty much stuck in front of the computer screen while programming. (Incidentally, we wrote both a Fortran version and S-Plus version. The Fortran program took 16 s for 10,000 runs. The S-Plus program took 9 min; not swift but not terrible either.)

3. THE EXAMPLES

The authors chose three examples to compare the methods. We agree that using test examples is the best way to compare methods. Example 2, however, may not be a good choice for comparing methods since it is nonidentifiable.

In fact, Example 2 has two sources of nonidentifiability. One source of nonidentifiability comes from "label-switching," that is, it is impossible to say which table is to be called (1) and which one is to be called (2) without further restriction. The authors have imposed the restriction that $\theta \leq 1/2$. However, there is another source of nonidentifiability. A 3×3 table has only 8 degrees of freedom, while the authors fit a 9 parameter model. This second source of nonidentifiability manifests itself as a ridge of high likelihood. (The label-switching produces a second, mirror-image ridge of high likelihood. Oddly enough, both of these ridges include substantial portions with $\theta \leq 1/2$. Each of these portions represents that portion of the other ridge where $\theta \geq 1/2$ and the labels are switched.) As the sample size goes to infinity, the posterior converges to a singular distribution on a lower dimensional manifold (rather than a point). To put it another way, even if we knew the true cell proportions for the 3×3 table, we could still decompose this table as the mixture of two independence tables in infinitely many ways. Hence, none of the parameters has a "true value" and it is not meaningful to produce point estimates of θ.

All the methods used by the authors for this example assume that the posterior is unimodal and well behaved. None is appropriate because of the peculiar shape of the posterior. It is worth remarking that we discovered this problem by observing several different runs of MCMC some of which migrated to one of the areas of high likelihood and some of which migrated to the other one.

We have to agree with the authors that Example 3 does not lend itself to a MCMC solution. The algorithm of Schervish (1984) is not designed for achieving absolute accuracy of 10^{-10} for six-dimensional problems in any reasonable amount of time. However, when converted to double precision, it does consistently produce the answer 0.166626×10^{-4} when one requests absolute accuracy of 10^{-3}, 10^{-4}, 10^{-5} or 10^{-6}. (This last took 12.4 h, so we did not continue the sequence.) The absolute error bound is apparently far too conservative in this example. It is unfortunate that we could not know that in advance.

In Example 1, the authors used both independence and random walk Metropolis chains. In the latter case, it appears that the authors used a multivariate random walk. In our experience, these two MCMC methods—independence chains and multivariate random walks—are not very effective. Indeed, these were found to be less accurate than Gibbs sampling by the authors. A third type of Metropolis chain, that often does better, is the combination of Metropolis and Gibbs that the authors attribute to Mueller (1991). In this chain, one generates a candidate for a single coordinate of θ (often with a symmetric distribution centered at the current value), accepting or rejecting it with probabilities given by the Metropolis–Hastings method. Then one cycles through the coordinates over and over as in Gibbs sampling. The advantage to this method is that it does not require the user to estimate any of the joint distributional information about θ in order to implement the chain. Typically, one merely adjusts the scaling of the candidate distributions to meet a desired acceptance rate for each coordinate.

4. NORMALIZING CONSTANTS

In Section 8, the authors say that Gibbs sampling does not provide an easy method to estimate the normalizing constant of the posterior. Actually, there are several methods for estimating the normalizing constant from MCMC output. DiCiccio, Kass, Raftery and Wasserman (1995) review five such methods including simulation-based and Bartlett-corrected Laplace estimators, reciprocal importance sampling (Gelfand and Dey, 1994) and bridge estimators (Meng and Wong, 1993). The latter actually requires two simulations, for example, a MCMC run followed by an importance sampling run. However, the importance sampler can be chosen using the

MCMC output. Nobile (1994, Section 4.3) presents a conditional importance sampling method tailor-made to Gibbs sampling. Space does not permit a detailed account of these methods here. Explanations of these and other methods can be found in the aforementioned papers as well as in Carlin and Chib (1995), Gelman and Meng (1994), Green (1994), Phillips and Smith (1994), Raftery (1995) and Verdinelli and Wasserman (1995).

We applied a few of these methods to Example 1. The conditional importance sampling estimator gives $\hat{I}(1) = 0.115 \times 10^{-21}$, the simulation Laplace gives 0.133×10^{-21}, reciprocal importance sampling gives 0.130×10^{-21} and a particular version of the bridge estimator gives 0.132×10^{-21}. We note that the first three use only the MCMC output.

5. ANOTHER EXAMPLE: THE VARIANCE COMPONENTS MODEL

A practical extension of the linear model discussed in Example 1 is the mixed linear model

$$y = X\beta + Z\gamma + \varepsilon,$$

where y denotes the data vector, β is a vector of unknown fixed-effects parameters with known design matrix X, γ is a vector of unknown random-effects parameters with known design matrix Z and ε is a vector of error disturbances. The most common assumptions associated with this model are that γ and ε are independent multivariate Gaussian vectors with zero means and variance–covariance matrices G and R, respectively.

Although the assumption of a Student distribution is certainly possible for γ and/or ε, we focus on the familiar Gaussian case for ease of exposition. Furthermore, we initially assume that $R = \sigma^2 I$, where I is an identity matrix, and G is a diagonal matrix of variance components. Using θ to denote the vector of all unknown variance parameters, we assume a flat prior for β and a Jeffreys prior for θ (Zellner, 1971; Box and Tiao, 1973; Harville, 1974; Broemeling, 1985).

The conditional posterior distribution of β given θ is multivariate normal with mean $(X'V^{-1}X)^{-}X' \cdot V^{-1}y$ and variance–covariance $(X'V^{-1}X)^{-}$, where $V = ZGZ' + R$ and the superscript minus $(^{-})$ denotes a generalized inverse to account for possible rank deficiencies in X. Unfortunately the remaining posterior results are difficult to derive analytically (Gelfand, Hills, Racine-Poon and Smith, 1990; Searle, Casella and McCulloch, 1992). This model is therefore an excellent candidate for application of the approximation methods described by Evans and Swartz.

As a practical exercise, let us briefly investigate the feasibility of the five classes of methods described by Evans and Swartz with reference to the variance components model. Even though Evans and Swartz focus on integration, a general objective associated with this model is to carry out Bayesian inferences on a variety of functions of β and θ. As they discuss, often these inferences are in the form of an integral with respect to the joint posterior density β and θ.

Because of the analytical difficulties associated with this model, especially with unbalanced data, the asymptotic methods appear to be too challenging to pursue in general. As mentioned by Gelfand et al. (1990), these calculations are usually very function specific and thus typically have to be performed again from scratch for each new function of interest. Similar arguments apply to the multiple quadrature methods, in spite of the fact that the dimension of θ is usually less than 6.

The sampling-based methods have much more promise. They have the attractive property of allowing one to construct any function of interest once a sample from the joint posterior density of β and θ has been generated. The problem then becomes one of selecting among the plethora of sampling-based methods, and Evans and Swartz are to be commended for helping to sort out the differences between them (see also Tierney, 1994, who is excellent in this regard). They divide the sampling-based methods into three categories: importance sampling, adaptive importance sampling and Markov chain methods (including the Gibbs sampler). Each of these categories contains numerous possible algorithms, and selecting one among them is not a trivial task. An ideal algorithm would possess the following attributes:

- simple in form and construction;
- robust across a variety of data sets and models;
- computationally efficient;
- easily implementable.

It would also preferably generate an independent sample from the joint posterior of β and θ to avoid the difficulties associated with analyzing a dependent sample.

In an attempt to meet these algorithmic goals for the variance components model, it seems imperative to exploit the inherent structure of the model. This involves the natural separation between β and θ and the fact that the conditional density of β given θ is multivariate normal. Thus if the algorithm can generate a sample from the marginal posterior density of θ, then one for β can be easily obtained as with the Gibbs sampler.

Based on analytical work (Hill, 1965, and the preceding references), the joint marginal posterior density of the variance components can be closely approximated using products of the inverse gamma density

$$\mathrm{IG}(x; a, b) \propto x^{-(1+a)} e^{-b/x}.$$

Because the inverted gamma is often very right-skewed, importance samplers based on the Student or split-t do not seem appropriate; however, one can easily generate samples directly from inverted gamma densities. One algorithm that suggests itself here is to simply generate a sample from an appropriate importance sampling density and output its values along with the weights computed as ratios with the true density.

Gelfand et al. (1990) also make use of the inverted gamma density in discussing the Gibbs sampler for the variance components problem. The method works well for the example they consider and goes a long way toward achieving the goals of an ideal algorithm. Some difficulties with the Gibbs sampler for this problem are as follows:

- The conditional densities for unbalanced models with multiple variance components can be complex.
- Automatically assessing convergence of the algorithm can be tricky.
- The final sample is not independent.

An algorithm that avoids these difficulties is one which is not mentioned by Evans and Swartz: rejection sampling (Ripley, 1987; Smith and Gelfand, 1992; Tierney, 1994). This algorithm works by generating a pseudo random observation from a convenient base distribution (chosen to be as close as possible to the posterior) and then retaining that observation in the final sample with probability proportional to the ratio of the two densities times a bounding constant.

Whether rejection sampling is a better algorithm for the variance components model than importance sampling or some Markov chain method remains an open question. Perhaps Evans and Swartz could provide recommendations in this regard, especially with a view toward implementation in a commercial software package such as the SAS MIXED procedure (SAS Institute Inc., 1992). The rejection sampling algorithm has already been implemented in Release 6.11 and appears to work well for many common examples with rejection rates usually less than 10%.

Another popular mixed model worthy of consideration is the random coefficients model (Laird and Ware, 1982; Rutter and Elashoff, 1994). Here G is typically block diagonal, with blocks containing variance and covariance components for random intercepts, slopes and so forth. Recommendations for this model would be welcome as well.

6. FINAL REMARKS

The survey of numerical methods given by Evans and Swartz will be very useful to researchers faced with difficult integration problems. Apart from minor disagreements about some of the examples, we found this to be an excellent paper. It would be a service to the statistical community if the authors were to put some of their software for adaptive importance sampling on *Statlib*. Some S-Plus routines for doing subregion adaptive quadrature are already there.

Rejoinder

Michael Evans and Tim Swartz

A number of interesting and significant points have been made by the discussants and we are grateful for their attention to the paper. There are few mathematical certainties with numerical work and by necessity one's own experience is always somewhat limited. The points of view of others are therefore particularly important. There is still a lot to learn about the "right" way to approximate integrals and we remain convinced that none of the methods that we have discussed should be uniformly discarded in favor of some universal approach. Overall we believe the conclusions that we have drawn remain sound. If we had been in

Michael Evans is Professor, Department of Statistics, University of Toronto, Ontario M5S 1A1, Canada. Tim Swartz is Associate Professor, Department of Mathematics and Statistics, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada.

possession of the discussions before writing the paper, however, we would have done some things differently.

1. RESPONSE TO THE DISCUSSION OF SHAW

We look forward to the results of Professor Shaw's research on new multiple quadrature methods and the multikernel approach. The ultimate goal of all research in this area is to develop code that can be used reliably by practitioners and BINGO may be helpful in this regard.

The univariate importance sampling family mentioned in his comments does seem to us to be useful, but we have reservations concerning the recommended use of such families in high-dimensional contexts. For suppose we took as importance sampler on R^k a product of independent Cauchy densities; that is, $w(x) \propto \prod_{i=1}^{k}(1 + x_i^2)^{-1}$. Then along any coordinate axis the tail is going down like $||x||^{-2}$, but along any ray given by unit vector u with all $u_i \neq 0$, the density goes down like $||x||^{-2k}$. Thus the tails of such densities are much shorter between coordinate axes and this leads to a somewhat unnatural shape for the importance sampling distribution. In contrast, the multivariate Cauchy has tails like $||x||^{-2}$ in every direction.

As discussed in the paper there is a need to develop truly multivariate families of practical importance samplers. In particular there is a need to produce importance samplers to model skewness and other deviations from ellipsoidal shape. Presumably such families will also be found useful in the Markov chain context. As mentioned in the paper, we feel that the most promising candidate for such a family is given by mixtures of multivariate Student distributions, but more work remains to be done to extend its usage beyond that described in Oh and Berger (1992).

The simplicity of the Monte Carlo methods with respect to mathematical analysis, error estimation and the lack of dependence on dimension provide for us decided advantages over the use of quasirandom rules at this point. One possibility is to combine the methods into a hybrid as is done in Cranley and Patterson (1976) and Owen (1994). These are examples of a more general class of methods that includes antithetic sampling and were referred to generally as systematic sampling in the paper. As discussed in Evans and Swartz (1995) these methods are related to groups $\mathcal{T}$ of symmetries of the importance sampler and this is helpful in characterizing when these techniques will be useful in a given context. The most important point in ensuring a successful approximation is to get the importance sampler right, and when we do, systematic sampling is markedly less efficient than straight importance sampling. This is because the integrand is close to being invariant under the relevant symmetries in such a situation. The value in systematic sampling methods, from this point of view, arises when we have made a very poor choice of importance sampler and thus the integrand is far from being invariant under the subgroup. On the basis of this analysis we conclude that systematic sampling should be used with caution. For some recent work on the related *randomized quadrature rules*, see Monahan and Genz (1995). The research outlined in Professor Shaw's discussion could make a contribution to this class of techniques as well.

2. RESPONSE TO THE DISCUSSION OF GENZ AND MONAHAN

The comments concerning tail length strike us as being particularly appropriate. We wish we had more to say on this topic. Our choice of 5 degrees of freedom for the multivariate Student importance sampler was an attempt to be conservative without going too far. Given that there are typically many integrals to evaluate, we did not try to optimize the choice for each case. We have heard recommendations as low as 1 degree of freedom for the importance sampler but we agree with Genz and Monahan on the principle of trying to get the tail length reasonably correct so that sampling takes place primarily where the posterior is high. What is needed is a good automatic way to choose the degrees of freedom.

Diagnosing when an importance sampler has failed is a difficult issue. While the sum of squares of the normalized weights is an appropriate measure of the accuracy of the approximation of the integral, it is unlikely to work well as a diagnostic when the importance sampler is very bad. In fact it can be very misleading. We look forward to any additional work by the discussants that can shed light on this problem. We also agree with the comments concerning the need to model skewness; see our response to Professor Shaw.

In choosing Example 1 we wanted it to be representative of the typical level of difficulty one might encounter in implementing a Bayesian analysis of a linear model with nonnormal error. Thus it would not have been appropriate for us to exploit the special feature of the problem pointed out in the discussion. On the other hand, in developing software to be used by practitioners, it is important that such characteristics be recognized by the software and exploited. Many statistical problems will exhibit such

structure and the computational savings can be substantial.

We certainly believe that our usage of ADBAYS can be improved with adaptive modifications as described. As mentioned in the paper it is our impression that this is the algorithm of choice for relatively low-dimensional problems. Anything that extends the upper limit of the dimension for practical computation times is good news. It would also be helpful if less conservative error estimates could be obtained from ADBAYS.

RESPONSE TO THE DISCUSSION OF SCHERVISH, WASSERMAN AND WOLFINGER

(a) Machine Efficiency versus Human Efficiency

Programming effort is an important aspect to be considered when choosing among algorithms for a specific problem. This is a difficult issue upon which to make comparisons, however, as it depends to a certain extent on individual tastes and on available software. Hopefully, in the near future, there will be software available that will minimize the need to write code as much as possible for any of the methods we have discussed, so that this will not be an issue. To a certain extent this is already happening as summarized in the Conclusions section of the paper, but having such facilities available from a convenient environment like S would be valuable. The speed of simulations is a limiting factor, however. While 10,000 iterations may seem like enough for some problems, our experience suggests that this is not the case for many other problems. Furthermore, 9 minutes to wait for an outcome does not encourage experimentation and this is an important aspect of numerical work.

With many examples Gibbs sampling is straightforward to code and, as we said in the paper, this is a strong point in its favor in such contexts. In particular Gibbs sampling avoids the need to maximize and this eliminates the programming of derivatives. On the other hand, it is not in general easy to generate from the requisite conditional distributions, and the associated programming problems can be difficult or even intractable when efficient algorithms are desired.

(b) The Examples

The Gibbs sampling algorithm for Example 2, as mentioned by the discussants, is particularly straightforward to implement and we wish in retrospect that we had tried it. In particular this would have lead to modifications in all of the methods we applied to this example, including two Markov chain algorithms.

Whenever $2(I + J) - 3 > IJ - 1$ in an $I \times J$ table then, as noted in the discussion, this model suffers from nonidentifiability even when the switching symmetry is removed via the prior as we did. This is reflected by the posterior in this example, in the basic parameterization, having one-dimensional contours that are independent of the data obtained. This does not invalidate the use of this model, however, as it is still reasonable to ask whether such a model can fit the data and to consider a Bayesian analysis. There is an issue concerning what are relevant estimates to quote for model parameters, but we see nothing wrong with quoting estimates provided that it is acknowledged that all values in the same contour are inferentially equivalent. So the problem of computing various integrals for this model is, in our opinion, still relevant.

It is interesting, however, that the deviation in the results obtained by Gibbs sampling does not seem to be caused by this second type of nonidentifiability. For in the original parameterization it is true that the posterior will have a ridge where it is maximal, but after the reparameterization this ridge is removed because of the Jacobian factor, and the transformed posterior has a unique global maximium. It turns out, however, that there is another mode in the new parameterization, or equivalently another ridge of high posterior probability density in the original parameterization. This was discovered in our original computations, but its effect was not explicitly taken into account. Our claim of only two modes is based on searches starting from 10^4 uniformly chosen points in the basic parameterization. At the global maximum the transformed log posterior takes the value -271.90, while at the second mode it takes the value -273.06. Implementing the Student mixture importance sampling algorithm described in the paper gives a mixture weight of 0.647 to the multivariate Student at the global maximum. The estimates of $R(\theta)$ and $R(\theta^2)$ computed via this algorithm and 10^5 iterations equal 0.431 and 0.188, respectively. These results were verified by a brute force calculation based on a sequence of product Gauss–Jacobi rules of increasing order. Notice that all moments of the parameters can be computed exactly using such rules of appropriate orders (see Evans, Gilula and Guttman, 1989, for more details), but this requires far too much computation to implement. The Gibbs algorithm got stuck in each of the high posterior density regions depending on what starting values were given. It did not give accurate estimates of the pos-

terior means in practically meaningful computation times.

The estimates of model parameters obtained using integration strategies based on approximating the transformed posterior at the global maximum, as was done in all the algorithms used in the paper, provides a remarkable fit to the data. For example, the chi-squared goodness-of-fit statistic equals 1.26. We also obtain a good fit when we base our integration strategies on the second lesser mode, as the chi-squared statistic equals 1.97 in this case. On the other hand, when we accurately compute the posterior means of the parameter values using Student mixture importance sampling we obtain a chi-squared value of 22.98. Thus we are confronted with the general issue of what are appropriate inference methods for parameters when the posterior is multimodal. From the point of view of accurate computations, however, the only algorithm that we would recommend here is the importance sampling algorithm via a mixture of Students.

(c) Normalizing Constants

The additional references on the estimation of the normalizing constant from the output of a Markov chain algorithm are valuable and should have been included in our review. In many contexts, such as Example 1, importance sampling and adaptive importance sampling provide simpler methods for estimating the normalizing constants. This is because importance sampling methods avoid the need to assess convergence to stationarity, the selection of an appropriate error estimate and the specification of additional ingredients to the simulation. Of course in contexts where the importance sampling algorithms are not feasible the methods outlined in the discussion are necessary.

(d) The Variance Components Model

The rejection algorithm mentioned in the discussion can be considered as an alternative to importance sampling. For suppose that w is the density generating the base chain for the rejection algorithm as described in Tierney (1994). Further suppose that we have a constant c such that $f \leq cw$, where f is the unnormalized posterior. Then we can estimate $I(1)$ using the rejection algorithm by recording the proportion of acceptances in n steps. It is easily shown, however, that this estimate always has variance at least as large as the variance of the importance sampling estimate of $I(1)$ based on w. In the general rejection algorithm we may not have the inequality $f \leq cw$ holding everywhere and, in addition, we are often primarily interested in estimating posterior expectations $R(m)$. Therefore the above result does not say that importance sampling is necessarily uniformly better than the rejection algorithm. It seems likely, however, that in a problem where the rejection algorithm performs well, that is, the rejection rate is low, then we might expect importance sampling to do as well or better. In such a situation there are some definite advantages to importance sampling; for example, the avoidance of the need to assess convergence to stationarity and straightforward measures of error. Therefore, without knowing the specific details of the rejection algorithm used in the SAS implementation, it would appear to be reasonable to consider importance sampling here. We hesitate to make a more concrete recommendation without further study.

Importance sampling based on a multivariate Student would seem to be feasible for this problem after reparameterizing by taking logs of the variance components. Implementation of this requires maximizing the posterior; we have not tried this, but it appears to be a tractable problem. The virtue of this approach, rather than trying to mimic the marginal posterior of the variance components by a product of inverse gammas, is that it takes into account the posterior correlations among all the parameters and, in our experience, these can sometimes be substantial. This may result in a more complicated algorithm than that currently implemented. In a package, however, this is not a relevant issue as the coding need only be done once. The essential criteria for comparing algorithms in such contexts are accuracy within reasonable CPU times and the existence of reliable diagnostics for the assessment of accuracy.

ADDITIONAL REFERENCES

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

BROMELING, L. D. (1985). *Bayesian Analysis of Linear Models*. Dekker, New York.

CAMERON, P. J. and VAN LINT, J. H. (1991). Designs, graphs, codes and their links. *London Math. Soc. Stud. Texts* **22**.

CARLIN, B. P. and CHIB, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *J. Roy. Statist. Soc. Ser. B* **57** 473–484.

CONWAY, J. H. and SLOANE, N. J. A. (1988). *Sphere Packings, Lattices and Groups*. Springer, New York.

COX, D. A., LITTLE, J. B. and O'SHEA, D. (1992). *Ideals, Varieties, and Algorithms*. Springer, New York.

CRANLEY, R. and PATTERSON, T. N. L. (1976). Randomization of number theoretic methods for multiple integration. *SIAM J. Numer. Anal.* **13** 904–1014.

DICICCIO, T. J., KASS, R. E., RAFTERY, A. and WASSERMAN, L. (1995). Computing Bayes factors by combining simulation and asymptotic approximations. Technical Report 630, Dept. Statistics, Carnegie Mellon Univ.

EVANS, M., GILULA, Z. and GUTTMAN, I. (1989). Latent class analysis of two-way contingency tables by Bayesian methods. *Biometrika* **76** 557–563.

EVANS, M. and SWARTZ, T. (1995). Bayesian integration using multivariate Student importance sampling. *Computer Science and Statistics: Proceedings of the 27th Symposium on the Interface*. To appear.

FANG, K-T. and WANG, Y. (1993). *Number-Theoretic Methods in Statistics*. Chapman and Hall, London.

GELFAND, A. and DEY, D. (1994). Bayesian model choice: asymptotics and exact calculations. *J. Roy. Statist. Soc. Ser. B* **56** 501–514.

GELFAND, A., HILLS, S. E., RACINE-POON, A. and SMITH, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *J. Amer. Statist. Assoc.* **85** 972–985.

GELMAN, A. and MENG, X. (1994). Path sampling for computing normalizing constants: identities and theory. Technical Report, Dept. Statistics, Univ. California, Berkeley.

GENZ, A. (1993). A comparison of methods for numerical computation of multivariate normal probabilities. *Computing Science and Statistics* **25** 400–405.

GENZ, A. and KASS, R. (1993). Subregion adaptive integration of functions having a dominant peak. Technical Report 586, Dept. Statistics, Carnegie Mellon Univ.

GREEN, P. J. (1994). Reversible jump MCMC computations and Bayesian model determination. Technical Report, Dept. Mathematics, Univ. Bristol.

HARVILLE, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika* **61** 383–385.

HILL, B. M. (1965). Inference about variance components in the one-way model. *J. Amer. Statist. Assoc.* **60** 806–825.

LAIRD, N. and WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38** 963–974.

MENG, X. and WONG, W. (1993). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. Technical Report 365, Dept. Statistics, Univ. Chicago.

MONAHAN, J. F. (1993). Testing the behavior of importance sampling weights. *Computing Science and Statistics* **24** 112–117.

MONAHAN, J. and GENZ, A. (1995). A comparison of omnibus methods for Bayesian computing. *Computer Science and Statistics: Proceedings of the 27th Symposium on the Interface*. To appear.

MUELLER, P. (1991). A generic approach to posterior integration and Gibbs sampling. Technical Report 91-09, Dept. Statistics, Purdue Univ.

NAYLOR, J. C. and SHAW, J. E. H. (1991). *BAYES FOUR User Guide*. Commercial Centre, Trent Univ., Nottingham, U.K.

NOBILE, A. (1994). Bayesian analysis of finite mixture distributions. Ph.D. Dissertation, Dept. Statistics, Carnegie Mellon Univ., Pittsburgh.

OH, M.-S. and BERGER, J. (1992). Adaptive importance sampling in Monte Carlo integration. *J. Statist. Comput. Simulation* **41** 143–168.

OWEN, A. B. (1994). Randomly permuted (t, m, s)-nets and (t, s) sequences. Technical Report, Dept. Statistics, Stanford Univ.

PHILLIPS, D. B. and SMITH, A. F. M. (1994). Bayesian model comparison via jump diffusions. Technical Report 94-20, Dept. Mathematics, Imperial College.

RAFTERY, A. (1995). Hypothesis testing and model selection via posterior simulation. In *Practical Markov Chain Monte Carlo* (W. Gilks, S. Richardson and D. J. Spiegelhalter, eds.). Chapman and Hall, London.

RIPLEY, B. D. (1987). *Stochastic Simulation*. Wiley, New York.

RUTTER, C. M. and ELASHOFF, R. M. (1994). Analysis of longitudinal data: random coefficient regression modelling. *Statistics in Medicine* **13** 1211–1231.

SAS INSTITUTE INC. (1992). SAS/STAT software: changes and enhancements, Release 6.07. Technical Report P-229, SAS Institute Inc., Cary, NC.

SCHERVISH, M. J. (1984). Multivariate normal probabilities with error bound. *Applied Statistics* **33** 81–87; Correction (1985) **34** 103–104.

SEARLE, S. R., CASELLA, G. and MCCULLOCH, C. (1992). *Variance Components*. Wiley, New York.

SHAW, J. E. H. (1988a). Aspects of numerical integration and summarisation (with discussion). In *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) **3** 147–159. Oxford Univ. Press.

SHAW, J. E. H. (1993). Statistical computing for bayesian applications. Available by ftp from *crocus.warwick.ac.uk*.

SLOAN, I. H. and JOE, S. (1994). *Lattice Methods for Multiple Integration*. Oxford Univ. Press.

SMITH, A. F. M. and GELFAND, A. E. (1992). Bayesian statistics without tears: a sampling–resampling perspective. *Amer. Statist.* **46** 84–88.

TIERNEY, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* **22** 1701–1762.

VERDINELLI, I. and WASSERMAN, L. (1995). Computing Bayes factors using a generalization of the Savage–Dickey density ratio. *J. Amer. Statist. Assoc.* **90** 614–618.

ZELLNER, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. Wiley, New York.