INTEGRATION OF MULTIMODAL FUNCTIONS BY MONTE CARLO IMPORTANCE SAMPLING

by

Man–Suk Oh
University of California
Berkeley

and

James O. Berger
Purdue University

Technical Report # 91-31C

Department of Statistics
Purdue University

June 1991

Integration of Multimodal Functions by Monte Carlo Importance Sampling*

Man-Suk Oh
University of California, Berkeley

James O. Berger
Purdue University

April 25, 1991

Abstract

Numerical integration of a multimodal integrand $f(\theta)$ is approached by Monte Carlo integration via Importance Sampling. A mixture of multivariate t density functions is suggested as an importance function $g(\theta)$, for its easy random variate generation, thick tails, and high flexibility. The number of components in the mixture is determined by the number of modes of $f(\theta)$, and the mixing weights and location and scale parameters of the component distributions are determined by numerical minimization of a Monte Carlo estimate of the squared variation coefficient of the weight function $f(\theta)/g(\theta)$. Stratified Importance Sampling and control variates are shown to be particularly effective variance reduction techniques in this case. The algorithm is applied to a 10-dimensional example and shown to yield significant improvement over usual integration schemes.

*Research was supported by the National Science Foundation, grants DMS-8717799 and DMS-8923071.

Key Words: Mixture, Numerical integration, Stratified importance sampling, Control variate, Weight function.

1 Introduction

Many statistical inference problems reduce to the calculation of integrals of the form

$$E\varphi = \frac{\int \varphi(\theta) f(\theta) d\theta}{\int f(\theta) d\theta}, \tag{1}$$

where $\theta \in R^p$, $\varphi(\theta)$ is a measurable function, and $f(\theta)$ is proportional to a density function. In Bayesian analysis, $E\varphi$ arises as posterior expectation with respect to the posterior density proportional to $f(\theta) = l(\theta; x)\pi(\theta)$, $l(\theta; x)$ and $\pi(\theta)$ being the likelihood and prior for θ, respectively.

Among the many strategies developed for numerical integration of (1) are Monte Carlo Integration via Importance Sampling, briefly, Importance Sampling, recent references including Kloek and van Dijk (1978), Stewart (1979, 1983), van Dijk and Kloek (1980, 1983), van Dijk, Kloek, and Boender (1985), Geweke (1988, 1989), Evans (1989, 1991a, 1991b), Oh and Berger (1989); Quadrature methods, recent references including Naylor and Smith (1982, 1988), Smith et al (1985); sampling-based methods, recent references including Geman and Geman (1984), Tanner and Wong (1987), Gelfand and Smith (1990a, 1990b); and approximation methods, recent references including Tierney and Kadane (1986), Tierney, Kass and Kadane (1989).

Most of methods that have been developed are designed for a unimodal integrand f. When f is multimodal there can be serious difficulties in finding a good numerical scheme (see van Dijk and Kloek, 1978). In this paper, we approach the multimodal problem via Importance Sampling with a mixture importance function.

Importance Sampling can be described as follows. Choose a density function $g(\theta)$, called the importance function, and estimate (1) by

$$\hat{E}\varphi = \frac{\sum_{i=1}^{n} \varphi(\theta_i) w(\theta_i)}{\sum_{i=1}^{n} w(\theta_i)}, \tag{2}$$

where the weight function $w(\theta)$ is defined as $f(\theta)/g(\theta)$ and $\theta_i,\ i = 1, ..., n$, are identically distributed random samples from the density function $g(\theta)$. Under mild conditions, $\hat{E}\varphi$ converges to $E\varphi$ with probability one, as $n \to \infty$, and has asymptotic variance σ^2/n, where

$$\sigma^2 = \frac{1}{\int f(\theta) d\theta} \left[var_g(\varphi w) - 2(E\varphi) \cdot cov_g(\varphi w, w) + (E\varphi)^2 \cdot var_g(w) \right]; \tag{3}$$

the subscript g in $var_g()$ and $cov_g()$ indicates that the variances and covariances are taken with respect to the density function $g(\theta)$. The efficiency and accuracy of Importance Sampling clearly depends on the choice of $g(\theta)$. Typically desirable properties of $g(\theta)$ are; (i) it is easy to generate random samples from, (ii) it has tails that are heavier than those of f, (iii) it is a good approximation to f.

For a multimodal f, it is not easy to find a good importance function. There are some studied multimodal density functions that have been utilized for Importance Sampling, such as the poly-t density function (Bauens and Richard 1985), but it can be difficult to generate random variates from these densities and hard to fit them to the multimodal f. Dividing the region of integration into separate regions, each of which contains a mode of f, and applying separate Monte Carlo integration schemes to each region is sometimes used for multimodal f. However, this method has several difficulties. First, it is difficult to divide the region appropriately in many cases. Second, it can waste many of the generated random variates unless the modes are very well separated, not only because random variates outside of each region are *"rejected"* but also because it is difficult to approximate f within each region for Monte Carlo

Importance Sampling and difficult to allocate the random variate generations among the regions appropriately.

In the present paper, a mixture of multivariate t density functions is suggested as an importance function to deal with a multimodal integrand f, due to its easy random variate generation, thick tails, and the high flexibility of mixtures in matching multimodal f. (The basic algorithm could use mixtures of other densities if desired.) The number of component density functions in the mixture is determined by the number of modes of $f(\theta)$. One of the chief advantages of a mixture importance function is that one can then also use stratified Importance Sampling and control variates, greatly improving the accuracy of the Monte Carlo estimate.

The heart of the algorithm that is developed, overcoming the key difficulty in using a mixture importance function, is the development of a highly efficient method of selecting the mixing weights and the location and scale parameters of the component t-densities so that the mixture importance function "*fits*" the multimodal function $f(\theta)$. This will be done by minimizing a Monte Carlo estimate of the squared variation coefficient of the weight function, which can be argued to be the correct measure of "*fit*" for Importance Sampling.

The paper is organized as follows. Section 2 provides a brief description of a mixture density function as a candidate for the importance function. Also, stratified Importance Sampling, and use of control variates with the mixture importance function are described. The algorithm for matching the mixture to f is given in Section 3. A 10-dimensional example is presented in Section 4. Section 5 gives a summary and conclusions.

2 Use of Stratification and Control Variates for Mixture Importance Functions

2.1 Mixture Importance Functions

A mixture density function $g(\theta)$, $\theta \in \Theta$, will be written as

$$g(\theta) = \varepsilon_1 g_1(\theta) + \ldots + \varepsilon_m g_m(\theta), \quad \theta \in \Theta, \tag{4}$$

where

$$\varepsilon_i > 0,\ i = 1, ..., m; \quad \varepsilon_1 + \ldots + \varepsilon_m = 1;$$

and $g_i(\theta)$, $i = 1, ..., m$, are density functions. The parameters $\varepsilon_1, ..., \varepsilon_m$ are the mixing weights; $g_1(\theta), \ldots, g_m(\theta)$ are the component density functions of the mixture; and m is the number of components.

We will consider in this paper the case $g_i = g_{\xi_i} \in \mathcal{G} = \{g_\xi,\ \xi \in \Xi\}$, $\mathcal{G}$ being the set of multivariate t density functions. Here, $\xi_i = (\alpha_i, \mu_i, T_i)$ where α_i is the degrees of freedom, μ_i is the location vector and T_i is a lower triangular matrix such that $T_i T_i'$ is the scale matrix of the multivariate t density function g_i. Any other parametric family of density functions could be considered in the same framework, but the t-family should typically suffice.

Generation of a random variate θ from a mixture $g(\theta)$ is easy;

- Generate a uniform random variate u in the interval $(0, 1)$.
- If, defining $\varepsilon_0 = 0$,

$$\sum_{j=1}^{i-1} \varepsilon_j \leq u \leq \sum_{j=1}^{i} \varepsilon_j,$$

then generate θ from $g_i(\theta)$.

Thus, one needs to generate θ from only one of $g_i(\theta)$, $i = 1, ..., m$, plus generate a uniform random variate u. This makes the generation of random variates from a mixture very efficient compared with generation from other multimodal density functions.

General discussion about finite mixtures of density functions is given in Everitt and Hand (1981), Titterington, Smith, and Makov (1985).

2.2 Stratified Importance Sampling

With a mixture importance function, one may generate random samples from $g(\theta)$ as described in the previous section and estimate $E\varphi$ by (2). However, a great improvement in the accuracy and efficiency of Monte Carlo estimation with a mixture importance function can be made using stratified Importance Sampling. With $g(\theta) = \sum_{i=1}^{m} \varepsilon_i g_i(\theta)$, $E\varphi$ can be written as

$$E\varphi = \frac{\sum_{i=1}^{m} \varepsilon_i \, E_{g_i} \varphi w}{\sum_{i=1}^{m} \varepsilon_i \, E_{g_i} w}.$$

Stratified Importance Sampling approximates $E\varphi$ by

$$\hat{E}^s \varphi = \frac{\sum_{i=1}^{m} \varepsilon_i \, \overline{(\varphi w)}_i}{\sum_{i=1}^{m} \varepsilon_i \, \overline{w}_i}, \tag{5}$$

where

$$\overline{(\varphi w)}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \varphi(\theta_j^{(i)}) w(\theta_j^{(i)}) \tag{6}$$

$$\overline{w}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} w(\theta_j^{(i)}), \tag{7}$$

and $\theta_1^{(i)}, \ldots, \theta_{n_i}^{(i)}$ are i.i.d. random samples from $g_i(\theta)$, for $i = 1, ..., m$. As can be seen from Theorems A.1 and Corollary A.2 in the appendix, under some mild conditions the stratified Importance Sampling estimate $\hat{E}^s \varphi$ is consistent for $E\varphi$ and has

asymptotic variance

$$\sum_{i=1}^{m} \varepsilon_i^2 \sigma_i^2 / n_i, \tag{8}$$

where

$$\sigma_i^2 = \frac{1}{(\int f(\theta)d\theta)^2} \left(var_{g_i}(\varphi w) + (E\varphi)^2 var_{g_i}(w) - 2(E\varphi) \cdot cov_{g_i}(\varphi w, w) \right). \tag{9}$$

If $n = \sum_{i=1}^{m} n_i$ is the total number of random samples that can be taken, the optimal allocation of n_i, which minimizes the variance among all possible allocations, is $n_i \propto \varepsilon_i \sigma_i$. The asymptotic variance of $\hat{E}^s \varphi$ is then σ_o^2/n, where

$$\sigma_o^2 = (\sum_{i=1}^{m} \varepsilon_i \sigma_i)^2. \tag{10}$$

While optimal, the dependence of n_i on σ_i^2 causes two difficulties. First, the σ_i^2 would themselves need to be estimated. Second, different φ would yield different σ_i^2, while in practice one typically wants to use the same importance sample to compute $E\varphi$ for a variety of φ. A pragmatic solution to this second problem is to replace σ_i^2 by

$$\sigma_i^{w2} = var_{g_i}(w)/(\int f(\theta)d\theta)^2. \tag{11}$$

Note that σ_i^{w2} is the term of σ_i^2, when scaled by $(E\varphi)^2$, which does not involve $\varphi(\theta)$, and that the allocation $n_i \propto \varepsilon_i \sigma_i^w$ is optimal in estimating $\int f(\theta)d\theta$. Since $\varphi(\theta)$ is often more slowly varying than $f(\theta)$, we have found that in practice use of the σ_i^w is quite satisfactory.

To avoid estimation of the σ_i or σ_i^w, one could choose the proportional allocation $n_i \propto \varepsilon_i$. The asymptotic variance of $\hat{E}^s \varphi$ is then σ_p^2/n, where

$$\sigma_p^2 = \sum_{i=1}^{m} \varepsilon_i \sigma_i^2. \tag{12}$$

It can be easily shown (Cochran, 1963) that

$$\sigma_o^2 \leq \sigma_p^2 \leq \sigma^2.$$

Thus, this simple allocation has a larger variance than the optimal allocation but is superior to ordinary (non-stratified) Importance Sampling for any $\varphi(\theta)$.

2.3 Use of Control Variates

When the mixture importance function is well matched to f (the algorithm achieving this will be described in Section 3.1), one can consider using $E_g\varphi = \int \varphi(\theta)g(\theta)d\theta$ as a control variate to reduce the variance of the Importance Sampling estimate. Because $g(\theta)$ is a mixture of t density functions, analytic evaluation of $E_g\varphi$ is possible for many φ of interest. In such cases, one may estimate $E\varphi$ by

$$\hat{E}^c\varphi = \hat{E}^s\varphi - \hat{E}^s_g\varphi + \int \varphi(\theta)g(\theta)d\theta, \tag{13}$$

where $\hat{E}^s\varphi$ and $\hat{E}^s_g\varphi$ are the stratified Importance Sampling estimates of $E\varphi$ and $E_g\varphi$, respectively. Obviously, $\hat{E}^c\varphi$ converges to $E\varphi$ with probability one as $n \to \infty$ and the variance of $\hat{E}^c\varphi$ is given by

$$var_g(\hat{E}^c\varphi) = var_g(\hat{E}^s\varphi) + var_g(\hat{E}^s_g\varphi) - 2cov_g(\hat{E}^s\varphi, \hat{E}^s_g\varphi). \tag{14}$$

As is usual with control variates, $\hat{E}^c\varphi$ will have a smaller variance than $\hat{E}^s\varphi$ if f and g are similar in shape.

One can actually proceed quantitatively, using Cramer (1948), by noting that

$$\hat{E}^s\varphi \approx E\varphi + \frac{1}{\int f(\theta)d\theta}\left(\sum_{i=1}^m \varepsilon_i\overline{(\varphi w)}_i\right) - \left(\frac{E\varphi}{\int f(\theta)d\theta}\right)(\overline{w}_i),$$

and that the covariance term in (14) is

$$cov_g(\hat{E}^s\varphi, \hat{E}^s_g\varphi) \approx \frac{1}{\int f(\theta)d\theta}\sum_{i=1}^m \frac{\varepsilon_i^2}{n_i}\left[cov_{g_i}(\varphi w, \varphi) - (E\varphi)\cdot cov_{g_i}(w, \varphi)\right]. \tag{15}$$

Therefore one can estimate $var_g(\hat{E}^c\varphi)$ and compare it with $var_g(\hat{E}^s\varphi)$ and select the one among $\hat{E}^c\varphi$ and $\hat{E}^s\varphi$ with the smaller variance as a final estimate of $E\varphi$. Note also

that the extra computation of $\hat{E}_g^s\varphi$ and estimation of $var_g(\hat{E}_g^s\varphi)$ and $cov_g(\hat{E}^s\varphi, \hat{E}_g^s\varphi)$ involve only some additional summations and multiplications, and therefore are often inexpensive.

General discussion about control variates is given in Rubinstein (1981). See also Tew and Wilson (1988) and Swain and Schmeiser (1988, 1989).

3 Selection of the Mixture Importance Function

3.1 The Algorithm

A brief description of the algorithm for selecting the parameters of the mixture importance function is presented in this section. More details of and justification for the algorithm will be given in next section.

Define $\lambda = \{(\varepsilon_i,\ \mu_i,\ T_i),\ i = 1, \ldots, m\}$. It is convenient from here on to write $g(\theta, \lambda)$ instead of $g(\theta)$ and $w(\theta, \lambda) = f(\theta)/g(\theta, \lambda)$ instead of $w(\theta)$. The notation $\mathcal{T}_{\alpha_i}(0, I)$ will be used to denote the standard multivariate t distribution with α_i degrees of freedom.

- Step 1 (Initialization). Choose an initial $g(\theta, \lambda)$ by selecting m, the α_i, and and an initial λ. Also specify N, l, and η, three computational constants. See Section 3.2.2 for details of this initialization.

- Step 2. Generate N random samples, $\{z_j^{(i)}\}_{j=1}^N$, from $\mathcal{T}_{\alpha_i}(0, I)$, and store them in the i-th column of a table, for $i = 1, ..., m$.

- Step 3 Minimize $\hat{CV}^2(w; \lambda, \mathbf{z}, N)$ with respect to λ, where

$$\hat{CV}^2(w; \lambda, \mathbf{z}, N) \ = \ \frac{\sum_{i=1}^m \varepsilon_i \ \hat{var}_{g_i}(w)}{(\sum_{i=1}^m \varepsilon_i \ \overline{w}_i)^2}, \tag{16}$$

$$N_i = [\varepsilon_i N], \quad \text{the integer part of } \varepsilon_i N, \tag{17}$$

$$\overline{w}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} w(\theta_j^{(i)}, \lambda), \tag{18}$$

$$\hat{var}_{g_i}(w) = \frac{1}{N_i}\sum_{j=1}^{N_i} w^2(\theta_j^{(i)}, \lambda) - (\overline{w}_i)^2, \tag{19}$$

$$\theta_j^{(i)} = T_i z_j^{(i)} + \mu_i;\ j = 1, ..., N,\ i = 1, ..., m, \tag{20}$$

$$\mathbf{z} = \{z_j^{(i)},\ j = 1, ..., N,\ i = 1, ..., m\}, \tag{21}$$

but monitor the minimization after each l steps, and stop if the relative reduction in $\hat{CV}^2(w; \lambda, \mathbf{z}, N)$ from the previous to current monitored values is smaller than a desired accuracy η.

3.2 Justification for the Algorithm

3.2.1 The Criterion for Matching the Mixture to f and its Implementation

For the mixture importance function, it is desirable to choose the parameters, m, $\alpha_i, \varepsilon_i,\ \mu_i, T_i,\ i = 1, \ldots, m$, to satisfy the desirable properties (i)-(iii) described in Section 1. Property (i) is automatically satisfied by the mixture form and (ii) can be easily obtained by appropriate choice of the α_i (see Section 3.2.2). As previously mentioned, the number of components, m, will be chosen to be the number of modes of $f(\theta)$. The most crucial and difficult problem is thus the selection of $\varepsilon_i,\ \mu_i, T_i$, $i = 1, \ldots, m$, for given m and α_i's.

In Importance Sampling, the degree of mimicry of g to f is well reflected by the inverse of the squared variation coefficient of the weight function $w(\theta, \lambda)$ (see Evans

1991b), namely

$$CV^2(w;\lambda) = N \cdot CV^2(\hat{E}_g w;\lambda) = \frac{N \cdot var(\hat{E}_g w)}{(E_g w)^2}. \tag{22}$$

In stratified Importance Sampling, $\hat{E}_g w = \sum_{i=1}^m \varepsilon_i \overline{w}_i$ and $E_g w = \sum_{i=1}^m \varepsilon_i E_{g_i} w$, so that (22) becomes

$$CV^2(w;\lambda) = \frac{N \cdot \sum_{i=1}^m \varepsilon_i^2 \; var_{g_i}(w)/N_i}{(\sum_{i=1}^m \varepsilon_i \; E_{g_i} w)^2}. \tag{23}$$

Of course, we cannot calculate $CV^2(w;\lambda)$ exactly, much less minimize it over the parameters of g, but we can minimize a Monte Carlo estimate of it, using $g(\theta,\lambda)$ itself as the importance function in the Monte Carlo estimation. (Using a Monte Carlo estimate based on generation from g is needed for efficiency.) Considerable care must be taken, however, in constructing the Monte Carlo estimates. In particular, for efficiency and numerical stability in the minimization, the random variates used in the Monte Carlo estimates must be fixed for all λ. Fixing the random variates is straightforward; generate and table a random set $\mathbf{z}$ from the $\mathcal{T}_{\alpha_i}(0,I)$ initially, and keep using it to compute the estimate of $CV^2(w;\lambda)$ in each step of the minimization. Note that these variates are transformed in Step 3 so that they are effectively generated by the importance function $g(\theta,\lambda)$. Thus one simultaneously has the efficiency of Monte Carlo sampling and yet is essentially minimizing a fixed (non-random) function of λ.

The optimal allocation described in Section 2.2 is impractical to use here because, given λ in each step of minimization, one would need a Monte Carlo run to estimate σ_i^w and then, with $N_i \propto \varepsilon_i \sigma_i^w$, another Monte Carlo run to compute $\hat{CV}^2(w;\lambda,\mathbf{z},N)$. (In each step of minimization, one will get a different λ, hence a different σ_i^w.) Hence we adopted the proportional allocation $N_i \propto \varepsilon_i$. Thus the choice $N_i = [\varepsilon_i N]$, the integer part of $\varepsilon_i N$ for simplicity, is used in the algorithm, and the estimate of (23) becomes (16).

Control variates cannot be used here since assuming $g \propto f$ leads to a constant $CV^2(w;\lambda)$, independent of θ. Note, however, that it will be used for the actual Importance Sampling to compute $E\varphi$ after $g(\theta)$ is obtained from the algorithm in Section 3.1.

A possible theoretical concern is whether the minimum of $\hat{CV}^2(w;\lambda,\mathbf{z},N)$ converges to the minimum of $CV^2(w;\lambda)$ as $N \to \infty$. Conditions under which this can be assured are given in Theorem A.3 in the appendix.

3.2.2 Initialization

• Initialization of λ: Because $\hat{CV}^2(w;\lambda,\mathbf{z},N)$ may well have local minima, good initialization of λ is important. Fortunately, the maximum likelihood method seems to work well for initialization of μ_i and T_i. Assume that f has k modes, $\hat{\theta}_1,...,\hat{\theta}_k$, and minus inverse Hessians, $-\hat{I}_1^{-1},...,-\hat{I}_k^{-1}$, at the modes. It is reasonable to choose k components for the mixture importance function, i.e., $m = k$ (this will be discussed later). Most crucial to avoiding a bad local minimum is choice of a good starting location μ_i^0 for the μ_i. The modes, $\hat{\theta}_i$, $i = 1,...,k$, seem to work very well, and even if the procedure gets stuck in a local minimum with μ_i near $\hat{\theta}_i$, $i = 1,...,k$, it is likely to be a good approximation to f. The positive definite lower triangular matrix T_i^0 such that $(T_i^0)(T_i^0)' = -\hat{I}_i^{-1}$ seems to be a reasonable starting value for T_i. This choice of T_i^0 roughly minimizes the variance of $w(\theta)$ around $\hat{\theta}_i$, assuming that f has approximately normal shape around $\hat{\theta}_i$. Also, it seems to be a good compromise between matching the variance of g_i to that of f around $\hat{\theta}_i$ (assuming that the latter is well approximated by $-I_i^{-1}$) and matching the Hessian of g_i to that of f at $\hat{\theta}_i$. (See Oh, 1991, for more details.) Finally, $\varepsilon_i^0 \propto f(\hat{\theta}_i)|T_i^0|\alpha_i^{p/2}\Gamma(\alpha_i/2)/\Gamma((\alpha_i+p)/2)$ (simply $\varepsilon_i^0 \propto f(\hat{\theta}_i)|T_i^0|$, if the α_is are equal), where $|T_i^0|$ is the determinant of T_i^0, is a

reasonable initial value for ε_i, essentially matching the heights of g_i and f at $\hat{\theta}_i$.

• Selection of m: An obvious choice for m would be the number of modes, k, of $f(\theta)$. Of course, it might be possible to obtain a better approximation to f using a number larger than k, especially when f is skewed. Thus, after the algorithm in Section 3.1 is run with $m = k$, one could increase m by one, initialize the new parameters $\varepsilon_i, i = 1, .., m$, μ_m, T_m, add another column of random variates to the table, and repeat Step 3. One could continue increasing m in this manner until $\hat{CV}^2(w; \lambda, \mathbf{z}, N)$ is stabilized, hopefully at some overall mixture minimum. This has proven to be difficult to implement because of the lack of any natural initial values for these extra components.

• Selection of α_i: The degrees of freedom, α_i, can be chosen from a preliminary study of the tail rates of $f(\theta)$. If information about the tail rates of $f(\theta)$ is not available, rather low degrees of freedom should be used for safety.

• Selection of N: The total number of random samples, N, should be moderate since computation of $\hat{CV}^2(w; \lambda, \mathbf{z}, N)$ is needed at each step of the minimization over λ. Also, highly accurate $\hat{CV}^2(w; \lambda, \mathbf{z}, N)$ is not necessary here because the purpose of the minimization routine is only to find a good importance function to be used in the actual Importance Sampling. In practice, we simply run several Monte Carlos to compute $\hat{CV}^2(w; \lambda, \mathbf{z}, N)$, with the initial λ and several different random seeds (i.e., different sets $\mathbf{z}$), and choose N so that the resulting values of $\hat{CV}^2(w; \lambda, \mathbf{z}, N)$ are roughly stable.

• Selection of l: There is no clear rule to determine the number of steps between the monitorings of the progress of the minimization. From our experience, letting l equal about $5 \sim 10$ times the number of elements in λ, i.e., the number of variables in the minimization, seems to work well.

• Selection of η: Note that it is not necessary to actually find the g which minimizes (16), since we are only going to use it as an importance function and any g for which $\hat{CV}^2(w;\lambda,\mathbf{z},N)$ is close to its minimum will serve well. Thus, we do not need to carry out the minimization of $\hat{CV}^2(w;\lambda,\mathbf{z},N)$ to high precision; indeed, all that is practically necessary is to have a g for which $\hat{CV}^2(w;\lambda,\mathbf{z},N)$ is within a small fraction of optimum. Therefore, $\eta = 0.1$ or 0.2 will typically be satisfactory. (Note that the resulting g, when used in the actual Importance Sampling run, will only be sacrificing roughly 10% to 20% efficiency.)

3.2.3 Parameter Reductions

There are $m(p+p(p+1)/2+1)-1 \approx m(p+3/2)^2/2$ elements in $\lambda = (\varepsilon_i,\ \mu_i,\ T_i,\ i = 1,...,m)$. Thus, minimization with respect to λ can be very expensive when p is large. A solution to this problem is to remove unimportant parameters of the mixture from the minimization, essentially by fixing their values. Oh (1991) indicates that the covariance matrices have less effect on the accuracy of the Importance Sampling estimates than the location μ_i's. Thus, when p is large it would be reasonable to fix some elements of the T_i's and exclude them from the minimization.

An obvious possibility is to let $T_i = T_i^0 D_i$, where D_i is a variable diagonal matrix and T_i^0 is as given in Section 3.2.2, and minimize with respect to D_i instead of T_i, i.e., let $\lambda = (\varepsilon_i, \mu_i, D_i,\ i = 1,...,m)$. The initial value of D_i would be chosen to be the $p \times p$ identity matrix I_p. This reduces the number of variables in the minimization to $m(2p+1)$. Our experience indicates that the resulting importance function is virtually as good as that resulting from unconstrained T_i.

4 An Example

"$m - n$" poly-t density functions have posterior kernels which are ratios of m multivariate t density kernels to n multivariate t density kernels. They arise as posterior densities in econometric problems (Dreze and Richard 1983, Bauens and Richard 1985), and can be multimodal. Because the integration of poly-t density functions is analytically intractable, numerical integration is necessary to obtain desired characteristics (moments, univariate marginal density functions, etc.).

For illustration, assume that $f(\theta)$ has the form

$$f(\theta) = \prod_{i=1}^{4} f_i(\theta), \tag{24}$$

where $\theta \in R^{10}$, and the $f_i(\theta)$ are the density functions of the $\mathcal{T}_1(\xi_i, \Sigma_i)$ distributions specified by

$$\xi_1 = \begin{pmatrix} (0.00)\ \mathbf{1}_5 \\ (0.00)\ \mathbf{1}_5 \end{pmatrix},\ \xi_2 = \begin{pmatrix} (-1.34)\ \mathbf{1}_5 \\ (0.45)\ \mathbf{1}_5 \end{pmatrix},\ \xi_3 = \begin{pmatrix} (-1.34)\ \mathbf{1}_5 \\ (-1.56)\ \mathbf{1}_5 \end{pmatrix},\ \xi_4 = \begin{pmatrix} (0.00)\ \mathbf{1}_5 \\ (-1.56)\ \mathbf{1}_5 \end{pmatrix}.$$

$$\Sigma_1 = \Sigma_4 = \frac{1}{1.7} I_{10}, \quad \Sigma_2 = \Sigma_3 = \frac{1}{0.7} I_{10};$$

here $\mathbf{1}_n$ is the n-vector of 1's and I_n is the $n \times n$ identity matrix. Maximizations of $f(\theta)$, with each ξ_i above as a starting point, showed two modes,

$$\hat{\theta}_1 = ((-0.122)\ \mathbf{1}_5', (-0.109)\ \mathbf{1}_5')', \tag{25}$$

$$\hat{\theta}_2 = ((-0.134)\ \mathbf{1}_5', (-1.413)\ \mathbf{1}_5')'. \tag{26}$$

The corresponding minus inverse Hessians, $-\hat{I}_i^{-1}$, are omitted to save space, but their diagonal elements are only about 0.06. Therefore, a multimodal importance function seems to be necessary.

- **Selection of the mixture importance function**

Since f has two modes, we chose $m = 2$. From (24), $f(\theta)$ can be seen to have tail rates of $(\theta'\theta)^{-4(1+10)/2}$. Because thicker (but not-too-thick) tails are preferable for $g(\theta)$ to ensure convergence in Importance Sampling and to handle possible skewness, we chose $\alpha_1 = \alpha_2 = 4$ so that the $g_i(\theta)$ have tail rates of $(\theta'\theta)^{-(4+10)/2}$. Because p is large, we set $T_i = T_i^0 D_i$, where T_i^0 is a positive definite lower triangular matrix such that $T_i^0 T_i^{0\prime} = -\hat{I}_i^{-1}$ and D_i is a variable diagonal matrix. For initialization, $f(\hat{\theta}_i)|T_i^0|/\sum_{i=1}^m f(\hat{\theta}_i)|T_i^0|$, $\hat{\theta}_i$, and I_{10} were chosen for ε_i, μ_i and D_i, respectively, as described in Section 3.2.2. From a few preliminary calculations of $\hat{CV}^2(w;\lambda,\mathbf{z},N)$, $N = 300$ seemed to give a roughly stable $\hat{CV}^2(w;\lambda,\mathbf{z},N)$ for different random sets $\mathbf{z}$. Thus $N = 300$ was chosen. Also we set $l = 210$, about 5 times the number of variables in the minimization. Finally we set $\eta = 0.2$, completing Step 1 of the algorithm in Section 3.1.

At Step 2 of the algorithm, 600 random samples were generated from $\mathcal{T}_4(0, I)$ and stored in a table of 2 columns and 300 rows. At Step 3, the NAG minimization routine E04JAF with $l = 210$ was run and the resulting $\hat{CV}^2(w;\lambda,\mathbf{z},N)$ was 3.967 (before the minimization, $\hat{CV}^2(w;\lambda,\mathbf{z},N)$ was 13.139). Another $l = 210$ steps of the minimization were performed, resulting in $\hat{CV}^2(w;\lambda,\mathbf{z},N)$ =2.325. Since 3.139/2.325 is not less than $1 + \eta = 1.2$, we continued the minimization. The algorithm stopped after a total of 6×210 steps of the minimization, and the final $\hat{CV}^2(w;\lambda,\mathbf{z},N)$ was 0.533, about 25 times smaller than the initial $\hat{CV}^2(w;\lambda,\mathbf{z},N)$. The total computing time for this matching algorithm was about 600 seconds. (All computations in this article were done on a SunOS 4.1 workstation with floating point accelerator at the University of California, Berkeley.)

- **Actual Importance Sampling**

With this "*matched*" mixture importance function it is now possible to run the

actual Importance Sampling. Suppose that we are interested in the posterior mean of θ and the posterior marginal distribution functions of θ_1 and θ_{10}, so that the φ of interest are the vector θ and the indicator functions $I(\theta_1 \leq x)$ and $I(\theta_{10} \leq x)$ for various x's.

Table 1 shows the estimated posterior means and standard deviations of the estimates resulting from stratified Importance Sampling with the proportional allocation $n_i \propto \varepsilon_i$ and a total Monte Carlo sample size of $n = 10,000$. (One could, of course, estimate σ_i^w from a preliminary run with the given mixture and use the optimal allocation. Note that the matching process provides an estimate of σ_i^w, but this estimate does not seem to be accurate enough for use in the actual Importance Sampling.) For comparison, results *with* and *without* $\int \varphi(\theta)g(\theta)d\theta$ as a control variate are shown.

To indicate the value of the matching algorithm, Table 1 also gives the results when the mixture importance function without the matching process, i.e., the mixture importance function with the initial value of λ, is used. Use of a control variate gave very little improvement in this case, so only the results from stratified Importance Sampling without control variates are presented in Table 1.

Clearly the matching algorithm greatly improves the accuracy. For illustrative comparison of the efficiencies, suppose that one wants the estimated posterior means to have relative Monte Carlo accuracy of 1% with probability 0.95, i.e.,

$$P(\frac{|\hat{E}^s\theta_i - E\theta_i|}{|E\theta_i|} \leq 0.01) \geq 0.95, \tag{27}$$

for all $i = 1, ..., m$ (here θ_i is the i-th element of θ). Then from the approximate normality of the stratified Importance Sampling estimate $\hat{E}^s\theta_i$ (see Corollary A.2 in the appendix), the standard deviation of the estimate should satisfy

$$\frac{SD(\hat{E}^s\theta_i)}{|E\theta_i|} \leq 0.01/2 = 0.005. \tag{28}$$

To achieve the desired accuracy, (28), the algorithm proposed here would require about $7 \times 10,000$ Monte Carlo observations so that the total computing time required would be about $600 + 7 \times 35 = 845$ seconds (600 seconds for the matching algorithm and 35 seconds for each 10,000 observations). Using the mixture without the matching process would require about $240 \times 10,000$ observations, hence about $240 \times 35 = 8,408$ seconds of computing time.

Finally, we also considered basic Importance Sampling with a unimodal multivariate t importance function, using the adaptive scheme described in Oh and Berger (1989) (non-adaptive Importance Sampling would be worse because of the well-separated modes). In the adaptive scheme, the location and scale matrix of the importance function were updated every 1,000 iterations. It appears that the unimodal importance function is yielding spuriously small standard errors; for instance the "*mixture with control*" and "*unimodal*" estimates for θ_6 differ by over 3.6 standard errors. This is probably caused by the unimodal importance function being centered in the "*valley*" between two modes, resulting in very inaccurate estimated Monte Carlo variances.

The posterior marginal distribution functions of θ_1 and θ_{10} were also computed for 20 equally spaced points. Their graphs are given as solid lines in Figures 1 and 2. As a graphical demonstration of the performance of the matching algorithm, the marginal distribution functions of θ_1 and θ_{10} from the *unmatched* mixture importance function (i.e., the initialization mixture) are given as the dashed lines, and from the *matched* mixture importance function as the dotted lines, in Figures 1 and 2. The NAG subroutine G01ABF was used to compute the marginal distribution function of a t-distribution. The matching algorithm seems to give a very good fit.

5 Summary and Conclusions

Dealing with high-dimensional multimodal integrands is notoriously difficult; numerical methods that are in common use today can easily break down with such integrands (without it being realized that there is a problem), as was partially indicated in Section 4 for even a sophisticated adaptive Importance Sampling scheme. To attack the problem we have combined a number of standard and nonstandard techniques.

Basing analysis on Importance Sampling with a mixture importance function is rather natural, and has the advantages that easy random variable generation is possible and that stratified Importance Sampling and control variates can be easily employed. We have adopted the rather simple scheme of simply choosing the number of components in the mixture to be the number of modes of the integrand. Note that, at a minimum, we thus assume a capability to identify the modes (or at least the important modes) of the integrand; this can, of course, itself be a difficult task but any method of dealing with multimodal integrands will probably be based on the assumption that the modes have been located. Our algorithm readily accommodates the possibility of adding additional components to the mixture, which can for instance be valuable in dealing with skewness. Unfortunately, we experienced difficulties in attempting to add such additional components, primarily because of the need to find a good starting point for the matching algorithm; modes are the only natural starting points.

To match the mixture importance function to the integrand, we started with the simple idea of using the modes and respective Hessians as initial location and scale parameters for the components of the mixture. The mixing weights of the components of this initial mixture were computed using the Hessians, degrees of freedom, and

the relative heights of the integrand at the modes. These choices provided good starting values for the ensuing matching algorithm; good starting values were found to typically be necessary for successful implementation of the algorithm.

At this point, one could use the mixture determined in this initialization as the importance function for the desired Monte Carlo integration, but we found that this initial mixture is frequently quite inefficient as an importance function. This led to the major development in the paper, the algorithm for matching the mixture to the integrand. In developing a matching algorithm, analytic methods seemed to hold little promise, since analytic matching methods typically involve an integral measuring fit. Hence we considered simulation-based methods of fitting.

The first problem was to define a reasonable measure of fit of the mixture importance function to $f(\theta)$, the important part of the integrand. Drawing on experience obtained in Oh and Berger (1989), we settled on $CV^2(w;\lambda)$ as a good generic measure of the fit. ($CV^2(w;\lambda)$ measures how well the importance function works in the Monte Carlo computation of the integral.) Thus the goal is to minimize $CV^2(w;\lambda)$ over the parameters, λ, of the mixture. The difficulty in doing so is that computation of $CV^2(w;\lambda)$ itself must be done by Importance Sampling; we did so using the mixture as an importance function. But for efficiency and numerical stability, the random variates used in the simulation must be fixed for the minimization over λ. Implementation of this idea became Step 3 of the algorithm.

The details of implementation were based on the realization that obtaining an optimal fit of the mixture to $f(\theta)$ is not necessary, since a fit that is only slightly better means that the final Importance Sampling will be only slightly more efficient. Hence the rule used to stop the matching algorithm is to stop when additional steps in the minimization are leading to only slight improvement in fit.

Appendix

Theorem A.1 *If (i) the support of g contains the support of f, (ii) $E_{g_i}w$ and $E_{g_i}(\varphi w)$ exist for each $i = 1, ..., m$, and (iii) $n_i \to \infty$ as $n \to \infty$, for each $i = 1, ..., m$, then $\hat{E}^s\varphi$ converges to $E\varphi$ with probability one as $n \to \infty$.*

Proof: Straightforward. □

Corollary A.1 *Suppose the conditions of the above theorem are satisfied with $\varphi = 1$, $n_i = N_i$ and $n = N$, where N_i is given in (17), and $E_{g_i}(w^2)$ exists for each $i = 1, ..., m$. Then $\hat{CV}^2(w; \lambda, \mathbf{z}, N)$, given in (16), converges to $CV^2(w; \lambda)$ in (23) with probability one, as $N \to \infty$.*

Theorem A.2 *In addition to conditions (i), (ii), (iii) of Theorem A.1, suppose (iv) $E_{g_i}(w^2)$ and $E_{g_i}[(\varphi w)^2]$ exist for each $i = 1, ..., m$, (v) $\lim_{n\to\infty}(n_i/n)$ exists and is nonzero for all $i = 1, ..., m$. Let $\bar{X}_n = \sum_{i=1}^m \varepsilon_i \bar{X}_{n_i}$, $\mu = \sum_{i=1}^m \varepsilon_i \mu_i$, $\Sigma_n = \sum_{i=1}^m \varepsilon_i^2 V_i/n_i$, where*

$$\bar{X}_{n_i} = \begin{pmatrix} \overline{(\varphi w)}_i \\ \overline{w}_i \end{pmatrix}, \quad \mu_i = \begin{pmatrix} E_{g_i}(\varphi w) \\ E_{g_i}(w) \end{pmatrix}, \quad V_i = \begin{pmatrix} var_{g_i}(\varphi w) & cov_{g_i}(\varphi w, w) \\ cov_{g_i}(\varphi w, w) & var_{g_i}(w) \end{pmatrix}, \tag{29}$$

where $\overline{(\varphi w)}_i$ and $\overline{w}_i$ are defined in (6) and (7). Then, for a continuous differentiable function h whose first partial derivative Δh is continuous at μ,

$$[(\Delta h(\mu))^t \Sigma_n (\Delta h(\mu))]^{-1/2}(h(\bar{X}_n) - h(\mu)) \longrightarrow N(0, I), \tag{30}$$

where $\Delta h(\mu)$ and $[\Delta h(\mu)]^t$ are values of Δh at μ and its transpose.

Proof: From the Central Limit Theorem, for each i, $\sqrt{n_i}(\bar{X}_{n_i} - \mu_i) \to N(0, V_i)$ as $n \to \infty$. Because of the condition (v), $\lim_{n\to\infty}(\Sigma_n^{-1/2}/\sqrt{n_i})$ exists. Therefore,

$$\Sigma_n^{-1/2}(\bar{X}_n - \mu) \quad = \quad \sum_{i=1}^m \varepsilon_i (\Sigma_n^{-1/2}/\sqrt{n_i}) \cdot \sqrt{n_i}(\bar{X}_{n_i} - \mu_i)$$

$$
\begin{aligned}
&\longrightarrow \sum_{i=1}^{m} \varepsilon_i [\lim_{n\to\infty} (\Sigma_n^{-1/2}/\sqrt{n_i})]\ N(0, V_i), \\
&\qquad \text{by Slutsky's Theorem (Bickel and Doksum 1977, A.14.9)}, \\
&= N(0,\ \sum_{i=1}^{m} \varepsilon_i^2\ [\lim_{n\to\infty} (\Sigma_n^{-1/2}/\sqrt{n_i})]\ V_i\ [\lim_{n\to\infty} (\Sigma_n^{-1/2}/\sqrt{n_i})]) \\
&= N(0,\ \lim_{n\to\infty} [\Sigma_n^{-1/2} (\textstyle\sum_{i=1}^{m} \varepsilon_i^2 V_i/n_i) \Sigma_n^{-1/2}]), \text{ by Bartle (1976, Thm 15.6)}, \\
&= N(0, I).
\end{aligned}
$$

By the mean-value theorem,

$$
h(\bar{X}_n) - h(\mu) = [\Delta h(\bar{X}_n^*)]^t (\bar{X}_n - \mu),
$$

where $|\bar{X}_n^* - \mu| \leq |\bar{X}_n - \mu|$. Obviously, from Theorem A.1 and the assumption on h, $\Delta h(\bar{X}_n^*) \to \Delta h(\mu)$ with probability one. Also, if we let $\sigma_n = [(\Delta h(\mu))^t \Sigma_n (\Delta h(\mu))]^{1/2}$, then $\lim_{n\to\infty} \Sigma_n^{1/2}/\sigma_n$ exists because of the condition (v). Using Slutsky's Theorem and Theorem 15.6 of Bartle again,

$$
\begin{aligned}
\frac{1}{\sigma_n}(h(\bar{X}_n) - h(\mu)) &= \frac{1}{\sigma_n} (\Delta h(\bar{X}_n^*))^t \Sigma_n^{1/2} \cdot [\Sigma_n^{-1/2} (\bar{X}_n - \mu)] \\
&\longrightarrow [\lim_{n\to\infty} \frac{1}{\sigma_n} (\Delta h(\bar{X}_n^*))^t \Sigma_n^{1/2}]\ N(0, I),\ \text{as } n \to \infty, \\
&= (\lim_{n\to\infty} \Delta h(\bar{X}_n^*))^t\ \lim_{n\to\infty} (\Sigma_n^{1/2}/\sigma_n)\ N(0, I) \\
&= (\Delta h(\mu))^t\ \lim_{n\to\infty} (\Sigma_n^{1/2}/\sigma_n) N(0, I) \\
&= N(0,\ \lim_{n\to\infty} [(\Delta h(\mu))^t \Sigma_n (\Delta h(\mu))/\sigma_n^2]), \\
&= N(0, I). \ \square
\end{aligned}
$$

Corollary A.2 *If conditions (i)-(v) of Theorems A.1 and A.2 are satisfied,*

$$
\sqrt{\textstyle\sum_{i=1}^{m} \varepsilon_i^2 \sigma_i^2/n_i}\ (\hat{E}^s \varphi - E\varphi) \longrightarrow N(0, I)
$$

as $n \to \infty$, where $\hat{E}^s \varphi$ and σ_i^2 are given in (5) and (9), respectively.

Theorem A.3 *Suppose that*

(i) Λ *is a compact subset of* $\mathcal{R}^q$, *where* q *is the number of elements in* λ,

(ii) $w(\theta, \lambda)$ *is continuous in both* θ *and* λ,

(iii) there exists a measurable function $h(\theta)$ *such that*

$$\sup_{\lambda \in \Lambda} w(\theta, \lambda) \leq h(\theta) \quad \textit{and} \quad \int h(\theta) f(\theta) d\theta < \infty.$$

Then

$$\lim_{n \to \infty} \hat{CV}^2(w; \lambda_N(\mathbf{z}), \mathbf{z}, N) \;=\; \min_{\lambda} CV^2(w; \lambda). \tag{31}$$

where $\lambda_N(\mathbf{z})$ *minimizes* $\hat{CV}^2(w; \lambda, \mathbf{z}, N)$ *over* $\lambda \in \Lambda$.

Proof: From Corollary A.1 and assumptions (i)-(iii) above, for almost every sequence $\mathbf{z}$,

$$\hat{CV}^2(w; \lambda, \mathbf{z}, N) \;\stackrel{N \to \infty}{\longrightarrow}\; CV^2(w; \lambda), \tag{32}$$

uniformly in λ (refer to Jenrich 1969). The result can be proved in a manner similar to that of Theorem 1 of Shao (1989). □

References

[1] Bartle, R. (1976), *The Elements of Real Analysis*, John Wiley & Sons, New York.

[2] Bauens, W., and Richard, J.F. (1985), "A 1-1 Poly-t Random Variable Generator with Application to Monte Carlo Integration," *Journal of Econometrics*, **29**, 19–46.

[3] Bickel, P., and Doksum, K. (1977), *Mathematical Statistics: Basic Ideas and Selected Topics*, Holden Day Inc.

[4] Cochran, W.G. (1963), *Sampling Techniques* (2nd ed.), John Wiley & Sons, New York.

[5] Dreze, J.H., and Richard, J.F. (1983), "Bayesian Analysis of Simultaneous Equation Systems," *Handbook of Econometrics*, Griliches, Z. and Intriligator, M., Editors, North-Holland, Amsterdam.

[6] Evans, M., and Gilula, Z., and Guttman, I. (1989), "Latent Class Analysis of Two-way Contingency Tables by Bayesian Methods," *Biometrika*, **76**, 557-563.

[7] Evans, M. (1991a), "Chaining via Annealing," *Annals of Statistics*, **19**, 382-393.

[8] Evans, M. (1991b), "Adaptive Importance Sampling and Chaining," *American Mathematical Society Contemporary Mathematics*, Flournoy, N. and Tsutakawa, R., editors, to appear.

[9] Everitt, B.S., and Hand, D.J. (1981), *Finite Mixture Distributions*, Chapman and Hall, London.

[10] Gelfand A., and Smith A.F.M. (1990a), "Sampling Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, **85**, 398-409.

[11] Gelfand A., and Smith A.F.M. (1990b), "Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling," *Journal of the American Statistical Association*, **85**, 972-985.

[12] Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions and Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721-741.

[13] Geweke, J. (1988), "Antithetic Acceleration of Monte Carlo Integration in Bayesian Inference," *Journal of Econometrics*, **38**, 73–89.

[14] Geweke, J. (1989), "Bayesian Inference in Econometric Models Using Monte Carlo Integration," *Econometrica*, **46**, 1–20.

[15] Jenrich, R.I. (1969), "Asymptotic Properties of Non-linear Least Squares Estimation," *Annals of Mathematical Statistics*, **40**, 633–643.

[16] Kloek, K., and van Dijk, H.K. (1978), "Bayesian Estimates of Equation System Parameters: An Application of Integration by Monte Carlo," *Econometrica*, **46**, 1–20.

[17] Naylor, J.C., and Smith, A.F.M. (1982), "Application of a Method for the Efficient Computation of Posterior Distributions," *Applied Statistics*, **31**, 214–225.

[18] Naylor, J.C., and Smith, A.F.M. (1988), "Econometric Illustrations of Novel Numerical Integration Strategies for Bayesian Inference," *Journal of Econometrics*, **38**, 103-125.

[19] Oh, M.S., and Berger, J.O. (1989), "Adaptive Importance Sampling in Monte Carlo Integration," *Tech. Report #89-19c*, Department of Statistics, Purdue University, to appear *Journal of Statistical Computing and Simulation.*

[20] Oh, M.S. (1991), "Statistical Multiple Integration via Monte Carlo Importance Sampling: Dimensionality Effect and an Adaptive Algorithm," *American Mathematical Society Contemporary Mathematics*, Flournoy, N. and Tsutakawa, R., editors, to appear.

[21] Rubinstein, R.Y. (1981), *Simulation and the Monte Carlo Method*, Wiley, New York.

[22] Shao, J. (1989), "Monte Carlo Approximation in Bayesian Decision Theory," *Journal of the American Statistical Association*, **84**, 727-732.

[23] Smith, A.F.M., Sken, A.M., Shaw, J.E.H., Naylor, J.C., and Dransfield, M. (1985), "The Implementation of the Bayesian Paradigm," *Communications in Statistics - Theory and Method*, **14**, 1079–1102.

[24] Stewart, L. (1979), "Multiparameter Univariate Bayesian Analysis," *Journal of the American Statistical Association*, **76**, 684–693.

[25] Stewart, L. (1983), "Bayesian Analysis Using Monte Carlo Integration – A Powerful Methodology for Handling Some Difficult Problems," *The Statistician*, **32**,. 195–200.

[26] Swain, J.J., and Schmeiser, B.W. (1988), "Control Variates for Monte Carlo Analysis of Nonlinear Statistical Models II; Raw Moments and Variances," *Journal of Statistical Computing and Simulation*, **30**, 39-56.

[27] Swain, J.J., and Schmeiser, B.W. (1989), "Control Variates for Monte Carlo Analysis of Nonlinear Statistical Models I: Overview," *Communications in Statistics*, **18**, 1011-1036.

[28] Tanner M., and Wong W. (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, **82**, 528-550.

[29] Tew, J.D., and Wilson, J.R. (1988), "Estimating Simulation Metamodels Using Integrated Variance Reduction Techniques," *American Statistical Association Proceedings of Statistical Computing Section*, 23 - 32.

[30] Tierney, L., and Kadane, J.B. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, **81**, 82–86.

[31] Tierney, L., Kass, R.E. and Kadane, J.B. (1989), "Fully Exponential Laplace Approximation to Expectations and Variances of Nonpositive Functions," *Journal of the American Statistical Association*, **84**, 710–716.

[32] Titterlington, D.M., Smith, A.F.M., and Makov, U.E. (1985), *Statistical Analysis of Finite Mixture Distributions*, Wiley, New York.

[33] van Dijk, H.J., and Kloek, T. (1980), "Further Experience in Bayesian Analysis Using Monte Carlo Integration," *Journal of Econometrics*, **14**, 307–328.

[34] van Dijk, H.K., and Kloek, T. (1983), "Experiments with Some Alternatives for Simple Importance Sampling in Monte Carlo Integration," *Bayesian Statistics 2*, North-Holland, Amsterdam, 511–530.

[35] van Dijk, H.K., and Kloek, T., and Boender, C.G.E. (1985), "Posterior Moments Computed by Mixed Integration," *Journal of Econometrics*, **29**, 3–18.

Table 1: Posterior Mean of θ

	mixture with matching process		mixture without matching process	unimodal with adaptive scheme
n	10000		10000	10000
time(sec)	35		35	31
$\hat{CV}^2(w;\lambda,\mathbf{z},N)$	1.129×10^{-4}		3.295×10^{-3}	2.88×10^{-4}
	w/o control var.	w/ control var.		
$\hat{E}\theta$ $(\hat{SD}(\hat{E}\theta_i))$	-0.4588 (0.0068)	-0.4632 (0.0059)	-0.4476 (0.0347)	-0.4428 (0.0081)
	-0.4600 (0.0073)	-0.4639 (0.0061)	-0.4192 (0.0257)	-0.4514 (0.0088)
	-0.4687 (0.0068)	-0.4703 (0.0058)	-0.4159 (0.0292)	-0.4404 (0.0085)
	-0.4576 (0.0067)	-0.4692 (0.0057)	-0.4404 (0.0272)	-0.4471 (0.0078)
	-0.4673 (0.0078)	-0.4592 (0.0064)	-0.4562 (0.0327)	-0.4659 (0.0097)
	-0.6935 (0.0082)	-0.6955 (0.0071)	-0.6459 (0.0277)	-0.6478 (0.0110)
	-0.6972 (0.0077)	-0.6943 (0.0074)	-0.7196 (0.0249)	-0.6567 (0.0104)
	-0.7027 (0.0084)	-0.6999 (0.0072)	-0.6709 (0.0268)	-0.6746 (0.0106)
	-0.6949 (0.0078)	-0.6906 (0.0069)	-0.6793 (0.0224)	-0.6634 (0.0112)
	-0.7054 (0.0084)	-0.7015 (0.0072)	-0.6997 (0.0259)	-0.6741 (0.0113)

Fig 1: Marginal Distribution Function of theta_1

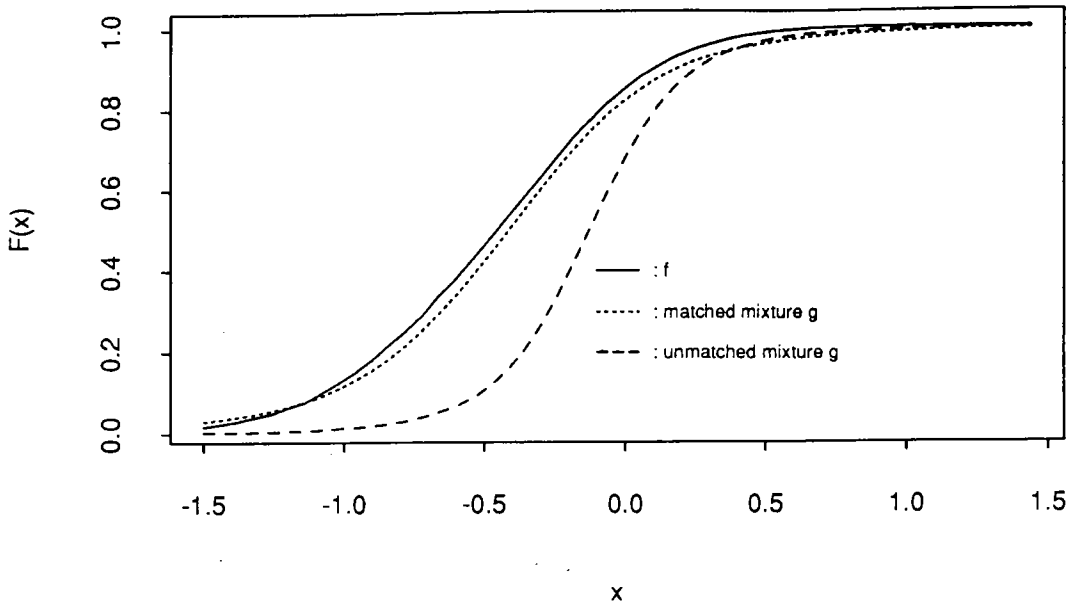

Fig 2: Marginal Distribution Function of theta_10