Bayesian Integration Using Multivariate Student Importance Sampling

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

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

Abstract

Multivariate Student importance sampling is a commonly used technique for Bayesian integration problems. For many problems, for example, when the integrand has ellipsoidal symmetry, it is a feasible and perhaps preferable, integration strategy. It is widely recognized, however, that the methodology suffers from limitations. We examine the methodology in conjunction with various variance reduction techniques; e.g. control variates, stratified sampling and systematic sampling, to determine what improvements can be achieved.

1 Introduction

It is only in the simplest Bayesian models where posterior quantities can be evaluated in closed form. For most Bayesian problems we are faced with the need to approximate integrals of the form

$$I(m) = \int_{R^k} m(\boldsymbol{x}) f(\boldsymbol{x})\, d\boldsymbol{x}$$

for some $m : R^k \to R$ and $f : R^k \to R$. Here $f \geq 0$ is the product of the likelihood and the prior and m determines the posterior quantity of interest. For example, $I(1)$ is the norming constant and $I(\boldsymbol{x})/I(1)$ is the vector of posterior means.

There exist a wide array of methods available for the approximation of $I(m)$ and the ratios $R(m) = I(m)/I(1)$. Evans and Swartz (1995a) provide a survey of the major integration techniques currently in use; namely asymptotic approximations, importance sampling, adaptive importance sampling, multiple quadrature and Markov chain methods. In problems where f has a dominant peak and approximate ellipsoidal symmetry, importance sampling, based on the multivariate Student distribution centered at the mode of f and scaled by the inverse Hessian matrix of $-\log f$ at the mode, proves to be a very competitive algorithm and in some ways is even preferable. For example, Markov chain methods require the determination of the attainment of stationarity. Further once stationarity has been achieved, error estimation is complicated by the correlation which exists between variates. These problems are avoided with importance sampling provided that we have a good importance sampler.

In this paper we discuss importance sampling in conjunction with variance reduction techniques. In particular, we are interested in the extent to which easily implemented variance reduction techniques can improve a possibly poor importance sampler. The general importance sampling algorithm proceeds as follows: generate $\boldsymbol{x} \sim w$ and estimate $I(m)$ by

$$\hat{I}(m) = \frac{m(\boldsymbol{x}) f(\boldsymbol{x})}{w(\boldsymbol{x})}.$$

This estimate is then averaged over a sample of size N from w. We estimate the ratio $R(m)$ by $\hat{R}(m) = \hat{I}(m)/\hat{I}(1)$.

Importance sampling methodology has limitations. For example, in a given problem, we need to choose an importance sampler that ensures estimates with finite variances. Moreover, an importance sampler should adequately mimic the integrand so that the variances are small. It is also necessary that the importance sampler permit efficient variate generation.

The importance sampling approach which we consider in this paper is that based on the multivariate Student importance sampler. The asymptotic normality of posterior distributions suggests that the multivariate Student will often be a reasonable choice as an importance sampler. Our goal here is to see whether meaningful variance reductions can be obtained when we tailor some common variance reduction techniques to work in tandem with Student importance sampling. In Section 2 we develop and investigate the use of additive and ratio control variates. Section 3 is devoted to stratified sampling and Section 4 to systematic sampling. Concluding remarks are given in Section 5.

For Student importance sampling we first use an optimization routine to obtain the value $\hat{\boldsymbol{u}}$ which maximizes f. This serves to appropriately locate the importance sampler. For the appropriate scaling we calculate $\hat{\Sigma} = (-\partial^2 \log f(\hat{\boldsymbol{u}})/\partial x_i \partial x_j)^{-1}$. The importance sampler is then given by

$$\boldsymbol{x} \sim \hat{\boldsymbol{u}} + \sqrt{\frac{\lambda - 2}{\lambda}}\, \hat{\Sigma}^{1/2}\, t_{k,\lambda} \qquad (1)$$

where $t_{k,\lambda}$ is the standard k-dimensional Student distribution with λ degrees of freedom. The choice of the parameter λ is unclear but the general idea is to choose it low so that hopefully, all estimates have finite variance. If the variances of both $\hat{I}(m)$ and $\hat{I}(1)$ are finite then the asymptotic variance of $\hat{R}(m)$, based on a sample of size N, is

$$\frac{1}{N}(\frac{\mu_1}{\mu_2})^2[(\frac{\sigma_1}{\mu_1})^2 + (\frac{\sigma_2}{\mu_2})^2 - 2(\frac{\sigma_1}{\mu_1})(\frac{\sigma_2}{\mu_2})\rho_{12}] \qquad (2)$$

where μ_1 is the mean of $\hat{I}(m)$, μ_2 is the mean of $\hat{I}(1)$, σ_1^2/N is the variance of $\hat{I}(m)$, σ_2^2/N is the variance of $\hat{I}(1)$ and ρ_{12}/N is the correlation between $\hat{I}(m)$ and $\hat{I}(1)$. The asymptotic variance (2) is easily estimated from the generated data.

We remark that when Student importance sampling is successful it is often possible to obtain variance reductions through the related strategy of adaptive Student importance sampling. The basic idea is to consider the class of Student importance samplers parametrized by the mean vector and covariance matrix. As sampling continues, the Student importance sampler is updated to agree more closely with the integrand of interest with respect to these characterisitics. Typically this results in improvements; see Evans and Swartz (1995a).

Throughout the paper, we illustrate the discussion with an example. We apply each of the proposed methods to the Bayesian analysis of the linear model $y = \beta_1 z_1 + \cdots + \beta_9 z_9 + \sigma e$ where the z_i are dummy variables, $e \sim Student(3)/\sqrt{3}$ and we use the flat prior on the unknown parameter $(\beta_1, \ldots, \beta_9, \log \sigma)$. We use simulated data, generating 5 observations from each of 9 populations, prescribed by setting $\sigma = 1$, $\beta_1 = 1$ and $\beta_i = 0$ for $i \neq 1$. See Example 1 of Evans and Swartz (1995a) for more details. This is a 10-dimensional integration problem. We set the degrees of freedom λ in the Student importance sampler equal to 5 and examine the approximation of the posterior expectations of various functions of the parameters.

2 Control Variates

For a control variate we use a function g that resembles f and is such that $\int_{R^k} m(\boldsymbol{x}) g(\boldsymbol{x})\, d\boldsymbol{x}$ can be evaluated in closed form. The idea is that some of the variation in f can be removed via g. As a natural choice, we let $g(\boldsymbol{x}) = I_{lap}(1)\phi(\boldsymbol{x})$ where $I_{lap}(1)$ is the Laplace approximation of $I(1)$ and ϕ is the $N_k(\hat{\mu}, \hat{\Sigma})$ density. This control variate was suggested in Evans and Swartz (1995a). Using the Student importance sampler given by (1), the estimator of $I(m)$ based on using g as an additive control variate is

$$\frac{1}{N}\sum_{i=1}^{N} m(\boldsymbol{x}_i)\, \frac{(f(\boldsymbol{x}_i) - g(\boldsymbol{x}_i))}{w(\boldsymbol{x}_i)} + \int_{R^k} m(\boldsymbol{x}) g(\boldsymbol{x})\, d\boldsymbol{x}. \qquad (3)$$

Similarly, the estimator of $I(m)$ based on using g as a ratio control variate, when $\int_{R^k} m(\boldsymbol{x}) g(\boldsymbol{x}) d\boldsymbol{x} \neq 0$, is

$$\left(\int_{R^k} m(\boldsymbol{x}) g(\boldsymbol{x})\, d\boldsymbol{x}\right) \frac{\sum_{i=1}^{N} m(\boldsymbol{x}_i) f(\boldsymbol{x}_i)/w(\boldsymbol{x}_i)}{\sum_{i=1}^{N} m(\boldsymbol{x}_i) g(\boldsymbol{x}_i)/w(\boldsymbol{x}_i)}. \qquad (4)$$

Assuming $Corr(mf/w, mg/w) \approx 1$, it is straightforward to show that the absolute value of the coefficient of variation of (3) is approximately equal to

$$\frac{1}{N}\left|CV(mf/w) - \frac{E(mg/w)}{E(mf/w)} CV(mg/w)\right| \qquad (5)$$

where CV denotes the coefficient of variation and all expectations are taken with respect to the importance sampler w. Also, for large sample sizes, the absolute value of the coefficient of variation of (4) is approximately equal to

$$\frac{1}{N}\left|CV(mf/w) - CV(mg/w)\right|. \qquad (6)$$

From (5) and (6) it is possible to deduce circumstances when the additive control variate is better than the ratio control variate and vice-versa.

For the additive approach, standard errors for $R(m)$ can be calculated by using the asymptotic variance formula (2) with f replaced by $f - g$. With the ratio control variate, we have a ratio of ratios, and the asymptotic variance is $h'Vh/N$ where the vector h is given by $h' = (\mu_2\mu_3)^{-1}(\mu_4,\ -\mu_1\mu_4/\mu_2,\ -\mu_1\mu_4/\mu_3,\ \mu_1)$, the matrix $V = (\sigma_{ij})$ and the indices $1, 2, 3$ and 4 refer to the estimators mf/w, mg/w, f/w and g/w respectively.

Introducing control variates to our linear model example resulted in minor savings. For example, the average efficiency of the control variate technique in estimating the vector of posterior means is 1.1 when compared to straight Student importance sampling. For the ratio variate technique, the average efficiency is 1.2. These small savings are the result of a poor approximation of the function f by the Laplace approximation. Indeed, when we modify the example to have 20 replicates instead of 5, the posterior becomes more normal and the efficiencies become greater. For example, the average efficiency in estimating the vector of posterior means is 4.7 using the additive control variate and 5.1 using the ratio control variate. In terms of computing overhead, we observe that the control variate methods require only slightly more computing time than Student importance sampling. Thus these control variates are primarily useful when the Laplace approximation is good, and when this is the case, we also expect the Student importance sampler to perform well.

3 Stratified Sampling

For stratified sampling, suppose that the importance sampler w can be decomposed as

$$w(\boldsymbol{x}) = p_1 w_1(\boldsymbol{x}) + \cdots + p_m w_m(\boldsymbol{x})$$

where w_i is a density with support S_i and $\{S_1, \ldots, S_m\}$ is a partition of R^k. Then $p_i = \int_{S_i} w(\boldsymbol{x})\, d\boldsymbol{x}$ and $p_1 + \cdots + p_m = 1$. Referring to the S_i as strata, we generate $\boldsymbol{x}_{j1}, \ldots \boldsymbol{x}_{jn_j}$ from w_j and calculate the stratified estimator

$$\sum_{j=1}^{m} \frac{1}{n_j} \sum_{i=1}^{n_j} m(\boldsymbol{x}_{ji}) f(\boldsymbol{x}_{ji}) / w_j(\boldsymbol{x}_{ji}). \tag{7}$$

The variance of the stratified estimator (7) is given by

$$\sum_{j=1}^{m} \frac{1}{n_j} Var_{w_j}[mf/w_j]$$

and can be estimated from the Monte Carlo samples.

Natural questions concern the choice of the strata and the choice of the n_j given the strata. It is well known that the optimal choice of the n_j, given the strata, is the Neyman allocation with

$$n_j = N \frac{SD_{w_j}(mf/w_j)}{\sum_{i=1}^{m} SD_{w_i}(mf/w_i)}.$$

This requires knowledge of the $SD_{w_i}(mf/w_i)$, however, and is therefore impractical. It is also well known that proportional allocation; i.e. $n_j = p_j N$, always leads to a variance reduction when compared to straight importance sampling based on w. A reasonable strategy then is to begin with proportional allocation, estimate the $SD_{w_i}(mf/w_i)$ based on a preliminary sample, and then use the Neyman allocation.

We now consider the selection of the strata for Student importance sampling. For this, suppose that we have transformed from $\boldsymbol{x}$ to $\boldsymbol{y} = \hat{\Sigma}^{-1/2}(\boldsymbol{x} - \hat{\boldsymbol{u}})$ in the original expression for $I(m)$. Then for the importance sampler we use the density w of $\boldsymbol{y} \sim \sqrt{\frac{\lambda-2}{\lambda}}\, t_{k,\lambda}$. Because of the spherical symmetry of w it is natural to decompose w by choosing the strata to be a set of annular rings that partition R^k. That is, w_i is the density proportional to w on $l_i \le \boldsymbol{y}'\boldsymbol{y} \le l_{i+1}$ where the (l_i, l_{i+1}) are specified constants, $i = 1, \ldots, m$ with $l_1 = 0$ and $l_{m+1} = \infty$.

To sample from the rotationally symmetric density w_i, we first generate $\boldsymbol{v} \sim Unif(S^{k-1})$. This can be accomplished by generating a sample $z_1, \ldots, z_k$ from the standard normal distribution and setting $\boldsymbol{v} = \boldsymbol{z}/||\boldsymbol{z}||$. We then generate $r^2 \sim k\frac{\lambda-2}{\lambda} F(k, \lambda)$ conditioned to (l_i, l_{i+1}) and set $\boldsymbol{y} = r\boldsymbol{v}$.

An optimal choice for the boundary points $l_2, \ldots, l_m$ is an open problem. We have chosen the boundary points such that each annular ring in the Student density has equal probability.

Using an obvious notation we estimate $R(m)$ by

$$\frac{\sum_{j=1}^{m} \hat{I}_j(m)}{\sum_{i=1}^{m} \hat{I}_j(1)}.$$

The asymptotic variance is then given by

$$\left(\frac{\sum_{j=1}^{m} \mu_{1j}}{\sum_{j=1}^{m} \mu_{2j}}\right)^2 \sum_{j=1}^{m} \frac{1}{n_j} \left[\frac{\sigma_{11}^{(j)}}{\mu_{1j}^2} + \frac{\sigma_{22}^{(j)}}{\mu_{2j}^2} - 2\frac{\sigma_{12}^{(j)}}{\mu_{1j}\mu_{2j}}\right]$$

where μ_{1j} is the mean of $\hat{I}_j(m)$, μ_{2j} is the mean of $\hat{I}_j(1)$, $\sigma_{11}^{(j)}/n_j$ is the variance of $\hat{I}_j(m)$, $\sigma_{22}^{(j)}/n_j$ is the variance of $\hat{I}_j(1)$ and $\sigma_{12}^{(j)}/n_j$ is the covariance between $\hat{I}_j(m)$ and $\hat{I}_j(1)$.

We applied stratified sampling to our linear model example using $m = 25$ strata. Using proportional allocation, we first generated a small sample of size 100 in each stratum. We then switched to Neyman allocation using the estimates obtained from the initial sample. We again observed variance reductions. For example, the average efficiency of stratified sampling in estimating the vector of posterior means is 2.8 when compared to Student importance sampling.

In this same example we also tried different boundary points based on the observation that the estimators of the component strata $\hat{I}_j(m)$ have greatly different standard errors. We therefore began with 2 strata having equal probability and generated a small sample of size 100 in each stratum. We then split the stratum having the largest standard error into 2 sub-strata of equal probability. This splitting process was continued until $m = 25$ strata were constructed. At this point we switched to Neyman allocation. However, this strategy did not result in any meaningful variance reductions.

We remark that it is extremely important that a good algorithm be used to generate from the conditional distributions of r^2. For example, using a crude inversion technique and factoring in the additional computational times that this requires, the efficiency of stratified sampling over Student importance sampling reduces from 2.8 to 1.3. Therefore to achieve the full efficiencies inherent in this stratified algorithm a much better generator is required. An alternative approach is to use a variation of adaptive rejection sampling due to Gilks and Wild (1992) to construct a more efficient generator. While the Gilks and Wild algorithm is applicable to log-concave densities, and in general $F(k, \lambda)$ densities are not log-concave, we note that the log of the density of the $F(k, \lambda)$ distribution has a single point of inflection at $\tilde{x} = (\lambda/k)\sqrt{k-2}/(\sqrt{k+\lambda} - \sqrt{k-2})$. Thus by choosing one of the boundary points $l_2, \ldots, l_m$ equal to $k(\lambda-2)\tilde{x}/\lambda$, we are faced with sampling from densities which are either log-concave or log-convex. The Gilks and Wild (1992) algorithm can be modified to handle log-convex densities.

4 Systematic Sampling

The use of systematic sampling for the integration problem leads to many different algorithms; see, for example, Hammersley and Morton (1956), Fishman and Huang (1983) and Geweke (1988). Basically all of these algorithms take the form of generating a value $\boldsymbol{x}$ from some w and then computing

$$\frac{(mf)^{\mathcal{T}}(\boldsymbol{x})}{w(\boldsymbol{x})} = \frac{1}{m}\sum_{i=1}^{m} \frac{m(T_i(\boldsymbol{x})f(T_i(\boldsymbol{x})J_{T_i}(\boldsymbol{x})}{w(\boldsymbol{x})}$$

where $\mathcal{T} = \{T_1, \ldots, T_m\}$ is a class of transformations $T_i : R^k \to R^k$. The benefit in this technique is that, with an appropriate choice of $\mathcal{T}$, substantial variance reductions can occur. Choosing $\mathcal{T}$ to be a subgroup of the group of volume-preserving symmetries of w ensures that $(mf)^{\mathcal{T}}$ and w are more alike as $(mf)^{\mathcal{T}}$ is also invariant under this subgroup. This can be viewed as the primary motivation for the method. Thus we require that the Jacobian determinant of T_i is 1, $w \circ T_i = w$, and $T_i \circ T_j \in \mathcal{T}$ for all i and j. We restrict attention here to the application of this technique to Student importance sampling. A more general treatment can be found in Evans and Swartz (1995b).

Again suppose that we have made the transformation to $\boldsymbol{y} = \hat{\Sigma}^{-1/2}(\boldsymbol{x} - \hat{\boldsymbol{u}})$ in the original expression for $I(m)$. Let f^* and m^* denote the transformed functions f and m respectively and use the importance sampler given by $\mathrm{y} \sim \sqrt{\frac{\lambda-2}{\lambda}}t_{k,\lambda}$. We then make the polar transformation $r = \|\boldsymbol{y}\|$, $\boldsymbol{v} = \boldsymbol{y}/\|\boldsymbol{y}\|$ and write

$$I(m) = \int_{S^{k-1}} \int_0^{\infty} r^{k-1} m^*(r\boldsymbol{v}) f^*(r\boldsymbol{v})\, dr d\boldsymbol{v}$$

where S^{k-1} is the surface of the $(k-1)$-dimensional unit sphere and $d\boldsymbol{v}$ is Lebesgue measure on S^{k-1}. Notice that the above Student distribution for $\boldsymbol{y}$ is equivalent to $\boldsymbol{v} \sim Unif(S^{k-1})$ statistically independent of $r^2 \sim k\frac{\lambda-2}{\lambda}F(k, \lambda)$.

Let the density and distribution function of r be denoted by h and H respectively. Then making the change of variables $s = H(r)$, we have that Student importance sampling is equivalent to generating $s \sim Unif(0,1)$, $\boldsymbol{v} \sim Unif(S^{k-1})$ and averaging

$$\frac{(H^{-1}(s))^{k-1} m^*(H^{-1}(s)\boldsymbol{v}) f^*(H^{-1}(s)\boldsymbol{v})}{h(H^{-1}(s))(2\pi^{k/2}/\Gamma(k/2))^{-1}}. \tag{8}$$

Now consider a finite group $\mathcal{T}$ consisting of symmetries of the importance sampler w defined on $(s, \boldsymbol{v})$. We specify a group formed as the direct product $\mathcal{T} = \mathcal{T}_1(m_1) \times \mathcal{T}_2(m_2)$ of a group $\mathcal{T}_1(m_1)$ of symmetries of the $U(0,1)$ distribution for s and a group $\mathcal{T}_2(m_2)$ of symmetries of the $Unif(S^{k-1})$ distribution for $\boldsymbol{v}$. For s, let $\mathcal{T}_1(m_1) = \{T_{11}, \ldots, T_{1m_1}\}$, where $T_{1,i}(s) = \frac{i-1}{m_1} \oplus s$ and $\oplus$ denotes addition modulo 1, as this group leaves the $U(0,1)$ distribution invariant. This group is essentially rotation sampling as discussed in Fishman and Huang (1983). For $\boldsymbol{v}$, let $\mathcal{T}_2(m_2) = \{T_{21}, \ldots, T_{2m_2}\}$ be a finite subgroup of $O(k)$, the group of $k \times k$ orthogonal matrices and note that $O(k)$ leaves the distribution of $\boldsymbol{v}$ invariant. The systematic sampling estimator is given by

$$\frac{1}{m_1 m_2} \sum_{i_1=1}^{m_1} \sum_{i_2=1}^{m_2} q(T_{1i_1}(s), T_{2i_2}(\boldsymbol{v})) \tag{9}$$

where $s \sim Unif(0,1)$, $\boldsymbol{v} \sim Unif(S^{k-1})$ and the function q is given by (8). The estimator (9) is then averaged over subsequent samples.

When compared to (8) the systematic sampling estimator (9) can be shown to always provide a variance reduction. However (9) requires $m_1 m_2$ times as many function evaluations as (8). Thus we say that systematic sampling gives a true variance reduction only when it's variance multiplied by $m_1 m_2$ is less than the variance of (8).

The immediate question concerns the choice of the groups $\mathcal{T}_1(m_1)$ and $\mathcal{T}_2(m_2)$. The theory in Evans and Swartz (1995b) suggests that we should attempt to find groups under which (8) is "far" from being invariant. Hopefully in a given problem, the structure is such that insight is available on the choice of the group. For example, when a high correlation exists between two variables it makes sense to consider groups of rotations on the cross-sectional plane as such rotations remove the correlations; see Evans and Swartz (1995b).

Systematic sampling does not always lead to true variance reductions. For example, if for fixed $\boldsymbol{v}$ the integrand (8) varies little as s changes, then $\mathcal{T}_1(m_1)$ just wastes function evaluations. The real value of the technique arises when this is not the case; i.e. when the importance sampler is poor. In general, systematic sampling can be viewed as a kind of insurance against poor choices of importance samplers. It may lead to a degradation of performance, bounded above by the size of the group, but it can lead to unboundedly large improvements in contexts where the importance sampler is not very good.

We tried various systematic sampling estimators in our linear model example. Keeping in mind the asymptotics that suggest approximate normality for the posterior, it is perhaps not reasonable to expect huge variance reductions in this problem. Table 1 provides estimated efficiencies of systematic sampling compared to (8) based on equivalent numbers of function evaluations. The three efficiency readings refer to the estimation of $I(1)$, $R(x_{10})$ and $R(x_7^2 - x_9^2)$. The group $\mathcal{T}_1(m_1)$ is an attempt to integrate the radial component of the integrand and the table indicates only modest asymmetry for this component from the uniform. The group $\mathcal{T}_2(4)$ acting on $\boldsymbol{v}$ was introduced with the idea that the 10-th coordinate in the linear model is somewhat problematic. For example, the Laplace approach has a relative error of 218% in estimating $R(x_{10})$. The group $\mathcal{T}_2(4)$ consists of the identity, a reflection in the 10th coordinate axis, a reflection through the origin and a simultaneous reflection in the first 9 axes. Effectively this group symmetrizes x_{10} and integrates all spherical harmonics of odd degree; see Evans and Swartz (1995b). We also considered the group $\mathcal{T}_3(m_2)$ acting on $\boldsymbol{v}$ which keeps all coordinates fixed except for v_7 and v_9 which are rotated through m_2 equally spaced angles. This group was developed as being appropriate for the estimation of $R(x_7^2 - x_9^2)$ as it can be shown, see Evans and Swartz (1995b), that such a group exactly integrates such a quadratic form under certain circumstances. As expected, this group did extremely well.

Group	Efficiencies		
$\mathcal{T}_1(2) \times \mathcal{T}_2(1)$	1.13	1.02	1.70
$\mathcal{T}_1(10) \times \mathcal{T}_2(1)$	.45	.43	.61
$\mathcal{T}_1(50) \times \mathcal{T}_2(1)$	.09	.08	.13
$\mathcal{T}_1(1) \times \mathcal{T}_2(4)$	.64	.70	1.53
$\mathcal{T}_1(2) \times \mathcal{T}_2(4)$	.89	.70	1.44
$\mathcal{T}_1(1) \times \mathcal{T}_3(3)$	.36	.39	5.36

Table 1

Experience with these groups and others suggests that only low order groups be used unless there is a specific reason to do otherwise. The natural alternative of sampling from high-order groups has been shown in Evans and Swartz (1995b) to lead to no true variance reductions whatsoever. Evans and Swartz (1995b) discuss the construction of other low-order subgroups of $O(k)$ to integrate out specific spherical harmonics. While the technique is promising it cannot be used blindly as it can be substantially deleterious. Further research is warranted on this issue.

5 Conclusions

There are many situations in applied Bayesian inference where Student importance sampling is not only adequate but preferable over other methods such as asymptotics or Markov chain methods; see Evans and Swartz (1995a) for further discussion of this issue. This argues for the inclusion of Student importance sampling as part of any general purpose software for Bayesian integration problems. As indicated, the standard variance reduction techniques can be helpful when applied to Student importance sampling but this cannot be said without caveats. At the very least, any general purpose software package should therefore offer these as options for the practitioner. We note that there is one further general class of variance reduction methods that we have not discussed; namely, randomized quadrature rules. This technique is discussed in Genz and Monahan (1994).

6 References

Evans, M. and Swartz, T. (1995a). Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. To appear in Statistical Science.

Evans, M. and Swartz, T. (1995b). Systematic Sampling in Monte Carlo. Manuscript.

Fishman, G.S. and Huang, B.D. (1983). Antithetic variables revisited. Communications of the ACM, Vol. 26, No. 11, 964-971.

Genz, A. and Monahan, J. (1995). Random integration rules for statistical computation. Computing Science and Statistics, 26, 135-138.

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

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

Hammersley, J.M. and Morton, (1956). A new Monte Carlo technique: antithetic variates. Proc. Camb. Phil. Soc., 52, 449-475.