Methods for Approximating Integrals in Statistics with Special Emphasis on Bayesian Integration Problems

Michael Evans
Department of Statistics
University of Toronto
Toronto, Ontario M5S 1A1
Canada

Tim Swartz
Department of Mathematics and Statistics
Simon Fraser University
Burnaby, British Columbia V5A 1S6
Canada

Abstract

This paper is a survey of the major techniques and approaches available for the numerical approximation of integrals in statistics. We classify these into five broad categories; namely, asymptotic methods, importance sampling, adaptive importance sampling, multiple quadrature and Markov chain methods. Each method is discussed giving an outline of the basic supporting theory and particular features of the technique. Conclusions are drawn concerning the relative merits of the methods based on the discussion and their application to three examples. The following broad recommendations are made. Asymptotic methods should only be considered in contexts where the integrand has a dominant peak with approximate ellipsoidal symmetry. Importance sampling, and preferably adaptive importance sampling, based on a multivariate Student should be used instead of asymptotics methods in such a context. Multiple quadrature, and in particular subregion adaptive integration, are the algorithms of choice for lower dimensional integrals. Due to the difficulties in assessing convergence to stationarity and the error in estimates, Markov chain methods are recommended only when there is no adequate

alternative. In certain very high-dimensional problems, however, Markov chain methods are the only hope. The importance of the parameterization of the integral is noted for the success of all the methods and several useful reparameterizations are presented.

Keywords and phrases : asymptotics, importance sampling, adaptive importance sampling, multiple quadrature, subregion adaptive integration, Markov chain methods.

1. Introduction

Reliably approximating the values of integrals is a problem of substantial concern for statisticians. We will refer to this problem hereafter as the *integration problem*. As the dimension of these integrals rises, the difficulty of the integration problem increases. Even in relatively low dimensions, however, approximating these values can be difficult.

A wide variety of methods are aimed at solving the integration problem. However, novices can be confused about which method to use. There are advantages and disadvantages to each of the methods, and it is part of our purpose to elucidate these. Further, we hope that the paper leaves a potential practitioner with the confidence to begin using these methods.

In many statistical problems we need to approximate a number of integrals where only part of the integrand changes from one problem to another. Often it is the common part of the integrand that is of primary importance in choosing an approximation technique. For this reason we denote the integrals of interest as

$$I(m) = \int_{R^k} m(\theta) f(\theta) d\theta \tag{1}$$

where $m : R^k \to R$, $f : R^k \to R^+$ and f represents the common part. For example, in Bayesian contexts f is the product of the likelihood and prior and m is some function whose posterior expectation $R(m) = I(m)/I(1)$ is required. An arbitrary integration problem can be represented as in (1) by taking $f = 1$. We assume hereafter that the integrand mf is integrable, however, the

verification of this in a particular problem can be difficult.

A characteristic of many of the integrals encountered in statistical practice is the dependence of the integrand on sample data. In this situation we encounter a different integral for each data set. This dependence influences the degree of accuracy that we require in our approximations. For example, suppose that we wish to approximate a posterior mean and we have only two-decimal accuracy in the data. Then it makes no sense to require accuracy in the approximation higher than the amount of variation we would observe in the posterior mean if the data were varied in the third decimal place. This is in sharp contrast to the attitude commonly encountered in the numerical analysis community where the goal is high accuracy. High accuracy is relevant, however, to certain problems encountered in statistics. An example is given by the evaluation of the multivariate normal distribution function; i.e. m is the indicator function for $\prod_{i=1}^{k}(-\infty, x_i]$ and f is the $N_k(\mu, \Sigma)$ density. For high-dimensions and a requirement of low relative error, evaluation of the multivariate normal distribution function is still a largely unsolved problem; see for example Schervish (1984) and Genz (1992a).

A very important aspect of the integration problem is the assessment of error in an approximation. The only absolutely certain method is to compare the approximation with the correct answer, but this is impossible in practice. The various integration methods that we discuss each have special characteristics with respect to error assessment. In general, there is no guaranteed method for error assessment. With iterative methods, we have the natural method of examining the approximations at successive stages and stopping the iteration when the changes become small for a number of iterations. For a discussion of this issue in the general context of iterative algorithms, see Linz (1988). Further discussion on error assessment can be found in Lyness (1983) where automatic and standard quadrature are contrasted with respect to the behaviour of the error. Also Bernsten (1989) discusses methods for improving error assessment when using adaptive multiple quadrature

routines. Perhaps the best way to be confident that we have accurately approximated a particular integral is to use two very different methods and see if the results agree.

This paper is a substantial amplification of Evans and Swartz (1992). That paper outlined the major methods used in approximating integrals in statistics but did not consider specific examples and did not attempt to provide conclusions concerning the relative merits of the methods. A review paper by Haber (1970) reflects the views of the numerical analysis community on the integration problem at that time and in particular emphasizes multiple quadrature approaches. By contrast this paper is much more heavily weighted towards Monte Carlo methods. Also Thisted (1988), Flournoy and Tsutakawa (1989), Smith (1991) and Tanner (1993) contain general discussions of the integration problem.

In Section 2 we present some statistical examples where the integration problem is relevant. In Section 3 we discuss the use of asymptotic approximations. Section 4 is devoted to importance sampling and Section 5 to adaptive importance sampling. Section 6 reviews multiple quadrature and subregion adaptive methods. In Section 7 we discuss Markov chain methods. A final section summarizes our conclusions. All computation times in the paper refer to work carried out on a Sparc 20 workstation.

2. Examples

An issue of fundamental importance to the success of any of the integration techniques that we discuss is the *parameterization* chosen for the integral. By a reparameterization we mean any one-to-one, continuously differentiable transformation of the variables of integration. As we will see, certain parameterizations are preferable to others for particular methods. For example, choosing a parameterization that makes a normalized f as much like a multivariate normal density as possible is important for the success of asymptotic methods. Transformations that locate and scale variables appropriately and reduce correlations are important for all Monte Carlo and multiple quadrature

methods. Some methods require that variables range over specific sets; e.g. $[0,1]$ or R^1, and so a transformation is essential. In general, however, the choice of a parameterization to ensure the success of a particular method of integration is an unsolved problem. Helpful comments on this problem can be found in Achcar and Smith (1990) and Hills and Smith (1992). Throughout this paper we provide further comments on the choice of parameterization for an integral.

The following examples are used in subsequent sections.

Example 1. The integrals in this example arise from a Bayesian analysis of a linear model with simulated data. The statistical model for the observed response value $\mathbf{y}$ is specified as follows. We have $\mathbf{y} = X\beta + \sigma\mathbf{z}$ where $\mathbf{y} \in R^{45}, X \in R^{45\times 9}$ has $x_{ij} = 1$ for i and j satisfying $5(j-1)+1 \le i \le 5j$ and 0 otherwise, $\beta \in R^9, \sigma \in (0,\infty)$ and $\mathbf{z} \in R^{45}$ is a sample of 45 from the Student*(3) distribution where Student$^*(\lambda)$ denotes the Student(λ) distribution standardized to have variance 1. The data $\mathbf{y}$ were obtained by setting $\beta_i = 0$ for every $i \neq 1$, $\beta_1 = 1.8$, $\sigma = 1$ and the z's were generated from a Student*(3) distribution using the S generator for the Student(3) distribution and dividing each of these values by $\sqrt{3}$. We denote the y-value corresponding to the j-th observation in the i-th cell by y_{ij}. Table 1 gives the generated z_{ij} values.

We place a flat prior on the β's and independently the prior $1/\sigma$ on σ and fix $\lambda = 3$. Using a Student distribution for the error is what makes this analysis difficult because nothing can be worked out in closed form. Rather we are forced to use numerical techniques to implement the inference methods. It is convenient to transform from σ to $\nu = \log\sigma$ as ν is unconstrained in R^1. Thus the integrals that we wish to evaluate take the form given by (1) where $\theta \in R^{10}$ with components $\theta_i = \beta_i$ for $i = 1,\ldots,9$ and $\theta_{10} = \nu$, $f(\theta) = \exp\{-9\,n\,\theta_{10}\}\prod_{i=1}^{9}\prod_{j=1}^{n} g_\lambda\left(\frac{y_{ij}-\theta_i}{\exp(\theta_{10})}\right)$ is the likelihood times the prior, m is a function on R^{10} such that (1) is finite and

$$g_\lambda(z) = \frac{\Gamma(\frac{\lambda+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{\lambda}{2})}\left(1+\frac{z^2}{\lambda-2}\right)^{-\frac{\lambda+1}{2}}\frac{1}{\sqrt{\lambda-2}}$$

i	j=1	j=2	j=3	j=4	j=5
1	0.152	1.086	0.591	-0.068	-0.378
2	-0.638	0.187	-0.010	0.495	0.262
3	2.833	1.098	-0.476	-0.383	1.129
4	0.042	0.174	-0.582	0.447	-0.071
5	-0.616	-0.514	-1.321	0.043	0.073
6	-0.817	0.935	-0.156	-0.172	-0.174
7	-0.879	0.028	-0.110	-1.193	0.974
8	-0.611	-0.549	-0.545	-0.403	-1.300
9	0.290	0.755	-0.150	0.746	-3.001

Table 1: The values of z_{ij} used in the example.

with $\lambda = 3$ and $n = 5$. We will approximate the posterior expectations $R(m)$ corresponding to $m(\theta) = \theta_i$ and $m(\theta) = \theta_i^2$ for $i = 1, 2, 4$ and 10. The exact values of these quantities are given in Table 2 and were computed by an extensive computation described in Section 4.

Example 2. In this example, we carry out a Bayesian analysis of a contingency table using data from Wing (1962). The data involves the cross-classification of 132 long-term schizophrenic patients into three row categories concerning the frequency of hospital visits and three column categories concerning the length of stay. The (i, j)-th cell probability takes the form $p_{ij} = \theta\alpha_i(1)\beta_j(1) + (1-\theta)\alpha_i(2)\beta_j(2)$ where $\theta \in [0, 1/2]$ and both $(\alpha_1(i), \alpha_2(i), \alpha_3(i))$ and $(\beta_1(i), \beta_2(i), \beta_3(i))$ comprise probability distributions for $i = 1, 2$. This is known as a latent class model where rows and columns are statistically independent given a single latent variable and we have only the count data. For further discussion of this model see Evans, Gilula and Guttman (1989) and the references therein. The likelihood function equals $\prod_{i=1}^{3}\prod_{j=1}^{3}(\theta\alpha_i(1)\beta_j(1) + (1-\theta)\alpha_i(2)\beta_j(2))^{f_{ij}}$ where f_{ij} is the count in the (i, j)-th cell. For the prior distribution we take $\theta \sim U(0, 1/2)$, $(\alpha_1(i), \alpha_2(i), \alpha_3(i)) \sim$ Dirichlet$(1, 1, 1)$ and $(\beta_1(i), \beta_2(i), \beta_3(i)) \sim$ Dirichlet$(1, 1, 1)$ for $i = 1, 2$ and assume independence. Therefore f in (1) equals the likelihood multiplied by these priors. We are interested in the posterior expectations $R(\theta) = .422$ and $R(\theta^2) = .181$. These values were obtained

via an extensive computation described in Section 4.

While the above parameterization is the natural one, it is not useful for any of the integration methods we discuss. Rather we use a *probability transform* to transform to $[0,1]^9$. We define this transformation generally. Let $x = (x_1, \ldots, x_k) \in R^k$ have density w and let $W_i(x_i|x_1, \ldots, x_{i-1})$ be the conditional distribution function of x_i given $x_1, \ldots, x_{i-1}$. Then putting $\theta_i = W_i(x_i|x_1, \ldots, x_{i-1})$ we have that $\theta \sim U([0,1]^k)$ and this is the probability transform. Of course any permutation of the coordinates can be used. For this example, following Evans, Gilula and Guttman (1989), we transform to $[0,1]^9$ using the probability transform of the prior with the natural ordering of the variables.

Example 3. In this example, we consider the computation of an orthant probability in R^6; i.e. $k = 6$, with high accuracy. For this integral $m(\theta) = I_{[0,\infty)^6}(\theta)$ is the indicator function of $[0,\infty)^6$ and f is the density of the $N_6(0,\Sigma)$ distribution where $\Sigma^{-1/2} = \mathrm{diag}(0,1,2,3,4,5) + ee'$ and $e = (1,1,1,1,1,1)'$. The exact value of this integral to 10 decimal places is $I(m) = .166625 \times 10^{-4}$ and this was computed via an extensive computation described in Section 4. We make use of a sequence of reparameterizations discussed in Genz (1992a). The first transformation is from $\theta \in [0,\infty)^6$ to $u = C^{-1}\theta$ where C is the lower triangular Cholesky factor of Σ. The second transformation is from $u \in R^6$ to $v = \Phi(u)$ where Φ is the $N(0,1)$ distribution function applied coordinate-wise. The interval of integration for v_i is $(a_i, 1)$ where $a_1 = 1/2$ and $a_i = \Phi(\{-c_{i1}\Phi^{-1}(v_1) - \ldots - c_{ii-1}\Phi^{-1}(v_{i-1})\}/c_{ii})$ for $i > 1$. The third transformation is given by $w_i = (v_i - a_i)/(1 - a_i)$ for $i = 1, \ldots, 6$. The integral now has a constant integrand with domain of integration for w equaling $[0,1]^6$ and the integration with respect to w_6 can be carried out exactly.

3. Asymptotics

One of the most commonly used asymptotic approximation techniques is the *Laplace method*. For this we suppose that (1) can be written in the form $I(m) = \int_D m(\theta)e^{-\tau h(\theta)}d\theta$ where D is

open, h has a continuous Hessian matrix H on D, h has a strict global minimum at $\hat{\theta} \in D$ and m is continuous in a neighbourhood of $\hat{\theta}$ with $m(\hat{\theta}) \neq 0$. We then approximate (1) using *Laplace's formula*

$$\hat{I}(m) = m(\hat{\theta})(2\pi)^{k/2}|\tau H(\hat{\theta})|^{-1/2}\exp\{-\tau h(\hat{\theta})\} \tag{2}$$

since $\lim_{\tau\to\infty} I(m)/\hat{I}(m) = 1$; for a proof see Wong (1989). Under slightly more restrictive conditions the proof follows from a form of Taylor's Theorem which says we can write $h(\theta) = h(\hat{\theta}) + \frac{1}{2}(\theta-\hat{\theta})'H(\hat{\theta})(\theta-\hat{\theta}) + \frac{1}{2}r(\theta)||\theta-\hat{\theta}||^2$ where r is continuous and $r(\theta)\to 0$ as $\theta\to\hat{\theta}$. Thus we are effectively approximating (1) by $(2\pi)^{k/2}|\tau H(\hat{\theta})|^{-1/2}\exp\{-\tau h(\hat{\theta})\}$ times the expectation of m with respect to the $N_k(\hat{\theta}, (\tau H(\hat{\theta}))^{-1})$ distribution for large τ. Intuitively Laplace's formula is saying that as $\tau\to\infty$; i.e. as the variance of the multivariate normal goes to 0, most of the contribution to $I(m)$ is occuring near the maximum of $\exp\{-\tau h(\theta)\}$. Under additional assumptions Wong (1989) establishes that the relative error in (2) is $O(\tau^{-1})$.

It is not always possible to transform an integral to the requisite form for the use of (2). In Bayesian contexts, however, we can write $h_\tau(\theta) = -\frac{1}{\tau}\log f(\theta)$ where $\tau\to\infty$ as the size of the data set grows. Recall that f depends implicitly on the data here. Strictly speaking we cannot apply Laplace's formula because of the dependence of h_τ on the data and of course on τ as well. As it is typical, however, for $h_\tau(\theta)$ to converge almost surely as $\tau\to\infty$, it is intuitively reasonable that (2) should still provide a good approximation. This was part of the insight in Lindley (1961, 1980). Applying Laplace's formula to $R(m) = I(m)/I(1)$ gives the approximation $\hat{R}_1(m) = \hat{I}(m)/\hat{I}(1) = m(\hat{\theta}_\tau)$ where $\hat{\theta}_\tau$ maximizes $-h_\tau$.

The use of Laplace's formula is closely related to approximating a distribution via a normal distribution. Chen (1985) provides conditions for the asymptotic normality of a sequence of posterior densities $\exp\{-\tau h_\tau(\theta)\}/\int\exp\{-\tau h_\tau(\theta)\}d\theta$. These conditions also establish the validity of

Laplace's formula for the inverse norming constant $I(1)$ and, with some additional mild conditions, also establish the validity of Laplace's formula for $I(m)$. The *asymptotic normality criterion* approximates $R(m)$ by the expectation of m using the $N_k(\hat{\theta}_\tau, (\tau H_\tau(\hat{\theta}_\tau))^{-1})$ distribution. We denote this approximation by $\hat{R}(m)$.

A formal asymptotic error order for $\hat{R}_1(m)$, and generalizations of this estimate, are obtained by performing the following operations on the integrals $I(m)$ and $I(1)$:

(i) formally expand m and h_τ in a Taylor series about $\hat{\theta}_\tau$,

(ii) make the transformation $\theta \to v = \left(\tau H_\tau(\hat{\theta}_\tau)\right)^{-1/2}(\theta - \hat{\theta}_\tau)$,

(iii) keep the first three terms of the expansion for h_τ in the exponent, factor this out, noting that $\left(\frac{\partial h_\tau(\hat{\theta}_\tau)}{\partial \theta_i}\right) = 0$, and expand the remaining exponential about 0,

(iv) express all expansions and products of expansions in powers of $\tau^{-1/2}$,

(v) formally evaluate the integrals by taking expectations term by term with respect to the $N_k(0, I)$ distribution and note that terms involving odd powers of $\tau^{-1/2}$ disappear,

(vi) formally expand the ratio of the expansions in powers of τ^{-1}.

To see clearly how these operations work the reader need only try a simple example. These operations give that the formal absolute error in $\hat{R}_1(m)$ is $O(\tau^{-1})$. When m is bounded away from 0 the relative error is also $O(\tau^{-1})$. An estimate with formal asymptotic error equal to $O(\tau^{-2})$ is obtained by taking the first two terms in the final expansion. This is equal to

$$\hat{R}_2(m) = \hat{R}_1(m) - \frac{1}{2\tau}\sum_{i,j}\left[h_\tau^{(i,j)}(\hat{\theta}_\tau)\left(\frac{\partial m(\hat{\theta}_\tau)}{\partial \theta_i}\sum_{r,s} h_\tau^{(r,s)}(\hat{\theta}_\tau)\frac{\partial^3 h_\tau(\hat{\theta}_\tau)}{\partial\theta_r\partial\theta_s\partial\theta_j} - \frac{\partial^2 m(\hat{\theta}_\tau)}{\partial\theta_i\partial\theta_j}\right)\right]$$

where $\left(H_\tau(\hat{\theta}_\tau)\right)^{-1} = \left(h_\tau^{(i,j)}(\hat{\theta}_\tau)\right)$. A similar analysis, that does not involve expanding m, establishes that the formal asymptotic error in $\hat{R}(m)$ is $O(\tau^{-1/2})$. It would appear that $\hat{R}_1(m)$ is a better approximation than $\hat{R}(m)$. On the other hand the Laplace approximation requires some regularity

conditions on the function m whereas nothing is required in the normal approximation beyond integrability; e.g. the Laplace approximation does not apply if m is an indicator function since it is not continuous.

In Tierney and Kadane (1986) this technique is modified for the case when $m > 0$ by taking $h^*_\tau(\theta) = -\frac{1}{\tau}[\log f(\theta) + \log m(\theta)]$ and finding θ^*_τ which maximizes $-h^*_\tau$. The first order estimate is

$$R^*(m) = \frac{|H^*_\tau(\theta^*_\tau)|^{-1/2} \exp\{-\tau h^*_\tau(\theta^*_\tau)\}}{|H_\tau(\hat{\theta}_\tau)|^{-1/2} \exp\{-\tau h_\tau(\hat{\theta}_\tau)\}}$$

and this is called the *fully exponential* Laplace approximation. Tierney and Kadane (1986) argue formally as above that, when m is bounded away from 0, the absolute and relative errors in $R^*(m)$ are $O(\tau^{-2})$ because of a cancellation between the numerator and denominator. The R^* approximation requires second order derivatives and two maximizations. This is a distinct advantage over $\hat{R}_2$ which requires third order derivatives. For a nonpositive m Tierney, Kass and Kadane (1989a) recommend using the fully exponential approach to approximate $R(\exp\{sm(\theta)\})$, the moment generating function of m, for several values of s about 0 and then numerically differentiating to get an estimate of $R(m)$.

Reviews of asymptotic methods in Bayesian contexts can be found in Kass, Tierney and Kadane (1988, 1991). Kass, Tierney and Kadane (1990) provide conditions for the validity of these approximations and asymptotic errors. Tierney, Kass and Kadane (1989b) use the Laplace approach to estimate the marginal density function of $\beta(\theta)$ for any function β having a Jacobian of full rank l on a neighbourhood of the posterior mode $\hat{\theta}_\tau$. When $l = 1$, DiCiccio, Field and Fraser (1990) establish asymptotic approximations for tail probabilities. Fraser, Lee and Reid (1990) discuss some asymptotic methods with particular relevance to Example 1. Morris (1988) provides an approach to asymptotic approximations in one dimension using the Pearson family. Discussions of other

i	$R(\theta_i)$	$\hat{R}(\theta_i)$	$\hat{R}_1(\theta_i)$	$R(\theta_i^2)$	$\hat{R}(\theta_i^2)$	$\hat{R}_1(\theta_i^2)$	$R^*(\theta_i^2)$
1	2.043	2.018	2.018	4.263	4.141	4.073	4.313
2	0.095	0.116	0.116	0.081	0.061	0.014	0.062
4	0.018	0.029	0.029	0.069	0.046	0.001	0.039
10	-0.073	-0.232	-0.232	0.033	0.079	0.054	0.033

Table 2: Asymptotic estimates for Example 1 where R denotes the exact value of the posterior expectation, $\hat{R}$ denotes the normal approximation, $\hat{R}_1$ denotes the Laplace approximation and R^* denotes the fully exponential Laplace approximation.

asymptotic methods with relevance for statistical applications can be found in Barndorff-Nielsen and Cox (1989, 1994) and Reid (1988).

We now apply some of these approximations to the examples of Section 2.

Example 1. Here we have that $\tau = n$ and recall from Section 2 that $n = 5$ in this example. In Table 2 we give the estimates. The $\hat{R}(\theta_i)$ and $\hat{R}_1(\theta_i)$ approximations are reasonable except for $i = 10$ where the relative error is 218%. A similar comment applies to $\hat{R}(\theta_i^2)$. The $\hat{R}_1(\theta_i^2)$ approximations are generally not as good as the relative errors for $i = 2, 4$ and 10 are 83%, 99% and 64% respectively. The $R^*(\theta_i^2)$ approximations are generally better except for $i = 4$. The maximizations were of necessity done numerically. We coded the first and second derivatives in this example and used the IMSL function optimization routine DUMIAH. The R^* computations were sensitive to the choice of initial value as the function maximized for the numerator is multimodal. For the last column of Table 2 we started DUMIAH at the posterior mode. In general multimodality can cause difficulties for all of the asymptotic methods.

Example 2. Here $\tau = 132$ is the total of the cell frequencies. We used the reparameterization discussed in Section 2 and further transformed to R^9 using the $N_9(0, I)$ inverse probability transform. We obtained the approximations $\hat{R}_1(\theta) = .402$, $\hat{R}_1(\theta^2) = .162$ and $R^*(\theta^2) = .186$. Recalling that $R(\theta) = .422$ and $R(\theta^2) = .181$ we see that these approximations perform well. We used what we believe to be the global maximum based on the results from several starting values for the

optimization routines. Calculating the approximations at other modes gave very different answers. We only coded first derivatives in this problem as coding second derivatives was exceedingly time consuming.

Example 3. For this example we could not find an approach using asymptotics that we felt was applicable. In particular the regularity conditions of Erkanli (1994) are not satisfied.

4. Importance Sampling and Variance Reduction Techniques

Suppose that w is a density function on R^k such that $\text{supp}(f) \subseteq \text{supp}(w)$ where $\text{supp}(f)$ is the closure of the set of points where f is nonzero, and $\theta_1, \ldots, \theta_N$ is a sample from w. Then since $I(m) = \int_{R^k} \frac{m(\theta)f(\theta)}{w(\theta)} w(\theta)d\theta$, (1) can be estimated by $\hat{I}_w(m) = \frac{1}{N}\sum_{i=1}^{N} \frac{m(\theta_i)f(\theta_i)}{w(\theta_i)}$. The justification for this estimate lies in the Strong Law of Large Numbers which implies that $\hat{I}_w(m) \to I(m)$ almost surely as $N \to \infty$. The density w is referred to as an *importance sampler*. The basic idea is to choose w so that it generates values that are in the region where the integrand is large as this region is where the most important contributions are made to the value of the integral. For general discussions on importance sampling see Hammersley and Handscomb (1964), van Dijk, Kloek and Louter (1986), Glynn and Iglehart (1989), Hesterberg (1990) and Wolpert (1991). Objections to the use of Monte Carlo methods have been raised by O'Hagan (1987).

We choose w so that the rate of convergence of $\hat{I}_w(m)$ to $I(m)$ is as fast as possible. The Law of the Iterated Logarithm, see Durrett (1991), suggests that we can do no better than require that $\hat{I}_w(m)$ have as small a variance as possible. This is equivalent to minimizing

$$\text{Var}_w\left[\frac{mf}{w}\right] = \int \frac{m^2(\theta)f^2(\theta)}{w(\theta)}d\theta - \left(\int m(\theta)f(\theta)d\theta\right)^2. \tag{3}$$

By the Central Limit Theorem $\hat{I}_w(m) \sim AN\left(I(m), N^{-1}\text{Var}_w[mf/w]\right)$ when (3) is finite and $X_N \sim AN(\mu, \sigma^2/N)$ means that $\sqrt{N}(X_N - \mu)/\sigma$ converges in distribution to a $N(0,1)$ random

variable as $N \to \infty$. If we put $q_m(\theta) = |m(\theta)|f(\theta)/\int |m(\theta)|f(\theta)d\theta$ then we have $\mathrm{Var}_w\left[\frac{mf}{w}\right] = I^2(|m|)\,\mathrm{E}_w\left[\left(\frac{q_m - w}{w}\right)^2\right] + (I^2(|m|) - I^2(m))$. Therefore (3) is uniquely minimized when $w = q_m$ since then the *Chisquare distance* between q_m and w is 0; i.e. $\mathrm{E}_w\left[\left(\frac{q_m - w}{w}\right)^2\right] = 0$. If $I(m) \neq 0$ then a useful representation of the Chisquare distance, in terms of the coefficient of variation of mf/w, is given by

$$\mathrm{E}_w\left[\left(\frac{q_m - w}{w}\right)^2\right] = \left(\frac{I(m)}{I(|m|)}\right)^2 \mathrm{CV}_w^2\left[\frac{mf}{w}\right] + \left(\frac{I(m)}{I(|m|)}\right)^2 - 1.$$

We assess the effectiveness of w for $\int |m(\theta)|f(\theta)d\theta$, rather than for (1), as these integrals have the same optimal importance sampler. Writing $W_i = |m(\theta_i)|f(\theta_i)/w(\theta_i)$ and $W_i^* = W_i/(W_1 + \ldots + W_N)$ as the i-th normalized sample weight, it is easy to show that the method of moments estimator of $\mathrm{CV}_w\left[\hat{I}_w(|m|)\right]$ is given by $\frac{1}{\sqrt{N}}\left\{N\sum_{i=1}^N W_i^{*2} - 1\right\}^{1/2}$. Therefore we have that $\frac{1}{N} \leq \sum_{i=1}^N W_i^{*2} \leq 1$, with the lower bound attained if and only if all the weights are equal and the upper bound attained if and only if all but one of the weights is 0. From this we see that quoting the estimated coefficient of variation to assess accuracy is equivalent to examining the normalized weights to determine if any are relatively large. Rather than the coefficient of variation appearing to be too large, however, the more frequent, and dangerous, phenomenon is for it to be too low for small N. This occurs when w samples only over a sub-region where the integrand is nearly constant.

The optimal w is typically not a possible choice because it requires that we be able to sample from this density and have a closed form expression. In general we are restricted to using those w for which efficient generating algorithms and exact expressions exist. In fact there is no automatic algorithm we can use to generate a sample from an arbitrary multidimensional distribution let alone do this efficiently; see Devroye (1986) for a thorough discussion of random variate generation. In general, however, we try to choose w to agree with $|m|f$ as closely as possible.

There are very few families of distributions in the multidimensional context that are in common

use as importance samplers. When the problem is to integrate over R^k virtually the only family used is the multivariate Student family. This family has unimodal densities with ellipsoidal contours and varying tail-lengths. If we choose to mimic f, and this function has approximately these characteristics, then we take w as the density of $\theta = \hat{\theta} + C\omega$ where $\hat{\theta}$ is the mode of f, C is the Cholesky factor of the inverse Hessian of $-\log f$ at the mode and $\omega \sim \text{Student}_k^*(\zeta)$ for some $\zeta > 2$ where $\text{Student}_k^*(\zeta)$ denotes a standardized $\text{Student}_k(\zeta)$ distribution. In Bayesian contexts this implies that w and the posterior have approximately the same mean and variance. The parameter ζ is chosen low enough so that the estimate $\hat{I}_w(m)$ has finite variance. Typically it is not possible to say exactly what ζ has to be to guarantee this. For an exception to this see Evans and Swartz (1994).

There is a great need to develop families of multivariate distributions that exhibit a wide variety of shapes, have efficient algorithms for random variate generation and can be easily fitted to a specific integrand. Geweke (1989) developed the split-t family by modifying the $\text{Student}_k(\zeta)$ distribution, via rescaling the density for ω differentially along each of the $2k$ rays from 0 following a coordinate direction, to allow for skewness. The tail behaviour of this w is the same as a $\text{Student}_k(\zeta)$ distribution but it has a discontinuity at the mode. A closely related generalized Student family was developed in Evans (1988) by joining $\text{Student}_1(\zeta_1)$ and $\text{Student}_1(\zeta_2)$ densities at the origin by a rescaling and standardizing so that the distribution has mean 0 and variance 1. In k dimensions we take the product of such densities. This family allows for different tail-lengths along each ray from the mode and is everywhere differentiable, but has the disadvantage of having much shorter tails along rays from the mode that lie between coordinate directions. Oh and Berger (1993) modify Student importance sampling by allowing for mixtures of multivariate Students. For example, if f has modes at $\hat{\theta}(1), \ldots, \hat{\theta}(c)$ and the Cholesky factor of the inverse Hessian of $-\log f$ at $\hat{\theta}_i$ is given by C_i then we take $w = \epsilon_1 w_1 + \ldots + \epsilon_c w_c$ where w_i is the density of $\hat{\theta}(i) + C_i\omega_i$ with $\omega_i \sim \text{Student}_k^*(\zeta_i)$

and the ϵ_i are mixture probabilities determined by the requirement that $\epsilon_i \propto f(\hat{\theta}(i))/w_i(\hat{\theta}(i))$ for $i = 1, \ldots, c$. The difficulty with the use of mixtures lies in determining the parameters of the mixture when the modes are unknown.

The estimate $\hat{R}_w(m) = \hat{I}_w(m)/\hat{I}_w(1)$ converges almost surely to $R(m)$. If $R(m) \neq 0$ and $\hat{I}_w(m)$ and $\hat{I}_w(1)$ have finite variances, then the delta method gives that $\hat{R}_w(m)$ is asymptotically normal with asymptotic coefficient of variation equal to

$$\frac{\mathrm{sgn}(R(m))}{\sqrt{N}} \left\{ \mathrm{CV}_w^2 \left[\frac{mf}{w}\right] + \mathrm{CV}_w^2 \left[\frac{f}{w}\right] - 2\mathrm{CV}_w \left[\frac{mf}{w}\right] \mathrm{CV}_w \left[\frac{f}{w}\right] \mathrm{Corr}_w \left[\frac{mf}{w}, \frac{f}{w}\right] \right\}^{1/2}$$

and this is estimated in the obvious way. An alternative estimator of $R(m)$ can be obtained as follows using sampling/importance resampling or SIR; see Rubin (1988) and Smith and Gelfand (1992). Let $W_1^*, \ldots, W_N^*$ be the normalized weights obtained from estimating $I(1)$ based on a sample $\theta_1, \ldots, \theta_N$ from w. Then these weights can be used to approximately sample from the density $f/\int f(\theta)d\theta$ when N is large. For this we resample $N^* \ll N$ values $\theta_1^*, \ldots, \theta_{N^*}^*$ from $\{\theta_1, \ldots, \theta_N\}$ using the W_i^* probabilities. We can then estimate $R(m)$ by $\frac{1}{N^*}\sum_{i=1}^{N^*} m(\theta_i^*)$ but this estimator always has variance at least as large as $\hat{R}_w(m)$.

There are many variance reduction techniques that can be used in conjunction with importance sampling. This is an extensive area of research and so we briefly discuss just control variates and systematic sampling. Discussion of a wider class of techniques can be found in Hammersley and Handscomb (1964), Powell and Swann (1966), Cranley and Patterson (1970) and Ripley (1987).

The use of *control variates* requires that there be a closely related integral whose value is known. As an example of this, suppose that we have a function g such that $\int m(\theta)g(\theta)d\theta$ is known. For a single θ generated from w we estimate (1) by $m(\theta)\frac{(f(\theta)-g(\theta))}{w(\theta)} + \int m(\theta)g(\theta)d\theta$ and this has variance $\int m^2(\theta)\frac{(f(\theta)-g(\theta))^2}{w(\theta)}d\theta - (I(m) - \int m(\theta)g(\theta)d\theta)^2$. With a good choice of g this will be much smaller

than (3). A sensible choice for g in Bayesian contexts will be the normal approximation to the posterior times the Laplace estimate $\hat{I}(1)$. We refer to this as the *asymptotic normality control variate*. Using this control variate when the posterior is asymptotically normal leads to an estimate of $R(m)$ with asymptotic variance 0.

For *systematic sampling*, or antithetic variates, let $T_i : \text{supp}(f) \to \text{supp}(f)$ for $i = 1, \ldots, t$ be one-to-one transformations with Jacobian determinants at θ denoted by $J_{T_i}(\theta)$. Then a systematic sampling estimator for (1), corresponding to the set $\mathcal{T} = \{T_1, \ldots, T_t\}$, is obtained by generating θ from w and calculating $(mf)^{\mathcal{T}}(\theta)/w(\theta) = t^{-1} \sum_{i=1}^{t} m(T_i(\theta))f(T_i(\theta))J_{T_i}(\theta)/w(\theta)$. This formulation generalizes the antithetic variable techniques presented in Hammersley and Handscomb (1964), Fishman and Huang (1983) and Geweke (1988). Note that when $t = 1$ the choice of $\mathcal{T}$ is essentially the problem of choosing an appropriate parameterization. *Latin hypercube sampling* is closely related to systematic sampling. For a discussion of this see McKay, Conover and Beckman (1979), Stein (1987), Owen (1992) and Tang (1993).

We now use importance sampling on the examples of Section 2.

Example 1. We used Student importance sampling with the mode and Hessian computed in Section 3 and $\zeta = 5$. The results are reported in Table 3 for a Monte Carlo sample size of $N = 10^5$. These computations took about 50 seconds of CPU time. The estimates are accurate but the coefficient of variation for $\hat{R}_w(\theta_4)$ is quite high. We tried importance sampling using different choices of $\zeta \in (2, 20)$ but the differences in the efficiencies were not large. We also computed these integrals using $N = 8 \times 10^7$ iterations and have taken these throughout as the exact values. We feel confident that these results have 3 decimal places of accuracy. We implemented the asymptotic normality control variate but this resulted in a reduction in the variance of the estimates of the norming constants by only 3.2%. We note that $I(1) = .131 \times 10^{-21}$ is poorly approximated by $\hat{I}(1) = .527 \times 10^{-22}$.

i	$R(\theta_i)$	$\hat{R}_w(\theta_i)$	$R(\theta_i^2)$	$\hat{R}_w(\theta_i^2)$
1	2.043	2.043 (.001)	4.263	4.265 (.002)
2	0.095	0.096 (.023)	0.081	0.080 (.014)
4	0.018	0.017 (.127)	0.069	0.070 (.015)
10	-0.073	-0.073 (.017)	0.033	0.033 (.008)

Table 3: Importance sampling estimates and estimated absolute coefficients of variation for Example 1 where R denotes the exact value of the posterior expectation and $\hat{R}_w$ denotes the importance sampling estimate.

Example 2. For this example we used the parameterization, mode and Hessian of Section 3 and Student importance sampling with $\zeta = 5$. Based on a sample of $N = 10^5$ which took 2.5 minutes of CPU time, the estimates together with their estimated absolute coefficients of variation are $\hat{R}_w(\theta) = .423(.003)$ and $\hat{R}_w(\theta^2) = .182(.007)$. The exact values $R(\theta) = .422$ and $R(\theta^2) = .181$ are based on a sample of $N = 10^8$ which took 41 hours of CPU time.

Example 3. For this example we implemented the obvious importance sampling algorithm based on the $N_6(0, \Sigma)$ distribution where Σ is specified in Section 2. Sometimes this is called the *hit-or-miss* estimator. For $N = 10^6$ the estimate and its estimated absolute coefficient of variation equal $\hat{I}_w(m) = .190000 \times 10^{-4}(.229)$ and this took 1 minute of CPU time. For $N = 10^8$ we obtained $\hat{I}_w(m) = .163000 \times 10^{-4}(.025)$ and this took about 100 minutes of CPU time. In Evans and Swartz (1988a) a much more accurate importance sampling approach was developed and this method was used to obtain the exact answer $I(m) = .166625 \times 10^{-4}$.

5. Adaptive Importance Sampling

Throughout this section we assume $m \equiv 1$ for convenience. For a given problem there is typically a class $\mathcal{W} = \{w_\alpha \mid \alpha \in \mathcal{A}\}$ of candidate importance samplers and we must select one. For example, $\mathcal{W}$ could be the set of all distributions on R^k obtained from $\mu + \Sigma^{1/2}\omega$ where $\mu \in R^k$, $\Sigma \in R^{k \times k}$ is positive definite, $\omega \sim \text{Student}_k^*(\zeta)$ for some fixed ζ and thus $\alpha = (\mu, \Sigma)$ is $k + k(k+1)/2$ dimensional. The optimal choice of α is the value minimizing $v(\alpha) = \int f^2(\theta)/w_\alpha(\theta) d\theta$. Of course evaluating $v(\alpha)$,

let alone minimizing this function, is at least as hard as calculating (1). The compromise suggested in Section 4 is to select α so that w_α and f have some characteristics in common. For example, with Student importance sampling we chose μ and Σ so that w_α has mean and variance equal to the approximate mean and variance of $p = f/\int f(\theta)d\theta$. The general principle then is to approximate characteristics of p and choose $w \in \mathcal{W}$ so that its corresponding characteristics match.

There is no reason why we have to restrict ourselves to a single $w \in \mathcal{W}$. For if we have initially selected w_{α_1} then we can estimate characteristics of p, that can be represented as expectations, using a sample $\theta_1, \ldots, \theta_N$ from w_{α_1}. This information can then be used to make an improved choice w_{α_2}. To be more precise, let $C : R^k \to R^l$ and let $R(C) = \mathrm{E}_p[C]$ be a vector of such characteristics of p. Then estimate $R(C)$ by $\hat{R}_{(1)}(C) = \hat{I}_{w_{\alpha_1}}(C)/\hat{I}_{w_{\alpha_1}}(1)$ and obtain $w_{\alpha_2} \in \mathcal{W}$ by minimizing $||\mathrm{E}_{w_\alpha}[C] - \hat{R}_{(1)}(C)||$. For example, if $\mathcal{W}$ is the Student family described above then a sensible choice for C is the vector giving rise to means, variances and covariances. In this case the minimization problem can be easily solved. If the estimate $\hat{R}_{(1)}(C)$ is reasonable then p and w_{α_2} should be more alike than p and w_{α_1}. We then generate $\theta_{N+1}, \ldots, \theta_{2N}$ from w_{α_2}, compute the new estimate $\hat{R}_{(2)}(C) = (\hat{I}_{w_{\alpha_1}}(C) + \hat{I}_{w_{\alpha_2}}(C))/(\hat{I}_{w_{\alpha_1}}(1) + \hat{I}_{w_{\alpha_2}}(1))$ and minimize $||\mathrm{E}_{w_\alpha}[C] - \hat{R}_{(2)}(C)||$ to obtain w_{α_3}. This process is continued until the estimates $\hat{R}_{(i)}(C)$ change very little, say at the n-th iteration, whence we take w_{α_n} as the importance sampler for subsequent simulations. Note that we have suppressed the dependence of α_i on $\theta_1, \ldots, \theta_{(i-1)N}$. We call this algorithm *adaptive importance sampling by matching characteristics*. Several authors have considered algorithms closely related to this; see, for example, Kloek and van Dijk (1978), Smith, Skene, Shaw and Naylor (1987), Naylor and Smith (1988), Evans (1988, 1991a), Oh (1991) and Oh and Berger (1992).

Typically we want to compute many of the components of $R(C)$ so this approach can be viewed as a way of simultaneously improving the importance sampler and estimating quantities of interest. Of course the effectiveness of the algorithm is strongly influenced by the choice of C. While we have

no formal theory to determine C, using means, variances, covariances and probability contents for central regions is preferable to choosing tail-sensitive measures such as high-order moments.

The analysis of convergence for adaptive importance sampling is more difficult than importance sampling because of the dependence between iterations. Almost sure convergence of $\hat{R}_{(n)}(C)$ to $R(C)$, almost sure convergence of α_n to a unique $\alpha_* \in \mathcal{A}$, an asymptotic normality result for $\hat{R}_{(n)}(C)$ and a consistent estimator for the asymptotic variance were all established in Oh and Berger (1992). However, these results require the uniform boundedness of the weight function; namely, there must exist B such that $f(\theta)/w_\alpha(\theta) \leq B$ for every α and θ, and this almost never holds in an application. It is possible to state much weaker conditions under which these results will hold.

Of some practical importance is the estimation of the asymptotic variance. Assuming for convenience that $l = 1$ it can be shown under weak conditons that $\hat{R}_{(n)}(C) \sim AN(R(C), \sigma^2/nN)$ where σ^2 equals $R^2(C)$ times

$$\mathrm{CV}^2_{w_{\alpha_*}}\left[\frac{Cf}{w_{\alpha_*}}\right] + \mathrm{CV}^2_{w_{\alpha_*}}\left[\frac{f}{w_{\alpha_*}}\right] - 2\mathrm{CV}_{w_{\alpha_*}}\left[\frac{Cf}{w_{\alpha_*}}\right]\mathrm{CV}_{w_{\alpha_*}}\left[\frac{f}{w_{\alpha_*}}\right]\mathrm{Corr}_{w_{\alpha_*}}\left[\frac{Cf}{w_{\alpha_*}}, \frac{f}{w_{\alpha_*}}\right]. \qquad (4)$$

An almost sure consistent estimate of (4) is obtained by estimating $E_{w_{\alpha_*}}[mf/w_{\alpha_*}]$ by $\frac{1}{n}\sum_{i=1}^{n} \hat{I}_{w_{\alpha_i}}(m)$, for $m = 1$ and $m = C$, where n is the number of adaptive steps and $\hat{I}_{w_{\alpha_i}}(m)$ is the importance sampling estimate based on the i-th importance sampler. We then substitute these expressions into (4).

One difficulty with the adaptive approach is the necessity of obtaining a good starting value $\alpha_1 \in \mathcal{A}$. With a poor choice convergence will not take place for a practical number of iterations. The technique of *chaining* can be useful in dealing with this problem, see Evans(1988, 1991a, 1991b). For this suppose we have an additional parameter $\eta \in \mathcal{N}$ such that $f = f_{\eta_*}$ and a good starting

importance sampler w_{α_1} for the problem $\int f_{\eta_1}(\theta)d\theta$. We run adaptive importance sampling for this problem and obtain the updated importance sampler w_{α_2}. We then make a small change in η to η_2 and run adaptive importance sampling for $\int f_{\eta_2}(\theta)d\theta$, starting with w_{α_2} and this leads to the updated importance sampler w_{α_3}. We continue this process constructing a chain from η_1 to η_* where η_* specifies the problem of interest. In a general problem we can chain on $\eta = (\eta_1, \eta_2) \in R^2$ where $f_\eta = f^{1/\eta_1} w_{\alpha_1}^{1/\eta_2}$. For large η_1 and $\eta_2 = 1$, f_η is like w_{α_1}. For large η_2, f_η is like f. We mention a simple use of chaining in Example 1.

There are other approaches to adaptive importance sampling. A very natural approach is to use Monte Carlo to try to minimize $v(\alpha)$ or at least obtain a relatively small value. There are several papers where this approach has been considered; see, for example, Oh and Berger (1993) and Piccioni and Ramponi (1993). In West (1993) and Givens and Raftery (1993) another approach is taken using ideas from kernel density estimation. This results in importance samplers that are mixtures of many multivariate Student densities.

We now consider the application of adaptive importance sampling to the examples.

Example 1. For this example we took $\mathcal{W}$ to be the Student family with $\zeta = 5$. For $R(C)$ we used the vector of posterior means, variances and covariances. We started with the importance sampler of Section 4 and used a sample of $N = 1000$ from each w_{α_i} for $i = 1, \ldots, n$ adaptive steps. Table 4 contains the results at the end of the 100-th iteration. These computations took about 50 seconds of CPU time and thus the adaptation added no real computational burden. There is a clear benefit to adaptation as the estimated absolute coefficients of variation are lower than in Table 3. This is particularly noteworthy for $i = 4$ where the reductions exceed 50%. We note that in general, with the model of this example, it will make more sense to treat λ as a nuisance parameter varying in $(2, \infty)$ rather than fixed at some value. In this case it makes sense to use the marginal density of the data; i.e. the inverse norming constant, to select a value for λ where the inverse norming constant

i	$R(\theta_i)$	$\hat{R}_{(100)}(\theta_i)$	$R(\theta_i^2)$	$\hat{R}_{(100)}(\theta_i^2)$
1	2.043	2.042 (.001)	4.263	4.262 (.001)
2	0.095	0.093 (.013)	0.081	0.080 (.007)
4	0.018	0.020 (.057)	0.069	0.070 (.007)
10	-0.073	-0.073 (.011)	0.033	0.033 (.008)

Table 4: Adaptive importance sampling estimates and estimated absolute coefficients of variation for Example 1 where R denotes the exact value of the posterior expectation and $\hat{R}_{(100)}$ denotes the adaptive importance sampling estimate.

is high. Also for λ large it easy to get good starting values for the posterior mean and variance matrix using least-squares theory and this avoids the need to numerically optimize. Therefore we can chain on the parameter λ by starting with λ large and then reducing it it in stages to obtain estimates of the inverse norming constant as a function of λ.

Example 2. We used the same algorithm as in Example 1, with appropriate adjustments, but took $N = 20,000$ and $n = 5$. We obtained the estimates and estimated absolute coefficients of variation $\hat{R}_{(5)}(\theta) = .421(.002)$ and $\hat{R}_{(5)}(\theta^2) = .180(.003)$. Comparing with the results of Section 4 adaptation has resulted in increased accuracy by about 33-50%. For other variables the increase in accuracy was even greater.

Example 3. For this example we used the parameterization discussed in Section 2 followed by the $N_5(0, I)$ inverse probability transform. We then used adaptive importance sampling with a multivariate Student based on 5 degrees of freedom adapting to the mean and variance. We started the adaptive importance sampling by computing the mode and the inverse Hessian at the mode of the transformed integrand. By necessity all derivatives were computed numerically because of the complexity of the expressions. Based on $N = 10^4$ and $n = 10$ adaptive steps we obtained $\hat{I}_{(10)}(m) = .166672 \times 10^{-4}$ with coefficient of variation .002. This took .9 minutes of CPU time. This is an improvement over hit-or-miss by a factor of 366.

6. Multiple Quadrature and Subregion Adaptive Integration

First we discuss *quadrature rules*; i.e. rules for dimension $k = 1$. A quadrature rule of order n, for approximating (1), takes the form $\sum_{i=1}^{n} w_i h(\theta_i)$ where $h = mf/w$, w is a density function and the points θ_i and weights w_i are determined by some criterion. As with importance sampling w is chosen so that mf/w or f/w is approximately constant. Given that we have selected distinct points $\theta_1, \ldots, \theta_n$ a common method of determining the w_i is to integrate the product of w and the degree $n-1$ Lagrange interpolating polynomials corresponding to the θ_i, assuming of course that w has its first $n-1$ moments; i.e. $w_i = \int_{-\infty}^{\infty} p_i(x)w(x)dx$ where $p_i(x) = \prod_{j \neq i}(x - \theta_j)/\prod_{j \neq i}(\theta_i - \theta_j)$. Such a rule is called an *interpolatory rule* and it calculates the integral exactly whenever h is a polynomial of degree less than n. For example, when w is the uniform density on some interval and the θ_i are equispaced and include the end-points, we get the trapezoid rule for $n = 2$ and Simpson's rule for $n = 3$. Rules that are based on a density w with a finite interval for support are often *compounded*; i.e. the interval is divided into subintervals and the rule is applied in each subinterval with the appropriate location and scale changes being made to w. It can be shown that, if the basic rule integrates constants exactly and the maximum length of a subinterval goes to 0, then the compounded rule converges to (1) as the number of subintervals increases. An optimal choice of the n points can be made so that the rule exactly integrates all polynomials up to the maximal degree $2n - 1$ provided that w has $2n$ moments. These are called *Gauss rules*. When w is the $N(0, 1)$ density we get the Hermite rules, when w is the Gamma(α) density we get the Laguerre rules and when w is the Beta(α, β) density, after a location-scale transformation to the interval $[-1, 1]$, we get the Jacobi rules. In general it is not an easy numerical problem to obtain these rules accurately but for the common densities they are tabulated in software libraries such as IMSL and NAG. Davis and Rabinowitz (1984) gives a thorough treatment of quadrature rules.

The simplest method of constructing *multiple quadrature rules*; i.e. $k > 1$, is to form *product rules*. If $h(\theta_1, \ldots, \theta_k) = m(\theta_1, \ldots, \theta_k)f(\theta_1, \ldots, \theta_k)/w_1(\theta_1) \cdots w_k(\theta_k)$, where the w_i are densities,

then a product rule approximates (1) by

$$\sum_{i_1=1}^{n_1} \cdots \sum_{i_k=1}^{n_k} w_{1,i_1} \cdots w_{k,i_k} h(\theta_{1,i_1}, \ldots, \theta_{k,i_k}) \tag{5}$$

where the θ_{i,j_i} and w_{i,j_i} are the points and weights of a quadrature rule associated with w_i. Such rules suffer from the *curse of dimensionality*; namely to implement this rule requires $N = n_1 \cdots n_k$ function evaluations. Further it can be shown that if we take all the rules in (5) to be the compounded trapezoid rule then the error in (5) is $O(1/N^{2/k})$. It would appear that the product trapezoid rule is not competitive with importance sampling whenever $k > 4$.

An integrand typically will not be well-approximated by a $N_k(0, I)$ density times a low degree polynomial. Asymptotics in Bayesian problems suggest, however, that this will often be the case if we standardize the posterior density by the posterior mean and variance. Naylor and Smith (1982), Smith, Skene, Shaw, Naylor and Dransfield (1985) and Naylor and Smith (1988) use an adaptive approach which computes an approximation to the posterior means, variances and covariances, using product Hermite rules. They then transform the integrand using these quantities, so that the transformed posterior has approximate mean 0 and variance matrix I. This is iterated until the process stabilizes. In general, successful application of multiple quadrature requires that the integrand be appropriately located, scaled and not have the bulk of its mass concentrated near a hyperplane.

Various attempts have been made to avoid the curse of dimensionality. Hammersley (1960) suggests estimating (5) by sampling the terms in this sum. When each rule in the product is a Gauss rule then $w_{i,j_i} > 0$ and $w_{i,1} + \cdots + w_{i,n_i} = 1$ for $i = 1, \ldots, k$. Then we can sample the terms of (5) using the weights as a discrete probability distribution. Evans and Swartz (1988b) give an extensive analysis of this technique including a comparison with the analogous importance

sampling method. It is shown that accurate approximations can be obtained in samples where N is much smaller than $n_1 \cdots n_k$.

Another attempt at avoiding the curse is to construct rules in R^k with a minimal number of points that exactly calculate $\int_{R^k} \theta_1^{i_1} \cdots \theta_k^{i_k} w(\theta_1, \ldots, \theta_k) d\theta_1 \ldots d\theta_k$ for a class of monomials satisfying $i_1 + \cdots + i_k \leq d$ where w is a density. Such a rule is called a *monomial rule* of degree d for w. *Fully symmetric* monomial rules are obtained, when w is invariant under permutations and sign changes of the coordinates; e.g. the uniform density on $[-1, 1]^k$, by performing permutations and sign changes on a set of generators to obtain the points and then solving a system of polynomial equations for the generators and weights. Fully symmetric rules with the minimal or close to the minimal number of points have been obtained for various w. These are in general not easy problems to solve and sometimes the rules suffer from having some of the points lying outside $\text{supp}(w)$. Dellaportas and Wright (1991, 1992) discuss a related approach using fully symmetric rules constructed from subsets of the points in a product Hermite rule. Davis and Rabinowitz (1984), Stroud (1971) and Cools and Rabinowitz (1993) are good references on multiple quadrature rules.

To assess the error in a multiple quadrature approximation the calculation is typically repeated with a rule containing more points. Sometimes there are guaranteed error estimates available involving bounds on derivatives of the integrand. These are generally felt to be far too conservative and difficult to implement.

There are other approaches to constructing multiple quadrature rules. *Bayesian quadrature* is discussed in Diaconis (1985) and O'Hagan (1992). *Quasirandom rules* are discussed in Wozniakowski (1991), Traub and Wozniakowski (1992), Niederreiter (1992) and Spanier and Maize (1994). Statistical applications of quasirandom rules are presented in Shaw (1988), Johnson (1992) and Fang, Wang and Bentler (1994). While it is sometimes argued that the quasirandom approach is superior to Monte Carlo the flexibility of the latter method still provides substantial advantages

in our opinion.

Subregion adaptive algorithms are described in van Dooren and de Ridder (1976), Genz and Malik (1980), Bernsten, Espelid and Genz (1991) and Genz (1991, 1992b). For low dimensional contexts they are greatly recommended as they produce accurate answers over a wide class of problems and are highly efficient. Related algorithms are discussed in Lepage (1978) and Friedman and Wright (1981). For these algorithms the problem is transformed so that the domain of integration is $[0,1]^k$. The basic algorithm then proceeds iteratively as follows: let $\epsilon > 0$ be the global absolute error desired and let n_{max} be the maximum number of subregions where we start with one region $R_{11} = [0,1]^k$. At the n-th step $[0,1]^k$ has been partitioned into n subregions $R_{n1}, \ldots, R_{nn}$. Multiple quadrature rules have been applied in each subregion to get estimates $I_{n1}, \ldots, I_{nn}$ of the corresponding integrals and also error estimates $E_{n1}, \ldots, E_{nn}$. The error estimate E_{ni} is obtained by computing I_{ni} using rules of different orders and comparing the results. If $E_{n1} + \cdots + E_{nn} < \epsilon$ or if $n \geq n_{max}$, the algorithm stops. Otherwise the region R_{ni} is split if $E_{ni} = \max\{E_{nj} \mid 1 \leq j \leq n\}$. The partitioning algorithm takes the regions to be rectangles with sides parallel to the coordinate axes and uses a criterion to choose which side of the chosen rectangle to split to form the subrectangles. Different criteria can be used but they are all based on selecting the coordinate direction, in the chosen subregion, where the integrand is most variable. Subregion adaptive integration has been applied to a number of problems in statistics. See, for example, Genz and Kass (1991, 1994) and Genz (1992b).

We now apply subregion adaptive integration to the examples of Section 2. For these computations we used the routine ADBAYS supplied by A. Genz. This algorithm uses fully symmetric rules of orders 5, 7 and 9 for subregion estimates and for error estimates; see Genz and Malik (1983) for a discussion of the particular rules. Further a criterion based on fourth differences is used to decide how to split a chosen subregion. The program runs until a specified relative error is attained or

i	$R(\theta_i)$	$\hat{R}(\theta_i)$	$R(\theta_i^2)$	$\hat{R}(\theta_i^2)$
1	2.043	2.040	4.263	4.237
2	0.095	0.094	0.081	0.068
4	0.018	0.017	0.069	0.055
10	-0.073	-0.102	0.033	0.032

Table 5: Subregion adaptive estimates for Example 1 where R denotes the exact posterior expectation and $\hat{R}$ denotes the subregion adaptive estimate.

too many function evaluations have been carried out.

Example 1. Table 5 gives the results of using subregion adaptive integration. Following Genz and Kass (1991, 1994) and Genz (1992b) we first made the transformation $\theta \to u$ where $\theta = \hat{\theta} + Cu$, $\hat{\theta}$ is the mode and C is the lower triangular Cholesky factor of the inverse Hessian computed in Section 3. We then made the transformation $u \to v$ where $v = \Phi(u)$ and Φ denotes the $N(0,1)$ distribution function applied component-wise to u. We specified a relative error of .001 and set the maximum number of function evaluations to 2×10^7. This maximum number was reached without the error bound being attained. The computation took 2.75 hours. We repeated the calculations allowing for a maximum of 10^8 function evaluations with the same relative error bound. The results improved somewhat; e.g. the relative error in the estimate of $R(\theta_{10})$ went from 40% to 33%, but the computation time increased to 24 hours.

Example 2. For this example we made the transformation discussed in Section 5, calculated the mode and inverse Hessian matrix corresponding to the transformed posterior, standardized and then made the $N_9(0, I)$ probability transform. Then requesting a relative error of .1 we ran ADBAYS. This computation required 5×10^6 function evaluations and 3.6 hours of CPU time. The approximations and relative errors are given by $\hat{R}_{sa}(\theta) = .419(.007)$ and $\hat{R}_{sa}(\theta^2) = .178(.017)$. In this example the algorithm was successful but required a considerable amount of computation.

Example 3. For this example we used the transformation discussed in Section 2. Based on $N = 10^5$

function evaluations the approximation and its relative error are given by $\hat{I}(m) = .166626 \times 10^{-4}(.6 \times 10^{-6})$ and this required .2 minutes of CPU time. For $N = 10^6$ the approximation was exact to 10 decimal places and required 3.4 minutes of CPU time.

7. Markov Chain Methods

Another class of Monte Carlo algorithms has come into prominence in statistical applications during the past several years. These are collectively known as *Markov chain* methods. The characterizing feature of this approach is the construction of an ergodic Markov chain $\theta_1, \theta_2, \ldots$ on $\text{supp}(f)$ with unique stationary distribution having density $p = f / \int f(\theta) d\theta$. Thus this approach provides a method of approximately simulating from the probability distribution p. A strong law then gives the almost sure convergence of $\hat{I}_p(m) = \frac{1}{N} \sum_{i=1}^{N} m(\theta_i)$ to $I(m)$ as $N \to \infty$. There are many problems in statistics where the construction of an appropriate Markov chain is almost immediate. Some of these problems are so high-dimensional that it is hard to conceive of any other method working. Also, in certain examples, the convergence is very rapid. General references on Markov chain algorithms include Hammersley and Handscomb (1964), Tierney (1991), Geyer and Thompson (1992), Besag and Green (1993), Smith and Roberts (1993), Neal (1993) and Besag, Green, Higdon and Mengersen (1995).

Markov chain methods are generally related to the *Metropolis algorithm*. The basic version of this was first developed in Metropolis, Rosenbluth, Teller and Teller (1953). We present a generalization due to Hastings (1970); see also Peskun (1973). For this we specify an initial time homogeneous Markov chain on $\text{supp}(f)$ with transition density functions $r(\theta, \cdot)$ for each $\theta \in \text{supp}(f)$ and an initial state $\theta_0 \in \text{supp}(f)$ which may be random. A new Markov chain is then generated as follows: given that we are in state θ_n at time n, generate θ from $r(\theta_n, \cdot)$ and with *acceptance probability* $\alpha(\theta_n, \theta) = \min\{f(\theta) r(\theta, \theta_n) / f(\theta_n) r(\theta_n, \theta), 1\}$ set $\theta_{n+1} = \theta$ otherwise set $\theta_{n+1} = \theta_n$. See Tierney (1991) for a discussion of conditions on r necessary for p to be the unique stationary

distribution of this chain. The original Metropolis algorithm required that r be symmetric which implies $\alpha(\theta_n, \theta) = \min\{f(\theta)/f(\theta_n), 1\}$. Therefore a transition is made with high probability if θ lies in a region of high density for p relative to θ_n.

A user of the Metropolis algorithm must select an r. We want an r such that the chain wanders thoroughly and rapidly around the support of f. In general it is not clear how to make this choice. Tierney (1991) provides some suggestions for r. For example, we could take $r(\theta_n, \theta) = w(\theta)$ where w is the Student importance sampler discussed in Section 4. This is an example of an *independence chain*. Alternatively r could be the transition function of a random walk based on a multivariate Student density centered at 0 and with variance matrix equal to the inverse Hessian of $-\log f$. This is an example of a *random walk chain*.

A special case of the generalized Metropolis algorithm is the *Gibbs sampling* algorithm which was introduced in Geman and Geman (1984) and brought to the attention of the statistical community by Gelfand and Smith (1990). For Gibbs sampling $\alpha(\theta_n, \theta) \equiv 1$. We describe a particular version of the Gibbs algorithm. Let $p_i(\cdot \mid \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_k)$ denote the conditional posterior density of the i-th coordinate of θ given the remaining coordinates. The algorithm then proceeds as follows. We specify an initial state θ_0. Then given the n-th state $\theta_n = (\theta_{n,1}, \ldots, \theta_{n,k})'$ we obtain the $(n+1)$-st state in k steps :

generate $\theta_{n+1,1} \sim p_1(\cdot \mid \theta_{n,2}, \ldots, \theta_{n,k})$

generate $\theta_{n+1,2} \sim p_2(\cdot \mid \theta_{n+1,1}, \theta_{n,3}, \ldots, \theta_{n,k})$

$\vdots$

generate $\theta_{n+1,k} \sim p_k(\cdot \mid \theta_{n+1,1}, \ldots, \theta_{n+1,k-1})$.

There are numerous variations, such as replacing $l < k$ steps with generation from a single $l-$dimensional subvector; see Amit and Grenander (1991), which can be useful when the components of the subvector are highly correlated.

The Gibbs sampling algorithm has proven to be extremely useful in a wide variety of statistical contexts but particularly with hierarchical models in Bayesian analysis where it arises very naturally. Integrals of enormous dimension have been successfully handled. For a number of applications see Gelfand and Smith (1990), Gelfand, Hills, Racine-Poon and Smith (1990), Gelfand, Smith and Lee (1992) and Escobar (1994). Gibbs sampling requires algorithms for generating from each of a set of conditional distributions and obtaining such algorithms can be a very difficult problem. Sometimes, however, latent variables can be introduced, increasing the dimension of the integral, but making the implementation of the algorithm much simpler. This technique is based on ideas in Tanner and Wong (1987). Further Gilks and Wild (1992) have developed an efficient adaptive rejection generating algorithm for log-concave densities. A number of the standard distributions, and truncated versions of these, are log-concave. Algorithms which are a combination of Metropolis and Gibbs steps can also be used to deal with the generating problem; see Mueller (1991). As pointed out in Gelfand And Smith (1990) Gibbs sampling leads to efficient estimates of marginal densities via the technique of *Rao-Blackwellization*. For example, if we want to estimate the marginal density p_1 of the first coordinate then for realizations $\theta_1, \ldots, \theta_n$ of the chain we use $\hat{p}_1(x) = \frac{1}{N}\sum_{i=1}^{N} p_1(x|\theta_{i,2}, \ldots, \theta_{i,k})$. In general the Metropolis algorithm does not seem to provide such straight-forward density estimates.

Various mathematical convergence results for Markov chain methods have been obtained; see Tierney (1991), Schervish and Carlin (1992), Polson (1993) and Baxter and Rosenthal (1994). Convergence can be very slow, however. For a simple example of this, involving Gibbs sampling and a function f with two modes, see Evans, Gilula and Guttman (1993). Typically when f is unimodal with roughly ellipsoidal contours, experience suggests that a reasonable Markov chain algorithm will do well. Once again the issue of parameterization is extremely important; see Wakefield (1992) for more discussion of this issue.

The practical assessment of the convergence of the Markov chain to stationarity is a difficult problem. Typically this is done by monitoring estimates of characteristics of the stationary distribution p that can be represented as expectations or monitoring estimates of marginal densities. Convergence is claimed when these estimates stabilize for a number of iterations. It is not uncommon, however, for a chain to have appeared to have converged when in fact it is far from its stationary distribution; see Gelman and Rubin (1992). If p is multimodal then the chain can spend long periods in regions near a single mode. Further the outcomes in the Markov chain are correlated. If these serial correlations are high then we can see stable results even when we are far from stationarity. To detect the problem it seems sensible to start chains at a variety of starting values and check to see if the characteristics being monitored agree. This is part of the recommendation in Gelman and Rubin (1992) and it is analogous to running an optimization routine with a number of starting values to check for local optima. For an alternative view on this see Geyer (1992). The convergence diagnostics of Zellner and Min (1994), where we compare estimates with known quantities such as a ratio $p(\theta_1)/p(\theta_2)$, also seem very sensible. The methods described in Rosenthal (1994), for achieving a bound on the total variation distance between the stationary distribution and the distribution of the $N-th$ state, also work in particular examples.

Another issue is the choice of a method for assessing the accuracy of estimates. When the process $m(\theta_1), m(\theta_2), \ldots$ is stationary then results in Brockwell and Davis (1991) establish, with additional conditions, that $\frac{1}{N}\sum_{i=1}^{N} m(\theta_i) \sim AN\left(R(m), \frac{1}{N}\sigma^2\right)$ where $\sigma^2 = \sum_{i=-\infty}^{\infty} \gamma_m(i)$ and $\gamma_m(i)$ is the autocovariance at lag i. A relatively simple approach to estimating σ^2 is to use the method of *batching*. For this we divide the series $m(\theta_1), \ldots, m(\theta_N)$ into nonoverlapping batches of size l, calculate the mean in each batch and then estimate σ^2/l using the sample variance obtained from the batch means. The batch size l must be chosen large enough so that the batch means have low serial correlations but not so large as to leave few degrees of freedom for the estimation of

σ^2. Alternatively we can estimate $R(m)$ by taking the average of every $l-th$ value of the chain for N iterations. The asymptotic variance of this estimate is $\gamma_m(0)/N$, provided l is large enough to eliminate correlations, and $\gamma_m(0)$ is estimated by the sample variance. If we must take l large then this method can be very expensive. Also, as shown in MacEachern and Berliner (1994), this estimate of $R(m)$ always has variance at least as large as the estimate obtained from using the whole chain. For more on these and other methods for estimating σ^2 see Bratley, Fox and Schrage (1983), Ripley (1987), Geweke (1991), Geyer (1992) and Mykland, Tierney and Yu (1992).

We implemented several Markov chain algorithms on the examples of Section 2. It is often recommended that an initial part of the chain be discarded before begining estimation so that the chain has approximately achieved stationarity. This is sometimes called *burn-in*. In all of the examples this made very little difference and so it was ignored.

Example 1. Table 6 gives the results of Gibbs sampling. The generating problem was solved, as in Verdinelli and Wasserman (1991), by adding the independent latent variables $u_{ij} \sim N(0,1), v_{ij} \sim$ Gamma$(\lambda/2)$ for $i = 1,\ldots,9$ and $j = 1,\ldots,5$ such that $z_{ij} = \sqrt{\frac{\lambda-2}{2}}\frac{u_{ij}}{\sqrt{v_{ij}}} \sim$ Student$^*(\lambda)$. The posterior of $(\beta_1,\ldots,\beta_9,\sigma,v_{11},\ldots,v_{95})$ is proportional to

$$\sigma^{-nk-1}\prod_{i=1}^{9}\prod_{j=1}^{5}\exp\left\{-\frac{1}{2}\left(\frac{2}{\lambda-2}\right)v_{ij}\left(\frac{y_{ij}-\beta_i}{\sigma}\right)^2\right\}v_{ij}^{\frac{\lambda+1}{2}-1}e^{-v_{ij}}.$$

Then denoting, for example, the conditional distribution of β_1 given the remaining variables by $\beta_1|\cdot$, we have that

$$\beta_i|\cdot \sim N\left(\frac{\sum_{j=1}^{5} v_{ij}y_{ij}}{\sum_{j=1}^{5} v_{ij}}, \frac{\lambda-2}{2}\left(\sum_{j=1}^{5} v_{ij}\right)^{-1}\sigma^2\right),$$

$$\frac{1}{\sigma^2}|\cdot \sim \left(\frac{1}{\lambda-2}\sum_{i=1}^{9}\sum_{j=1}^{5} v_{ij}(y_{ij}-\beta_i)^2\right)^{-1}\text{Gamma}(45/2),$$

$$v_{ij}|\cdot \sim \left[\frac{1}{\lambda-2}\left(\frac{y_{ij}-\beta_i}{\sigma}\right)^2+1\right]^{-1}\text{Gamma}\left(\frac{\lambda+1}{2}\right).$$

This Gibbs sampler ran for approximately 50 seconds so that it used an equivalent amount of time

i	$R(\theta_i)$	$\hat{R}(\theta_i)$	$R(\theta_i^2)$	$\hat{R}(\theta_i^2)$
1	2.043	2.042 (.001)	4.263	4.260 (.002)
2	0.095	0.094 (.015)	0.081	0.081 (.008)
4	0.018	0.015 (.088)	0.069	0.069 (.008)
10	-0.073	-0.073 (.019)	0.033	0.033 (.010)

Table 6: Gibbs sampling estimates and the estimated absolute coefficients of variation in Example 1 where R denotes the exact posterior expectation and $\hat{R}$ denotes the Gibbs sampling estimate.

as the other Monte Carlo algorithms. This required $N = 50,000$ steps in the chain. The standard errors were estimated using batch size $l = 100$. We also tried $l = 1000$ and there was very little change in the estimates.

We also ran the Metropolis independence chain, using the importance sampler of Section 4. To provide a fair comparison with other methods we ran this for $N = 10^5$ steps. The standard errors were estimated using batch size $l = 100$ and the results were similar for $l = 1000$. This algorithm was substantially less accurate than Gibbs sampling. For example, the estimates and estimated absolute coefficients of variation, $\hat{R}(\theta_4) = .015(.206)$ and $\hat{R}(\theta_{10}) = -.071(.027)$ were obtained. There was evidence of substantial serial correlation as the estimates of the standard errors for $l = 20$ were very different. This was not the case with the Gibbs sampling algorithm. The acceptance rate for this algorithm; i.e. the proportion of generated values that resulted in a move for the chain, was .28. We also ran the random walk Metropolis chain and rescaled the variance by $2.4/\sqrt{10}$, following the advice given in Gelman, Roberts and Gilks (1994). Again we ran this for $N = 10^5$ steps and estimated the standard errors using batch size $l = 100$. The random walk chain gave less accuracy than the independence chain. The acceptance rate for this algorithm was .33. Varying the degrees of freedom for the Student produced only minor differences.

Example 2. For this problem we could not find an easily implemented Gibbs algorithm. We ran the Metropolis independence chain based on the starting importance sampler used in Section 5 for

$N = 10^5$ steps and estimated standard errors using a batch size of $l = 100$. This produced the estimates and estimated coefficients of variation, $\hat{R}(\theta) = .419(.002)$ and $\hat{R}(\theta^2) = .179(.003)$. Note that in both cases the true values of .422 and .181 are further than 3 standard errors from the estimate. This problem disappeared when we took $l = 1000$ as expected.

Example 3. We could not find a feasible Markov chain algorithm for this problem.

8. Conclusions

First we consider the relative performance of the various integration algorithms on the examples; then we will draw some general conclusions from our study. In Example 1, all the methods worked well except for subregion adaptive integration. Gibbs sampling and adaptive importance sampling were roughly equivalent with respect to efficiency, but Gibbs sampling is arguably easier to code here. We recommend Gibbs sampling in this context but note that if there is interest in estimating the norming constant, then adaptive importance sampling is the better choice because Gibbs sampling does not easily provide such an estimate. In Example 2, all the methods worked reasonably well but the subregion adaptive algorithm was again time consuming and we could not find an easily implemented Gibbs sampling algorithm. Adaptive importance sampling and the Metropolis independence chain were most efficient and roughly equivalent. The Metropolis algorithm suffered from high correlations, however. There is no saving in computational time or in programming time for this Markov chain algorithm over adaptive importance sampling here as both require maximizing f. In Example 3, subregion adaptive is by far the best and is our recommended algorithm although adaptive importance sampling also performed quite well.

The Laplace approach works well when it is applicable. Example 3 indicates that it is not always easy to apply the Laplace approximation in a given problem. Further it is clear that there is a close association between the success of the Laplace methods and the adequacy of the normal approximation to the integrand. Reparameterizatons then are recommended to ensure, for

example, that variables range over appropriate domains. At this point there is no methodology specifying a reparameterization for a particular problem. However, general recommendations, such as taking logs of nonnegative variables, should be helpful. In some problems, such as Example 2, more complicated reparameterizations are necessary. Computations are virtually instantaneous, but it can be extremely time consuming to code derivatives. Although there are optimization routines that numerically compute the derivatives, our experimentation with these sometimes lead to unsatisfactory results. In general, asymptotic methods require derivatives; this is a limitation of the methodology when the dimension is high and we do not have closed form expressions. The most serious drawback of this method is the lack of error estimates. Further, to avoid the problems associated with multimodality - and these occurred with both Examples 1 and 2 - it is necessary to try a number of different starting values when searching for the global maximum.

The success of the importance sampling algorithm that we have used is also highly dependent on having an integrand with a single dominant peak and approximately ellipsoidal contours. Of course asymptotics suggest that this is not a rare occurrence in statistical contexts, but complicated reparameterizations may often be necessary. It seems likely that Student importance sampling will be successful in any context where the asymptotic methods work. Further it is likely that the importance sampling algorithm will be more robust to deviations from ideal conditions than asymptotic methods. With this in mind, it is our recommendation that any time asymptotic methods are contemplated, then, at the very least, the Student importance sampling algorithm should be implemented as a check. There will be contexts where the asymptotic methods have a distinct practical advantage in terms of computation times, e.g. when there are many integrals to approximate. But even in such an instance a few integrals should be checked. For us, the possibility of obtaining useful assessments of the error in the approximation is an additional deciding factor in favour of importance sampling over asymptotics. Of course importance sampling suffers from

the same limitations with respect to derivatives as the asymptotic approaches, for it is always recommended that the importance sampler be centered at the mode and be scaled by the inverse Hessian at the mode. If more than one mode is found, then the methods of Oh and Berger (1993) are a possibility. Further, and as our examples show, in any context where Student importance sampling is going to be used, adaptive importance sampling is recommended because it leads to improvements in accuracy. The additional computational cost required to adapt to means, variances, and covariances is practically negligible.

For relatively low-dimensional problems, say $k \leq 6$, multiple quadrature is a competitive technique and is recommended for such problems. Again an appropriate parameterization is necessary. For example, the iterative product-rule approach described in Naylor and Smith (1982) depends on the validity of the normal approximation. When such an approximation is reasonable and the dimension is low, this approach is preferable to importance sampling because the convergence is much faster. The subregion adaptive algorithm is more robust to deviations in shape and yet is still efficient for most low dimensional problems. It is our recommended algorithm for low dimensions. As we have seen, particularily in Example 1, the curse of dimensionality remains a very big stumbling block with this algorithm even when we seem to have an appropriate parameterization. Of some concern is the inability of the user to predict how long the algorithm will run and the fact that it sometimes indicates success when it has actually failed. Error estimates are extremely conservative and for longer running examples, memory requirements can be substantial. We do not recommend this algorithm for problems of even modest dimension.

In many Bayesian problems Gibbs sampling arises very naturally. Sometimes these problems can involve thousands of dimensions and in such contexts it hard to imagine a competitor, simply because of the difficulties of implementation. A distinct advantage for Gibbs sampling over adaptive importance sampling is that there is no need to optimize first. In contexts where we have efficient

algorithms for generating from the conditionals, we recommend Gibbs sampling but note that this special structure is not enough to ensure rapid convergence. While there are many examples where convergence is extraordinarily rapid, there are others where it is so slow as to be impractical. Characterizing such contexts is an area of active and important research. Implementing Gibbs sampling is generally straightforward but sometimes requires the derivation of complicated conditional distributions. Although the generality of Metropolis algorithms can provide decided advantages over Gibbs sampling our experience with some of the more common implementations, as in Example 1, shows that this is not always the case. At this point we do not see Metropolis algorithms as a general solution for high-dimensional integration problems, but research may lead to the design of very efficient implementations for particular problems. Further we note that, as in Example 3, it is not always straightforward to find a reasonable Markov chain algorithm. Considerable disadvantages of Markov chain algorithms include the need to diagnose convergence to stationarity and the existence of serial correlations, as these substantially complicate error assessment. Some of our examples indicated that large batch sizes are needed to avoid this effect when analyzing the output using batching. This lack of independence is also true of adaptive importance sampling, but in adaptive importance sampling a consistent estimator of the error is easily obtained. We recommend parameterizing the problem to minimize the correlation effect, if possible.

There is well-tested software available for doing integrations. The NAG and IMSL packages contain routines, but also see Goel (1988) for a list of software primarly aimed at Bayesian problems. In particular we mention SBAYES which implements Bayesian asymptotic methods, BAYES FOUR which implements adaptive Hermite integration and adaptive importance sampling, BUGS which implements Gibbs sampling, and ADBAYS for subregion adaptive integration. In general, it is our view that a package aimed at solving practical integration problems should provide the user with the option of using any of the methods we have discussed.

Our study reveals advantages and disadvantages for each of the integration methods discussed in this paper. For low dimensions, a multiple quadrature approach is likely to be best. For modest dimensions, adaptive importance sampling or a Markov chain algorithm are good choices in well-behaved problems. For very high dimensions, an algorithm that can exploit special features of the integrand seems essential; e.g. Gibbs sampling when the full conditionals can be easily generated from. Of course other considerations beyond efficiency, such as convenience, may influence the choice of an approach. As for the crucial aspect of error assessment perhaps the best approach is to approximate the integral by several very different methods. This approach, however, leaves us with the question of what to do when results differ substantially.

Acknowledgements

The authors thank the editor, four referees and two additional readers for many constructive comments that lead to substantial improvements. Both authors were partially supported by grants from the Natural Sciences and Engineering Research Council of Canada.

References

Achcar, J.A. and Smith, A.F.M. (1990). Aspects of reparameterization in approximate Bayesian inference. Bayesian and Likelihood Methods in Statistics and Econometrics, eds. S. Geisser, J.S. Hodges, S.J. Press and A. Zellner, 439-452.

Amit, Y. and Grenander, U. (1991). Comparing sweep strategies for stochastic relaxation. J. Mult. Anal., 37, 197-222.

Barndorff-Nielsen, O.E. and Cox, D.R. (1989). Asymptotic Techniques for Use in Statistics. Chapman and Hall.

Barndorff-Nielsen, O.E. and Cox, D.R. (1994). Inference and Asymptotics. Chapman and Hall.

Baxter, J. and Rosenthal, J. (1994). Rates of convergence for everywhere positive Markov chains. Tech. Rep., No. 9406, Dept. of Stat., U. of Toronto.

Bernsten, J. (1989). Practical error estimation in adaptive multidimensional quadrature routines. J. of Comput. and Appl. Math., 25, 327-340.

Bernsten, J., Espelid, T.O. and Genz, A. (1991). An adaptive algorithm for the approximate calculation of multiple integrals. ACM Trans. on Math. Software, Vol. 17, No. 4, 437-451.

Besag, J. and Green, P.J. (1993). Spatial statistics and Bayesian computation. JRSSB, 55, 1, 25-37.

Besag, J., Green, P., Higdon, D. and Mengersen, K. (1995). Bayesian computation and stochastic systems. To appear in Stat. Sci.

Bratley P., Fox B.L. and Schrage, L.E. (1983). A Guide to Simulation. Springer-Verlag.

Brockwell, P. and Davis, R. (1991). Time Series: Theory and Methods. Second Edition. Springer-Verlag.

Chen, C-F. (1985). On the asymptotic normality of limiting density functions with Bayesian implications. JRSSB, 47, No. 3, 540-546.

Cools, R. and Rabinowitz, P. (1993). Monomial cubature rules since 'Stroud': a compilation. J. Comput. Appl. Math., 48, 309-326.

Cranley, R. and Patterson, T.N.L. (1970). A regression method for the Monte Carlo evaluation of multidimensional integrals. Numer. Math., 16, 58-72.

Davis, P.J. and Rabinowitz, P. (1984). Methods of Numerical Integration, Second Edition. Academic Press.

Dellaportas, P. and Wright, D. (1991). Positive embedded integration in Bayesian analysis. Stat. and Compt., 1, 1-12.

Dellaportas, P. and Wright, D. (1992). A numerical integration strategy in Bayesian analysis. Bayesian Statistics 4, 601-606, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith. Oxford University Press.

Devroye, L. (1986). Non-Uniform Random Variate Generation. Springer-Verlag.

Diaconis, P. (1988). Bayesian numerical analysis. In Statistical Decision Theory and Related Topics IV, Vol. 1., eds. S.S. Gupta and J. Berger, 163-175, J. Wiley.

DiCiccio, T.J., Field, C.A. and Fraser, D.A.S. (1990). Approximations of marginal tail probabilities and inference for scalar parameters. Biometrika, 77, 77-95.

van Dijk, H.K., Kloek, T. and Louter, A.S. (1986). An algorithm for the computation of posterior moments and densities using simple importance sampling. The Statistician, 35, 83-90.

van Dooren, P. and de Ridder, L. (1976). Algorithm 6. An adaptive algorithm for numerical integration over an n-dimensional cube. J. Comput. Appl. Math., 2, 207-217.

Durrett, R. (1990). Probability: Theory and Examples. Wadsworth and Brooks/Cole.

Erkanli, A. (1994). Laplace approximations for posterior expectations when the mode occurs at the boundary of the parameter space. JASA, Vol. 89, NO. 425, 250-258.

Escobar, M. (1994). Estimating normal means with a Dirichlet process prior. JASA, Vol. 89, No. 425, 268-277.

Evans, M. (1988). Monte Carlo computation of marginal posterior quantiles. Tech. Rep. No. 11, Dept. of Statistics, U. of Toronto.

Evans, M. (1991a). Adaptive importance sampling and chaining. Statistical Multiple Integration, eds. N. Flournoy and R.K. Tsutakawa, Contemporary Mathematics, 115, Amer. Math. Soc., 137-143.

Evans, M. (1991b). Chaining via annealing. Ann. Statist., Vol. 19, No. 1, 382-393.

Evans, M., Gilula, Z. and Guttman, I. (1989). Latent class analysis of two-way contingency tables by Bayesian methods. Biometrika, 76, 3, 557-563.

Evans, M., Gilula, Z. and Guttman, I. (1993). Computational issues in the Bayesian analysis of categorical data: log-linear and Goodman's RC model. Stat. Sin., Vol. 3, No. 2, 391-406.

Evans, M., Guttman, I. and Olkin, I. (1993). Numerical aspects in estimating the parameters of a mixture of normal distributions. J. Comp. and Stat. Graph., Vol. 1, 4, 351-365.

Evans, M, and Swartz, T. (1988a). Monte Carlo computation of some multivariate normal probabilities. J. Statist. Comput. Simul., Vol. 30, 117-128.

Evans, M. and Swartz, T. (1988b). Sampling from Gauss rules. SIAM J. Sci. Stat. Comput., Vol. 9, No. 5, 950-961.

Evans, M. and Swartz, T. (1992). Some integration strategies for problems in statistical inference. Computing Science and Statistics: Proceedings of the 24th Symposium on the Interface, Ed. J. Newton, Interface Foundation of America, 310-317.

Evans, M. and Swartz, T. (1994). Distribution theory and inference for polynomial-normal densities. Commun. Statist.- Theory Meth., 23(4), 1123-1148.

Fang, K-T., Wang, Y. and Bentler, P.M. (1994). Some applications of number-theoretic methods in statistics. Stat. Sci., Vol. 9, No. 3, 416-428.

Fishman, G.S. and Huang, B.D. (1983) Antithetic variates revisited. Commun. ACM, Vol. 26, No. 11, 964-971.

Flournoy, N. and Tsutakawa, R.K. (1991). Statistical Multiple Integration. Contemporary Mathematics 115, Amer. Math. Soc.

Fraser, D.A.S., Lee, H.S. and Reid, N. (1990). Nonnormal regression; an example of significance levels in high dimensions. Biometrika, 77, 2, 333-341.

Friedman, J.H. and Wright, M.H. (1981). A nested partitioning procedure for numerical multiple integration. ACM Trans, Math. Software, Vol. 7, No. 1, 76-92.

Gelfand, A. E., Hills, S. E., Racine-Poon, A. and Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. JASA, 85, 972-985.

Gelfand, A.E. and Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities. JASA, 85, 398-409.

Gelfand, A.E., Smith, A.F.M. and Lee, T-M. (1992). Bayesian analysis of constrained parameter and truncated data problems. JASA, 87, 523-532.

Gelman, A. and Rubin, D.B. (1992). Inference from iterative simulation. Stat. Sci., Vol. 7, No. 4, 473-483.

Gelman, A., Roberts, G. and Gilks, W. (1994). Efficient Metropolis jumping rates. Manuscript.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. on Pattern Analysis and Machine Intelligence, 6, 721-741.

Genz, A. (1991). Subregion adaptive algorithms for multiple integrals. Statistical Multiple Integration, eds. N. Flournoy and R.K. Tsutakawa. Contemporary Mathematics, 115, Amer. Math. Soc, 23-31.

Genz, A. (1992a). Numerical computation of multivariate normal probabilities. J. Comput. Graph. Stat. 1, 141-150.

Genz, A. (1992b). Statistics applications of subregion adaptive multiple numerical integration. In Numerical Integration: Recent Developments, Software and Applications, eds. T.O. Espelid and A. Genz. Kluwer Academic Publishers, 267-280.

Genz, A. and Kass, R. (1991). An application of subregion adaptive numerical integration to a Bayesian inference problem. Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface, Ed. E. Keramidas, Interface Foundation of America, 441-444.

Genz, A. and Kass, R. (1994). Subregion adaptive integration of functions having a dominant peak. Manuscript.

Genz, A. and Malik, A. (1980) Algorithm 019. Remarks on algorithm 006: An adaptive algorithm for numerical integration over an n-dimensional rectangular region. J. Comp. Appl. Math., 6, 295-302.

Genz, A. and Malik, A. (1983). An imbedded family of fully symmetric numerical integration rules. SIAM J. of Numer. Anal., 20, 580-587.

Geyer, C.J. (1992). Practical Markov Chain Monte Carlo. Stat. Sci., Vol. 7, No. 4, 473-483.

Geyer, C.J. and Thompson, E.A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion). JRSSB, 54, 657-700.

Geweke, J. (1988). Antithetic acceleration of Monte Carlo integration in Bayesian inference. J. of Econometrics, 38, 73-89.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. Econometrica, 57, 1317-1339.

Geweke, J. (1991). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. Bayesian Statistics 4, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith. Oxford Press

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. Appl. Statist., 41, 337-348.

Givens, G.H. and Raftery, A.E. (1993). Local adaptive importance sampling for multivariate densities with strong nonlinear relationships. Tech. Rep. No. 260, Dept. of Stat., U. of Washington.

Glynn, P.W. and Iglehart, D.L. (1989). Importance sampling for stochastic simulations. Mgmnt. Sc., Vol. 35, No. 11, 1367-1392.

Goel, P.K. (1988). Software for Bayesian analysis: current status and additional needs. Bayesian Statistics 3, 173-188, eds. J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith. Oxford Press.

Haber, S. (1970). Numerical evaluation of multiple integrals. SIAM Review, Vol. 12, No. 4, 481-526.

Hall, P. and Heyde, C.C. (1980). Martingale Limit Theory And Its Application. Academic Press.

Hammersley, J.M. (1960). Monte Carlo methods for solving multivariable problems. Ann. New York Acad. Sci., 86, 844-874.

Hammersley, J.M. and Handscomb, D.C. (1964). Monte Carlo Methods. Methuen and Co.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. Biometrika, 57, 1, 97-109.

Hesterberg, T. (1990). Importance sampling for Bayesian estimation. Computing and Graphics in Statistics, eds. A. Buja and P.A. Tukey.

Hills, S.E. and Smith, A.F.M. (1992). Parameterization issues in Bayesian inference. Bayesian Statistics 4, 227-246, eds. J.M Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, Oxford University Press.

Johnson, V.E. (1992) A technique for estimating marginal posterior densities in hierarchical models using mixtures of conditional densities. JASA, Vol. 87, No. 419, 852-860.

Kass, R.E., Tierney, L. and Kadane, J.B. (1988). Asymptotics in Bayesian computation. Bayesian Statistics 3, eds. J. Bernardo, M. DeGroot, D.V. Lindley and A.F.M. Smith. Oxford University Press.

Kass, R.E., Tierney, L. and Kadane, J.B. (1989). Approximate methods for assessing influence and sensitivity in Bayesian analysis. Biometrika, 76, 663-674.

Kass, R.E., Tierney, L. and Kadane, J.B. (1990). The validity of posterior expansions based on Laplace's method. Bayesian and Likelihood Methods in Statistics and Econometrics, eds.

S. Geisser, J.S. Hodges, S.J. Press and A. Zellner, Elsevier Science Publishers B.V. (North Holland), 473-488.

Kass, R.E., Tierney, L. and Kadane, J.B. (1991). Laplace's method in Bayesian analysis. Statistical Multiple Integration, eds. N. Flournoy and R.K. Tsutukawa. Contemporary Mathematics, 115, Amer. Math. Soc., 89-99.

Kloek, T. and van Dijk, H.K. (1978). Bayesian estimates of equation system parameters: an application of integration by Monte Carlo. Econometrica, Vol. 46, No. 1, 1-19.

Lepage, G.P. (1978) A new algorithm for adaptive multidimensional integration. J. Comput. Phys., 27, 192-203.

Lindley, D.V. (1961). The use of prior probability distributions in statistical inference and decisions. Proc. 4th Berkeley Symp. 1, 453-468.

Lindley, D.V. (1980). Approximate Bayesian methods. Bayesian Statistics, eds. J.M. Bernardo, M.H. De Groot, D.V. Lindley and A.F.M. Smith. Valencia, Spain : University Press.

Linz, P. (1988). A critique of numerical analysis. Bull. Amer. Math. Soc., Vol. 19, No. 2, 407-416.

Lyness, J.N. (1983). When not to use an automatic quadrature routine. SIAM Rev., Vol. 25, No. 1, 63-87.

MacEachern, S.N. and Berliner, L.M. (1994). Subsampling the Gibbs sampler. Amer. Statist., Vol. 48, No. 3, 188-190.

McKay, M.D., Beckman, R.J. and Conover, W.J. (1979). A comparison of three methods for

selecting values of input variables in the analysis of output from a computer code. Technometrics, Vol. 21, No. 2, 239-245.

Metropolis, N., Rosenbluth, A.W., Teller, A.H., and Teller, E. (1953). Equations of state calculations by fast computing machines. J. Chemical Physics, 21, 1087-1091.

Morris, C.N. (1988). Approximating posterior distributions and posterior moments. Bayesian Statistics 3, 327-344, eds. J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith. Oxford Press.

Mueller, P. (1991). A generic approach to posterior integration and Gibbs sampling. Tech. Rep. #91-09, Dept. of Stat., Purdue U.

Mykland P., Tierney, L. and Yu, B. (1992). Regeneration in Markov chain samplers. Tech. Rep. No. 585, School of Statistics, U. of Minnesota.

Naylor, J.C. and Smith, A.F.M. (1982). Applications of a method for the efficient computation of posterior distributions. Appl. Stat., 31, 214-225.

Naylor, J.C. and Smith, A.F.M. (1992). Econometric illustrations of novel integration strategies for Bayesian inference. J. of Econometrics, 38, 103-125.

Neal, R. (1993). Probabilistic inference using Markov chain Monte Carlo methods. Tech. Rep. CRG-TR-93-1, Dept. of Comp. Sci., U. of Toronto.

Niederreiter, H. (1992). Random Number Generation and Quasi-Monte Carlo Methods. CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM.

Oh, M-S. (1991). Monte Carlo integration via importance sampling: dimensionality effect and an adaptive algorithm. Statistical Multiple Integration, eds. N. Flournoy and R.K. Tsutakawa,

Contemporary Mathematics, 115, 165-187, Amer. Math. Soc..

Oh, M-S. and Berger, J. (1992). Adaptive importance sampling in Monte Carlo integration. J. Stat. Comp. and Simul., 41, 143-168.

Oh, M-S. and Berger, J.O. (1993). Integration of multimodal functions by Monte Carlo importance sampling. JASA, Vol. 88, No. 422, 450-456.

O'Hagan, A. (1987). Monte Carlo is fundamentally unsound. The Statistician, 36, 247-249.

O'Hagan, A. (1991). Bayes-Hermite integration. J. Statist. Plan. Inf., 29, 245-260.

Owen, A.B. (1992). A central limit theorem for latin hypercube sampling. J.R. Statist. Soc. B, 54, No. 2, 541-551.

Peskun, P.H. (1973). Optimum Monte Carlo sampling using Markov chains. Biometrika, 60, 3, 607-612.

Piccioni, M. and Ramponi, A. (1993). Minimum chisquare unilateral approximation of Gibbs random fields. Manuscript.

Polson, N. (1993). Convergence of Markov chain Monte Carlo algorithms. Working paper 93-148, Grad. Sch. Bus., U. of Chicago.

Powell, M.J.D. and Swann, J. (1966). Weighted uniform sampling - a Monte Carlo technique for reducing variance. J. Inst. Appl. Math., 2, 228-236.

Reid, N. (1988). Saddlepoint methods and statistical inference (with discussion). Stat. Sci., Vol. 3, No. 2, 213-238.

Ripley, B.D. (1987). Stochastic Simulation. John Wiley and Sons.

Rosenthal, J. (1994). Theoretical rates of convergence for Markov chain Monte Carlo. To appear in Computer Science and Statistics: Proc. of the 26-th Symp. on the Interface

Rubin, D.B. (1988). Using the SIR algorithm to simulate posterior distributions. Bayesian Statistics 3, 395-402, J.M. Bernardo, D.V. Lindley and A.F.M. Smith (eds.). Oxford University Press.

Schervish, M. (1984). Multivariate normal probabilities with error bound. Appl. Stat., 33, 81-87.

Schervish, M. and Carlin, B.P. (1992). On the convergence of successive substitution sampling. J. Comp. Graph. Stat., 1, 111-127.

Shaw, J.E.H. (1988). A quasirandom approach to integration in Bayesian statistics. Ann. Statist., Vol 16, No. 2, 895-914.

Smith, A.F.M. (1991). Bayesian computational methods. Phil. Trans. R. Soc. Lond., A 337, 369-386.

Smith, A.F.M. and Gelfand, A.E. (1992). Bayesian statistics without tears: a sampling-resampling perspective. Amer. Stat., Vol. 46, No. 2, 84-88.

Smith, A. F. M. and Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. J. R. Statist. Soc. B, 55, 3-23.

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

Smith, A.F.M., Skene, A.M., Shaw, J.E.H., Naylor and Dransfield, M. (1985). The implementation of the Bayesian paradigm. Commun. Statist.-Theor. Meth., 14(5), 1079-1102.

Spanier, J. and Maize, E.H. (1994). Quasi-random methods for estimating integrals using relatively small samples. SIAM Review, Vol. 36, No. 1, 18-44.

Stein, M. (1987). Large sample properties of simulations using latin hypercube sampling. Technometrics, Vol. 29, No. 2, 143-151.

Stroud, A.H. (1971). Approximate Calculation of Multiple Integrals. Prentice-Hall Inc.

Tang, B. (1993). Orthogonal array-based latin hypercubes. JASA, Vol. 88, No. 424, 1392-1397.

Tanner, M. (1993). Tools for Statistical Inference. Observed Data and Data Augmentation Methods. Second Edition. Springer-Verlag.

Tanner, M.A. and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation. JASA, 82, 528-540.

Thisted, R.A. (1988). Elements of Statistical Computing. Chapman and Hall.

Tierney, L. (1991). Markov chains for exploring posterior distributions. Technical Report No. 560, School of Statistics, University of Minnesota and to appear Ann. Statist.

Tierney, L. and Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities. JASA, Vol. 81, No. 393, 82-86.

Tierney, L., Kass, R. and Kadane, J.B. (1989a). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. JASA, 84, 710-716.

Tierney, L., Kass, R. and Kadane, J.B. (1989b). Approximate marginal densities of nonlinear functions. Biometrika, 76, 3, 425-433.

Traub, J.F. and Wozniakowski, H. (1991). Information-based complexity: new questions for mathematicians. Math. Int., Vol. 3, No. 2, 34-43.

Verdinelli, I. and Wasserman, L. (1991). Bayesian analysis of outlier problems using the Gibbs sampler. Stat. Comp., 1, 105-117.

Wakefield, J. (1992). Comment on Hills and Smith (1992). Bayesian Statistics 4, 243-244, eds. J.M Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, Oxford University Press.

West, M. (1993) Approximating posterior distributions by mixtures. JRSSB, 55, No. 2, 409-422.

Wing, J.W. (1962). Institutionalism in mental hospitals. Br. J. Social Clin. Psych., 1, 38-51.

Wolpert, R.L. (1990). Monte Carlo integration in Bayesian statistical analysis. Statistical Multiple Integration, eds. N. Flournoy and R.K. Tsutakawa, Contemporary Mathematics, 115, Amer. Math. Soc., 101-116.

Wong, R. (1989). Asymptotic Approximations of Integrals. Academic Press.

Wozniakowski, H. (1991). Average case complexity of multivariate integration. Bull. AMS, Vol. 24, No. 1, 185-194.

Zellner, A. and Min, C. (1994). Gibbs sampler convergence criteria. To appear in JASA.