Computing Normalizing Constants for Finite Mixture Models via Incremental Mixture Importance Sampling (IMIS)

Russell J. STEELE, Adrian E. RAFTERY, and Mary J. EMOND

This article proposes a method for approximating integrated likelihoods in finite mixture models. We formulate the model in terms of the unobserved group memberships, z, and make them the variables of integration. The integral is then evaluated using importance sampling over the z. We propose an adaptive importance sampling function which is itself a mixture, with two types of component distributions, one concentrated and one diffuse. The more concentrated type of component serves the usual purpose of an importance sampling function, sampling mostly group assignments of high posterior probability. The less concentrated type of component allows for the importance sampling function to explore the space in a controlled way to find other, unvisited assignments with high posterior probability. Components are added adaptively, one at a time, to cover areas of high posterior probability not well covered by the current importance sampling function. The method is called *incremental mixture importance sampling* (IMIS).

IMIS is easy to implement and to monitor for convergence. It scales easily for higher dimensional mixture distributions when a conjugate prior is specified for the mixture parameters. The simulated values on which the estimate is based are independent, which allows for straightforward estimation of standard errors. The self-monitoring aspects of the method make it easier to adjust tuning parameters in the course of estimation than standard Markov chain Monte Carlo algorithms. With only small modifications to the code, one can use the method for a wide variety of mixture distributions of different dimensions. The method performed well in simulations and in mixture problems in astronomy and medical research.

Key Words: Bayes factor; Bayesian model averaging; Dirichlet-multinomial; Defensive mixture importance sampling; Gibbs sampling; Label-switching; Markov chain Monte Carlo; Multimodality.

Russell J. Steele is Assistant Professor of Mathematics and Statistics, McGill University, 805 Sherbrooke O., Montreal, PQ, Canada H3A 2K6 (E-mail: *steele@math.mcgill.ca*; Web: *www.math.mcgill.ca/˜steele*). Adrian E. Raftery is Professor of Statistics and Sociology, University of Washington, Box 354322, Seattle, WA 98195-4322 (E-mail: *raftery@stat.washington.edu*; Web: *www.stat.washington.edu/raftery*). Mary J. Emond is Research Assistant Professor of Biostatistics, University of Washington, Box 357237, Seattle, WA 98195-7237 (E-mail: *emond@u.washington.edu*).

Journal of Computational and Graphical Statistics, Volume 15, Number 3, Pages 712–734
DOI: 10.1198/106186006X132358

1. INTRODUCTION

The integrated likelihood plays an essential role in Bayesian inference and testing as it is the central component of the Bayes factor for comparing two models. It also plays a role in Bayesian estimation, as the normalizing constant for the posterior distribution. The integrated likelihood of a model is

$$I \equiv \text{pr}(\mathbf{y}) = \int f(\mathbf{y}|\tau)p(\tau)d\tau, \tag{1.1}$$

where $\mathbf{y}$ denotes the observed data, $f(\mathbf{y}|\tau)$ is the likelihood function for the parameter τ under the model, and $p(\tau)$ is the density (or probability mass function) for the prior distribution of τ given the model.

Because the integrated likelihood often is not analytically tractable, a body of literature on the use of numerical methods for its calculation has developed. Evans and Swartz (1995) and Chen, Shao, and Ibrahim (2000) included methods based on quadrature rules, Laplace's method, importance sampling, and Markov chain Monte Carlo (MCMC). Combinations of MCMC with importance sampling and the Laplace method were considered by Rozenkranz and Raftery (1994), Raftery (1996b), and Lewis and Raftery (1997). The Bayesian information criterion (BIC) can be used as the basis for an asymptotic approximation to the Bayes factor (Schwarz 1978; Kass and Wasserman 1995; Raftery 1995).

For finite mixture models, however, none of these methods is fully satisfactory. Two features of mixture models make many current methods for approximating the integrated likelihood problematic. The first is that the model is not "regular" for testing and model-selection purposes. In regular models, the log-likelihood becomes approximately elliptically contoured when there are enough data, even when the true parameter values correspond to a lower-dimensional submodel that one is trying to test. In this standard situation, for example, the likelihood-ratio test statistic has an approximate asymptotic chi-squared distribution with degrees of freedom equal to the difference in the number of parameters. This does not hold in finite mixture models whenever one estimates a model with G components but the true number of components is smaller, so that the true parameter values lie on the edge of the parameter space (Lindsay 1995). A second feature is the "label-switching" problem, namely that the likelihood is invariant to relabeling of the mixture components, and so has $G!$ modes of the same height. Additional local modes are often present (Lindsay 1995; Titterington, Smith, and Makov 1985; Atwood, Wilson, Elston, and Bailey-Wilson 1992).

The Laplace method (e.g., Tierney and Kadane 1986) provides an analytic approximation to the integrated likelihood based on the assumption that the posterior distribution is approximately elliptically contoured (e.g., Raftery 1996a), and when this assumption holds it can provide approximations of remarkable quality (e.g., Tierney and Kadane 1986; Grunwald, Guttorp, and Raftery 1993; Lewis and Raftery 1997). However, for mixture models this assumption fails when the model being fit has G components and the actual number of components is smaller (Lindsay 1995), which is a situation of great interest for model comparison and testing. Thus the Laplace method does not work in this situation.

Markov chain Monte Carlo (MCMC) can be used to estimate mixture models, and associated methods can be used to approximate integrated likelihoods (e.g., Chib 1995;

Raftery 1996b). However, in addition to the usual problems with MCMC methods (dependent samples, convergence issues, complexity of programming and implementation), in mixture models they can easily fall foul of the label-switching problem (Celeux 1997; Celeux, Hurn, and Robert 2000; Stephens 1997, 2000b). For example, Neal (1998) pointed out that Chib's (1995) results for a mixture model were in error for this reason. The problem could be solved correctly using the methods of Chib and Jeliazkov (2001). However, assessing the accuracy of the estimated integrated likelihoods is not trivial with MCMC because of the dependence between successive samples.

Our goal in this article is to propose adaptive importance sampling methods for integrated likelihoods in mixture models that are easy to implement and that avoid the difficulties we have been discussing. The success of any importance sampling method depends critically on the importance sampling function. Here we develop importance sampling functions for the component labels (rather than the parameters of the component densities) that are themselves mixtures and are specified adaptively. We propose two approaches to this. The first takes defensive mixture importance sampling (Hesterberg 1995; Raghavan and Cox 1998) as a starting point, and the second is based on sampling via perturbation of an initial grouping that has high posterior probability.

One key advantage of our approach is that the algorithm does not become more complicated as the complexity of the underlying mixture densities increases. Implementation for higher dimensional mixtures is similar to that for one-dimensional mixtures, reducing the coding burden normally associated with adapting Markov chain Monte Carlo methods to various problems. It seems to provide quite good approximations to the integrated likelihood, reliable estimates of the associated standard error, and is easily monitored for convergence of the estimate.

Section 2 reviews mixture models, presents our importance sampling based estimators, and shows how they are applied in the context of multimodality due to the label-switching problem. Section 3 contains simulation results motivated by problems in astronomy and medical research. Section 4 discusses advantages, limitations, other methods and directions for future research.

2. INTEGRATED LIKELIHOODS FOR MIXTURE MODELS VIA IMPORTANCE SAMPLING

2.1 Integrated Likelihoods for Finite Mixture Models

Let $y = (y_1, \ldots, y_n)$ be a realization from a G-component mixture distribution. The corresponding likelihood is

$$\prod_{i=1}^{n} \sum_{j=1}^{G} \pi_j f_j(y_i|\theta_j) \equiv \prod_{i=1}^{n} p(y_i|\theta, \pi), \tag{2.1}$$

where the π_j's are mixing proportions that sum to 1, $\pi = (\pi_1, \ldots, \pi_G)$, and the θ_j's are component-specific parameter vectors with $\theta = (\theta_1', \ldots, \theta_G')'$. Each observation, y_i, arises from one of the G component densities, $f_j, j = 1, \ldots, G$, but the group memberships are unknown. The parameter π_j is the unknown probability of an observation arising from f_j.

To obtain the integrated likelihood, or marginal probability of the data, the joint distribution of y and $\tau = (\theta, \pi)$ is integrated with respect to the unknown parameters:

$$I \equiv \int_\tau \prod_{i=1}^n p(y_i|\theta, \pi) p(\theta, \pi) d\tau, \tag{2.2}$$

where $p(\theta, \pi)$ is the prior density for (θ, π). Analytic integration of (2.2) is usually not feasible.

The component membership for y_i may be thought of as an unobserved random variable. When the component membership is known, the likelihood takes a simpler form. Let $z_i \equiv (z_{i1}, \ldots, z_{iG})'$ be the vector that indicates component membership for the ith observation such that $z_{ij} = 1$ if y_i is from component j and 0 otherwise. The $n \times G$ matrix $z \equiv \{z_1, \ldots, z_n\}'$ gives the component membership for the entire sample. Then

$$\begin{aligned} I &= \int_\tau \prod_{i=1}^n \sum_{z_i} p(y_i|z_i, \theta, \pi) p(z_i|\theta, \pi) p(\theta, \pi) d\tau & (2.3) \\ &= \sum_z \int_\tau \prod_{i=1}^n \prod_{j=1}^G (\pi_j f_j(y_i|\theta))^{z_{ij}} p(\theta, \pi) d\tau. & (2.4) \end{aligned}$$

In (2.3), the summation is over the G possible values of z_i, and in (2.4) the summation is over all G^n possible values of z. Note that we have assumed that $p(z_i|\theta, \pi) = p(z_i|\pi) = \prod_{j=1}^G \pi_j^{z_{ij}}$. In the rest of this article, we also assume that θ and π are independent a priori, so that $p(\theta, \pi) = p(\theta)p(\pi)$ (where it is understood that the $p(\cdot)$'s refer to different functions, depending on the argument).

This formulation simplifies part of the problem because the inner integral of (2.4) (an integration with respect to $\tau = (\theta, \pi)$) often can be evaluated analytically, or at least closely approximated via the Laplace method or a similar approach, and may also be more amenable to numerical integration via quadrature. The problem then takes the following general form

$$I = \sum_z p(y|z) p(z). \tag{2.5}$$

Here, $p(y|z) = \int_\theta \prod_{i=1}^n p(y_i|z_i, \theta) p(\theta) d\theta$, and $p(z) = \int_\pi \prod_{i=1}^n p(z_i|\pi) p(\pi) d\pi$. For the purposes of this article, we will assume that integration with respect to (θ, π) can be done analytically. Desai (2000) and Desai and Emond (2004) treated cases where numerical methods are needed for integration with respect to (θ, π).

2.2 DEFENSIVE MIXTURE IMPORTANCE SAMPLING

Because of the large number of possible allocations of observations to components, sampling based methods are required to calculate the integral (2.5). A simple Monte Carlo approach to integration would sample z from its marginal distribution, $p(z)$. A Dirichlet prior on π induces a Dirichlet-multinomial prior distribution on z. If a group assignment

z is sampled from that induced prior distribution, then one could calculate an empirical average

$$\hat{I}_{\text{MC}} = \frac{1}{T}\sum_{t=1}^{T} p(y|z^t). \tag{2.6}$$

where the $z^t, t = 1 \ldots T$ are sampled from $p(z)$. However, the prior distribution on the component labels will be far too diffuse for many problems because many group assignments will have low posterior probability, yielding unstable $\hat{I}_{\text{MC}}$ estimates.

Hammersley and Handscomb (1964) first suggested sampling from an "importance sampling distribution," $g(z)$, that samples more often from "important" parts of the space of integration, yielding the importance sampling estimate

$$\hat{I}_{\text{IS}} = \frac{1}{T}\sum_{t=1}^{T} p(y|z^t)w(z^t) = \frac{1}{T}\sum_{t=1}^{T} p(y|z^t)\frac{p(z^t)}{g(z^t)}. \tag{2.7}$$

There is an optimal $g(z)$ from which to sample. If one could sample from $p(z|y)$ and had access to its analytic form, then

$$\hat{I}_{\text{IS}} = \frac{1}{T}\sum_{t=1}^{T} p(y|z^t)\frac{p(z^t)}{p(z^t|y)} = p(y), \tag{2.8}$$

is a zero-variance estimator of I. However, knowledge of $p(z|y)$ requires the unknown $p(y)$, and so it is not available. Still, this gives hope that one could find an importance sampling function close to the optimal $p(z|y)$ that would give estimates with lower variance than the Monte Carlo estimator.

Wei and Tanner (1990) suggested that $p(z|\hat{\tau}, y)$ would make a good substitute for $p(z|y)$ in an imputation context (where, in their case, z represented missing observations and $\hat{\tau}$ represented the value of τ maximizing $p(y|\tau)$). However, in the mixture problem, the likelihood is typically multimodal and this importance sampling function is often concentrated around sets of labelings corresponding to just one likelihood mode, and so may not be a good approximation to $p(z|y)$. An overly concentrated importance sampling function causes difficulties because it may *increase* the variance of estimates of I, due to the high variability in the weights. For instance, if $g(z)$ is small for a z that gives a large value of $p(z)$ and $p(y|z)$, then the importance sampling estimate may have a very large variance.

Hesterberg (1995) suggested a simple fix for this particular drawback of importance sampling. Although mixtures of importance sampling functions had been proposed in the past (Oh and Berger 1993; West 1993; Givens and Raftery 1996), Hesterberg (1995) was the first to suggest using the Monte Carlo sampling function, $p(z)$, as a component of the mixture importance sampling function $\delta p(z) + (1-\delta)g(z)$, giving the following importance sampling estimator

$$\hat{I}_{\text{DM}} = \frac{1}{T}\sum_{t=1}^{T} p(y|z^t)\frac{p(z^t)}{\delta p(z) + (1-\delta)g(z)} = \frac{1}{T}\sum_{t=1}^{T} p(y|z^t)w^*(z^t), \tag{2.9}$$

where $g(z)$ is the usual sampling function that covers important parts of the space as before. One of the appealing advantages of using this defensive mixture importance sampling function is that the importance sampling weights $w^*(z)$ are bounded by $1/\delta$. The choice $\delta = 1$ results in the $\hat{I}_{\mathrm{MC}}$ estimator, while $\delta = 0$ gives the $\hat{I}_{\mathrm{IS}}$ estimator with importance sampling function $g(z)$. Hesterberg notes that a K-component mixture could also be used

$$h(z) = \sum_{k=1}^{K-1} \delta_k g_k(z) + \delta_K p(z), \;\; \sum_{k=1}^{K} \delta_k = 1, \tag{2.10}$$

which would allow one to sample from different parts of the space.

2.3 CHOICES OF $g(z)$

One promising choice of $g(z)$ is Wei and Tanner's (1990) proposal, $p(z|\hat{\tau}, y)$. Sampling from $p(z|\hat{\tau}, y)$ is simple to do, as one need only sample each component label from a multinomial distribution with probabilities equal to the conditional probabilities of group membership for each observation. An advantage of sampling on the component labels is that the sampling does not depend on the dimensionality of the data or the underlying parameter space. Multinomial sampling is fast and computationally inexpensive.

Still, using $p(z|\hat{\tau}, y)$ has drawbacks. One is that it uses only one of the $G!$ possible specifications of $\hat{\tau}$. The likelihood surface has $G!$ modes corresponding to the $G!$ different component labelings. For a likelihood symmetric prior distribution on the parameters (the most common choice in the literature), using a particular $\hat{\tau}$ for $p(z|\hat{\tau}, y)$ would result in either an underestimation of $p(y)$ (because certain parts of the space would rarely be visited) or estimates of $p(y)$ with high empirical variance (because of the larger importance sampling weights associated with sampling rare component labelings under a particular $\hat{\tau}$). Some authors have addressed difficulties with the multiple likelihood modes due to label-switching by specifying ordering contraints on the parameters by, for example, constraining $\theta_1 > \theta_2$ for a two-component mixture (Richardson and Green 1997). There are two drawbacks to this sort of prior specification. First, ordering components can become complicated as the dimensionality of θ_i and the number of groups both increase. Second, other researchers have found that ordering the components can cause computational and inferential difficulties (see, e.g., Celeux et al. 2000 and Stephens 2000).

Another difficulty with sampling from $p(z|\hat{\tau}, y)$ is that $p(z|\hat{\tau}, y)$ often contains many values close to 1, which does not allow the importance sampling function to explore much of the space of component labels. In order to overcome the difficulties associated with sampling from $p(z|\hat{\tau}, y)$, we suggest two other importance sampling functions based on $p(z|\hat{\tau}, y)$ in the following subsections.

2.3.1 A Label-Switching Dependent Product of Multinomials

First, we introduce the following label-switching version of $p(z|\hat{\tau}, y)$. In most problems, there will be observations that have values of $p(z_i|\hat{\tau}, y)$ very close to 1. In fact, many of these values will essentially *be* 1 to within rounding error. We will use these points as representative points to set a particular labeling of the components, and then sample the rest of the observations according to $p(z|\hat{\tau}_s, y)$, where $\hat{\tau}_s$ has the maximum likelihood estimates labeled according to the labeling of the representative points.

We now more formally describe the algorithm. Let $\hat{z}$ be the $n \times G$ matrix of $p(z_{ij}|\hat{\tau}_1, y)$ resulting from the EM algorithm, where each row sums to 1. Assume the observations are ordered such that $\max_j \hat{z}_{ij} > \max_j \hat{z}_{(i+1)j}$, for all i. Now assign the observations to components in the following way. For the first observation, let

$$\Pr(z_{1j} = 1) = \frac{1}{G}, \quad j = 1, \ldots, G.$$

In other words, observation 1 will be assigned to the components uniformly. Next, assign observation 2 to a group according to the following:

$$\Pr(z_{2j} = 1) = \begin{cases} \dfrac{1 - \hat{z}_{2l_1}}{G-1} & \text{for} \quad j \neq k_1 \\ \hat{z}_{2l_1} & \text{for} \quad j = k_1, \end{cases}$$

where k_1 is the group to which observation 1 was assigned and $l_1 = \text{argmax}_j \hat{z}_{1j}$, that is, the $\hat{z}$ matrix column for which observation 1 has the highest conditional probability. Observation 2 has high probability of being assigned to the same group as observation 1 if they have high conditional probability for the same group label according to $\hat{z}$, otherwise observation 2 will be assigned uniformly to the remaining groups.

Now, if observation 2 is assigned to a different group than observation 1, assign observation 3 in the following way:

If $\text{argmax}_j \hat{z}_{2j} \neq k_1$:

$$\Pr(z_{3j} = 1) = \begin{cases} \hat{z}_{3l_1} & \text{for} \quad j = k_1 \\ \dfrac{1 - \hat{z}_{3l_1} - \hat{z}_{3l_2}}{G-2} & \text{for} \quad j \neq k_1, k_2 \\ \hat{z}_{3l_2} & \text{for} \quad j = k_2. \end{cases}$$

If $\text{argmax}_j \hat{z}_{2j} = k_1$:

$$\Pr(z_{3j} = 1) = \begin{cases} \dfrac{1 - \hat{z}_{3l_1}}{G-1} & \text{for} \quad j \neq k_1 \\ \hat{z}_{3l_1} & \text{for} \quad j = k_1, \end{cases}$$

where $l_2 = \text{argmax}_j z_{2j}$ and k_2 is the group to which observation 2 was assigned.

If observation 2 is assigned to group k_1, then assign observation 3 according to the following:

$$\Pr(z_{3j} = 1) = \begin{cases} \dfrac{1 - \hat{z}_{3l_1}}{G-1} & \text{for} \quad j \neq k_1 \\ \hat{z}_{3l_1} & \text{for} \quad j = k_1. \end{cases}$$

Continue assigning observations in this way until $G-1$ representatives have been assigned, which then leads to assignment of observations according to a permuted version of the original $\hat{z}$ matrix.

The optimal situation in which to use such a sampling method would be when the $\hat{z}$ matrix has $G-1$ observations such that each observation has probability very close to 1, but all for different groups. This way, one would be sampling from permuted versions of the $\hat{z}$ matrix, without incurring the extra computational expense necessary to sample from a specific $\hat{z}$ and then randomly permute the labels (as the importance sampling function $g(z)$ in this case would have $G!$ summands to be calculated at each iteration). The joint probability of any simulated z value, z^*, is easy to calculate, since $\Pr(z = z^*) = \Pr(z_1 = z_1^*)\Pr(z_2 = z_2^*|z_1 = z_1^*)\ldots\Pr(z_n = z_n^*|z_{n-1} = z_{n-1}^*, \ldots, z_1 = z_1^*)$, where each probability is determined by the algorithm and requires only recording the probability used at the time of sampling. It may also be viewed as a sequential importance sampling function (Liu, Chen, and Wong 1998; MacEachern, Clyde, and Liu 1999).

This dependent sampling method addresses the multimodality of the likelihood due to label-switching, but does not address a problem inherent in using a $\hat{z}$ matrix that contains many values very close to either zero or one. We propose a second potential candidate for importance sampling that addresses this drawback of using such a $\hat{z}$ matrix for multinomial sampling.

2.3.2 A Product of Dirichlet-Multinomials

One of the problems with sampling only from the prior on z, $p(z)$, is that many observations that "should" be in the same group will not be with high probability (or observations that "should not" be in the same group together will be with high probability). We propose using the $\hat{z}$ matrix to determine preliminary groupings of observations, and then applying the Dirichlet-Multinomial sampling function to each of these groups individually. By doing so, a weak dependency is built among observations which have high posterior probability of belonging to the same group.

As before, let $\hat{z}$ be the matrix of $p(z_{ij}|\hat{\tau}, y)$ for a specific permutation of the component labels. Create G groups by assigning observations, initially, to group l_i, where $l_i = \text{argmax}_j \hat{z}_{ij}$. Then, for each nonempty group $r, (r = 1, \ldots, G)$, sample η_r from a Dirichlet distribution with parameter vector $\alpha_r = (\alpha_{r1}, \ldots, \alpha_{rG})$. Now, for each group, reassign observations to groups according to their group-specific η_r.

One could envision many different values of the parameters α_r, but we found that taking $\alpha_{rj} = 1$ for all r, j works reasonably well. The reason for this is that the sampling distribution assigns observations symmetrically to groups (which obviates the difficulties due to label switching encountered with use of $p(z|\hat{\tau}, y)$ directly) and also gives a fair number of samples of z such that the G initial groupings remain roughly intact. The required probabilities $g_2(z)$ are

$$g_2(z) = \prod_{r=1}^{G} \frac{\Gamma(\sum_j \alpha_{rj})}{\Gamma(\sum_j n_{rj} + \alpha_{rj})} \frac{\prod_j \Gamma(n_{rj} + \alpha_{rj})}{\prod_j \Gamma(\alpha_{rj})}, \tag{2.11}$$

where n_{rj} is the number of observations from the rth group assigned to the jth group.

Our two proposed sampling distributions accomplish different goals. The label-switching dependent product of multinomials primarily samples label allocations that correspond to one specific likelihood mode of the parameters. For well-separated mixtures, one would expect these allocations to provide most of the mass in

$$I = \sum_z p(y|z)p(z).$$

However, due to the presence of local modes beyond those due to label-switching, sampling from the $\hat{z}$ matrix corresponding to just one mode will probably be inefficient. The product of Dirichlet-multinomials will sample other parts of the space more often, which helps to guard against large importance sampling weights.

2.4 INCREMENTAL MIXTURE IMPORTANCE SAMPLING (IMIS)

The importance sampling functions we have just described can miss areas of high posterior probability. To avoid this, we propose adaptively specifying the mixture importance sampling function by incrementally adding components to the mixture to capture parts of the space that have been missed. We use a mixture importance sampling function (as in Geyer 1991), based on several τ_j^*'s, where each $\tau_j^* = (\theta^*, \pi^*)$ corresponds to a local posterior mode in the parameter space. In the notation of Section 2.2, we will be adaptively constructing a function of the form

$$g_K(z) = \delta p(z) + \sum_{j=1}^{J} (\delta_{1j} g_{1j}(z) + \delta_{2j} g_{2j}(z)),$$

where J is the number of posterior modes, $g_{1j}(z)$ is a concentrated sampling functions centered at the jth mode, $g_{2j}(z)$ is a dispersed sampling function centered at the jth mode, and $p(z)$ is the prior distribution on z.

One could imagine many choices for δ and the δ_{ij}'s, but we have found that the sampling algorithm is not very sensitive to the precise choice of the δ_{ij}'s. We have found that one often needs only to bound δ away from 0 and 1, as suggested by Hesterberg (1995) and as discussed by Steele (2002). Therefore, in the examples that follow, we will use $\delta_{1j} = \delta_{2j} = 1/(2J)$ and $\delta = 0.5$.

Implementation requires a method for choosing the τ_j^*'s. Note that the posterior probability of any group assignment, $p(z|y)$, is proportional to $p(y, z)$. A missed area of high posterior probability would be indicated by a large value of the summand used in calculating $\hat{I}$, so that $p(y, z)/g(z)$ would be large. At each iteration of our incremental algorithm, a new importance sampling function is specified by adding a mixture component to the importance sampling function at the previous iteration. The component added corresponds to the z^t with the highest summand value at the previous iteration, $p(y, z^t)/g(z^t)$. We then add two additional components to the mixture corresponding to a new local mode, z^*, based on $p(z|\tau^*, y)$, where τ^* is the mode of $p(\tau|y, z^*)$.

It is impossible to know for certain when the IMIS algorithm has stabilized. The only way to guarantee convergence would be to include one component for each of the G^n allocations of observations to groups. It could always be the case that there is a pathological z that yields a very large value for the importance sampling weight and for $p(y|z)$. It should be noted that the problem of local modes is a problem that is shared by all importance sampling and Monte Carlo methods. However, as the number of components increases, the probability of missing isolated modes will decrease due to the sampling of labels from the prior distribution. We will assume that the IMIS algorithm has approached convergence when the variability of the $\hat{I}_k$ value and its associated coefficient of variation have stabilized.

If the addition of components to the distribution does not change the variance or the value of $\hat{I}$, then one can assume that the local modes corresponding to the later components have either already been covered by other components or do not contribute much to the estimate of I.

The add-one methodology retains most of the simplicity of the original proposed importance sampling method, but has the flexibility to sample from important parts of the space neglected by other importance sampling functions based on only one initial τ^*. We call the resulting algorithm *incremental mixture importance sampling* (IMIS).

3. EXAMPLES

We now present two applications of the adaptive importance sampling method. The first example is a common one-dimensional Gaussian mixture example showing that the method works well compared to a Markov chain Monte Carlo method. The second example is a three-dimensional Gaussian mixture example from the medical literature.

3.1 GALAXY DATA

The first example shows an attempt to approximate the integrated likelihood for a dataset involving velocities of galaxies (Postman, Huchra, and Geller 1986), discussed by Roeder (1990). The number of possible groups of galaxies is the question of interest for this dataset. Although Stephens (2000a) used a mixture of t-distributions, most other authors who have analyzed these data have used mixtures of Gaussians, and we will do likewise in order to facilitate comparisons with other methods.

Liang and Wong (2001) suggested a simulated annealing MCMC approach for calculating normalizing constants combined with bridge sampling (Meng and Wong 1996). Their method, called evolutionary Monte Carlo (EMC), requires running several (in their examples, 20) Markov chains, each of which samples from $f_i(\tau) = (p(y|\tau, G))^{u_i} p(\tau|G)$, where $u_i = 0, 0.05, \ldots, 1.0$ is different for each chain. Their primary contribution was to suggest an evolutionary Monte Carlo approach for swapping parts of the sample vector τ among chains. The annealed MCMC approach allowed the chain to visit the multiple modes of the likelihood. Neal (1998) pointed out that Chib's (1995) Gibbs sampling approach for the same galaxy dataset was inadequate because it did not visit all the modes of the mixture likelihood surface.

We compare IMIS to the method of Liang and Wong (2001), viewed as a state-of-the-art representative of MCMC approaches to the problem. Note that their method will become more difficult to implement as the dimensionality of the data (and therefore τ) increases.

For the galaxy data, we use a conjugate Normal-Inverse-χ^2 prior for the mean and variance parameters, namely

$$\begin{aligned} \mu_j|\sigma_j^2 &\sim N(20, \sigma_j^2) \\ \sigma_j^2 &\sim 100 \text{ Inverse-}\chi_6^2, \end{aligned}$$

Table 1. Comparison of Log-Integrated Likelihood Estimates for the Galaxy Data. To obtain actual log-integrated likelihood estimates, subtract 230 from each value in the table. The number in parentheses for all methods is the coefficient of variation for these 10 runs, namely the standard deviation over 10 runs of $\hat{I}$ for each method divided by the average $\hat{I}$ value for each method. Values for IMIS were obtained using a 51-component mixture with $T = 10{,}000$. Values for EMC short runs were obtained using 10 runs of the method for 125,000 iterations, thinned by 50 to obtain nearly independent values. Values for Monte Carlo estimates were obtained using 10 runs of one million iterations. Values for EMC long runs were obtained using 1 run of length 1.25 million, thinned by 50.

Clusters	*IMIS*	*EMC (short)*	*MC estimate using p(τ)*	*EMC (long)*
2	−2.96 (0.001)	−2.99 (0.08)	−2.89 (0.76)	−2.92
3	−2.14 (0.007)	−2.13 (0.07)	−2.11 (0.62)	−2.15
4	−2.36 (0.038)	−2.32 (0.11)	−2.39 (1.04)	−2.36
5	−2.81 (0.106)	−2.73 (0.07)	−2.82 (0.66)	−2.80
6	−3.28 (0.204)	−3.21 (0.12)	−2.29 (0.81)	−2.29
7	−3.76 (0.299)	−3.81 (0.10)	−2.64 (1.04)	−2.77

which is similar to the prior used by Chib (1995) and by Liang and Wong (2001), with the difference that they did not use a conjugate prior, but instead assumed prior independence of the mean and variance parameters. We used the standard uniform Dirichlet prior for the mixing parameters.

Table 1 shows the values of I estimated by IMIS, by Monte Carlo sampling from the prior on τ and, as a "gold" standard, by a much longer run of Liang and Wong's (2001) EMC method. The gold standard estimate consists of 25,000 sample points from a chain of length 1.25 million, taking every 50th value of τ. This was done in order to ensure high quality of the estimate and to avoid problems with dependence amongst values of τ. Still, because it consists of only one run, the estimate of I in Table 1 has no associated standard error. The short runs are an attempt to match the amount of computational time required for a reasonable run of the adaptive importance sampling method.

Table 1 shows that IMIS provides accurate estimates of $\log(\hat{I})$. The coefficients of variation are low and all runs produce approximate 95% confidence bands for $\hat{I}$ within 1.0 of the "gold" standard long EMC estimate. This is adequate for interpretation on the standard scale for interpreting Bayes factors (Jeffreys 1961; Kass and Raftery 1995), which views a Bayes factor of three or less as weak evidence or, in Jeffreys's words, "evidence not worth more than a bare mention." Sampling from the prior gives a reasonably good answer when averaged over the 10 trials, but the variability across those 10 trials is unacceptable, especially as the dimension of the problem increases.

Figure 1 shows how the estimates of $\hat{I}$ vary for a randomly selected run as components are added to the mixture importance sampling function. IMIS allows one to continue adding components until the estimates of $\hat{I}$ stabilize. Note that all the estimates are unbiased; the purpose of adding components is to improve precision. The plot shows that for $G = 2$ and $G = 3$, 10 to 15 components would be enough to obtain reasonably good estimates of I. For $G = 4$ or $G = 5$ components, 20 to 25 components might be needed, whereas for $G = 6$ and $G = 7$, it might be worth going beyond even the 51 components used here. Figure 2

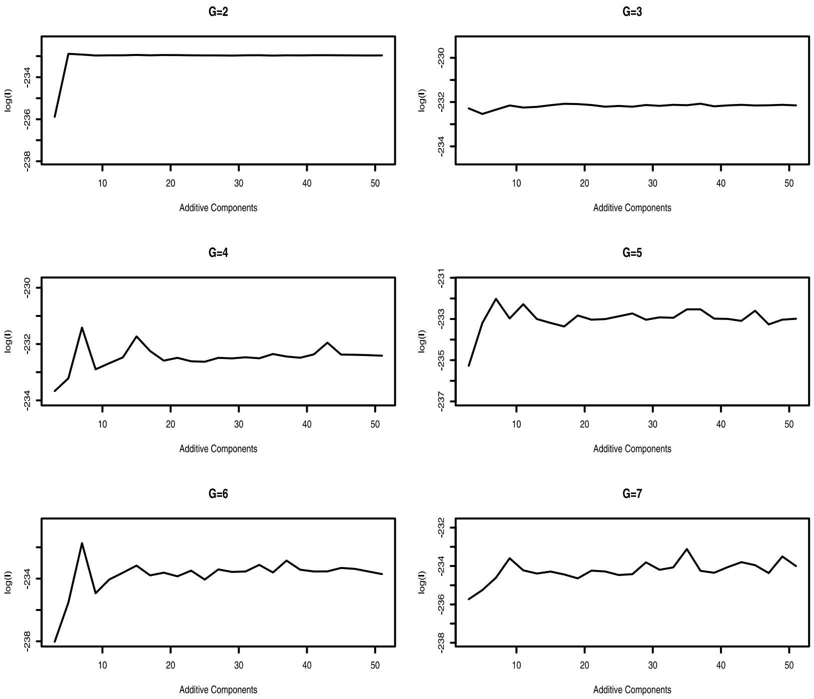

Figure 1. IMIS log$(\hat{I})$ trace plots for each number of components for the galaxy data. All runs used a maximum of 51 components based on 25 τ_j^*'s.

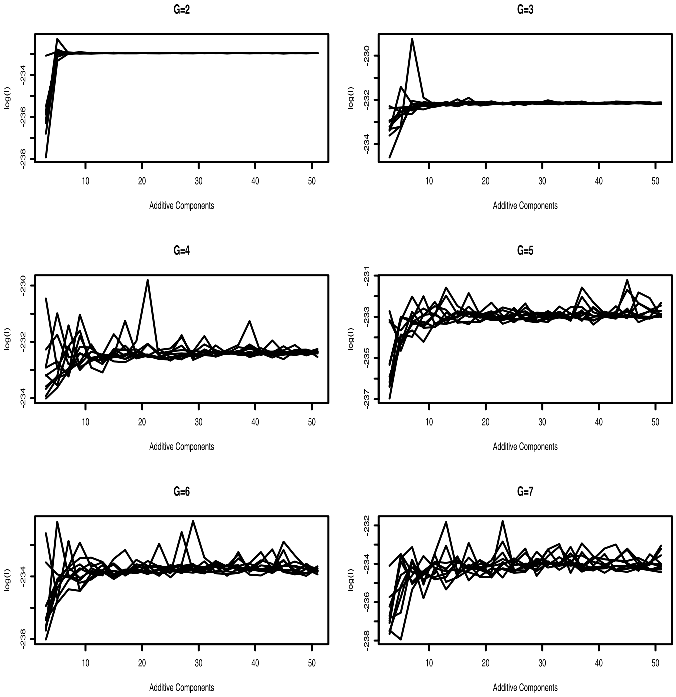

Figure 2. Trace plots for 10 runs of the IMIS method for the galaxy data. All runs used a maximum of 51 components based on 25 τ_j^*'s.

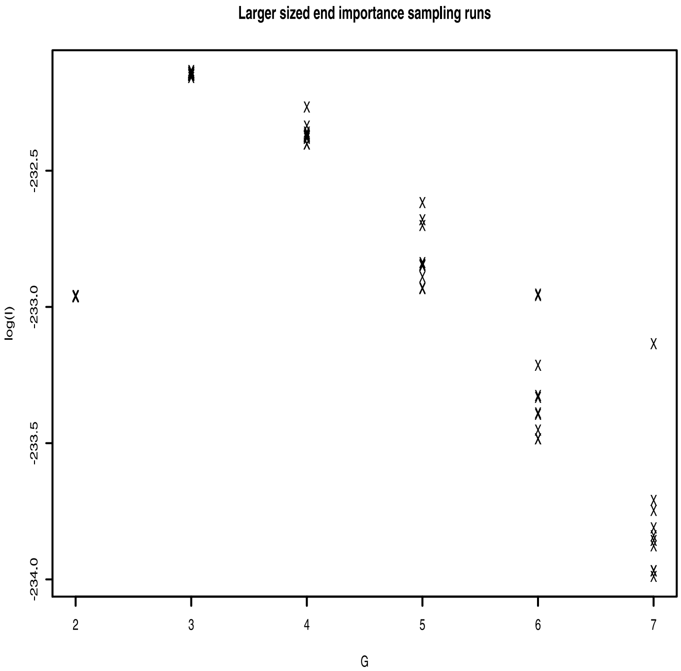

Figure 3. IMIS $\log(\hat{I})$ final estimates for the galaxy data for two to seven components. All points are based on 100,000 simulated values from a 51-component mixture importance sampling function.

shows trace plots for each of the 10 runs for $G = 2$ to 7 components, and Figure 3 shows the 10 IMIS final estimates using a larger number of simulations.

Table 2 shows the running times for each method. The table lists the time required for one run of the adaptive importance sampling method, a short run of the EMC method, and a single 1.25×10^5 run of the standard Monte Carlo sampling algorithm. The adaptive importance sampling method takes less time to run than a run of the EMC algorithm for each number of components. The Monte Carlo approach is faster for $G < 7$, but then actually takes longer than the adaptive importance sampling approach for $G = 7$. Of course, the high variability of Monte Carlo integration for this example makes it undesirable, and it is included here only for comparison.

3.2 DIABETES DATA

Next we consider a higher-dimensional example from the medical literature (Reaven and Miller 1979). The dataset consists of blood measures of insulin, glucose, and insulin resistance levels (SSPG) for 145 diabetes patients; the pairs plot of the data is shown in Figure 4. Fraley and Raftery (1998) analyzed the dataset using model-based clustering. Even with only three dimensions, the problem becomes hard to analyze using the evolutionary Monte Carlo method. Proposing good covariance matrices is hard to do and also expensive. The Monte Carlo and basic importance sampling methods fail because the dimensionality of the problem is too high, even for a model with $G = 2$ components, which has $6 + 12 + 1 = 19$

Table 2. CPU Times for the Galaxy Data. Table cells indicate seconds of CPU time required for running each method once to estimated the log-integrated likelihood. The methods were implemented as described in the note to Table 1, so that the various runs estimates with relatively equal precision.

# Clusters	*IMIS*	*EMC*	*MC estimate using p(τ)*
2	299	388	180
3	370	498	260
4	464	567	340
5	492	625	410
6	550	684	500
7	570	740	590

parameters. IMIS requires no special adjustments or tuning, because the sampling is done only on the component labels, and the only changes that need to be made to go from one dimension to more are the likelihood and prior functions.

Table 3 and Figure 5 shows the IMIS-estimated log-integrated likelihoods for G from 2

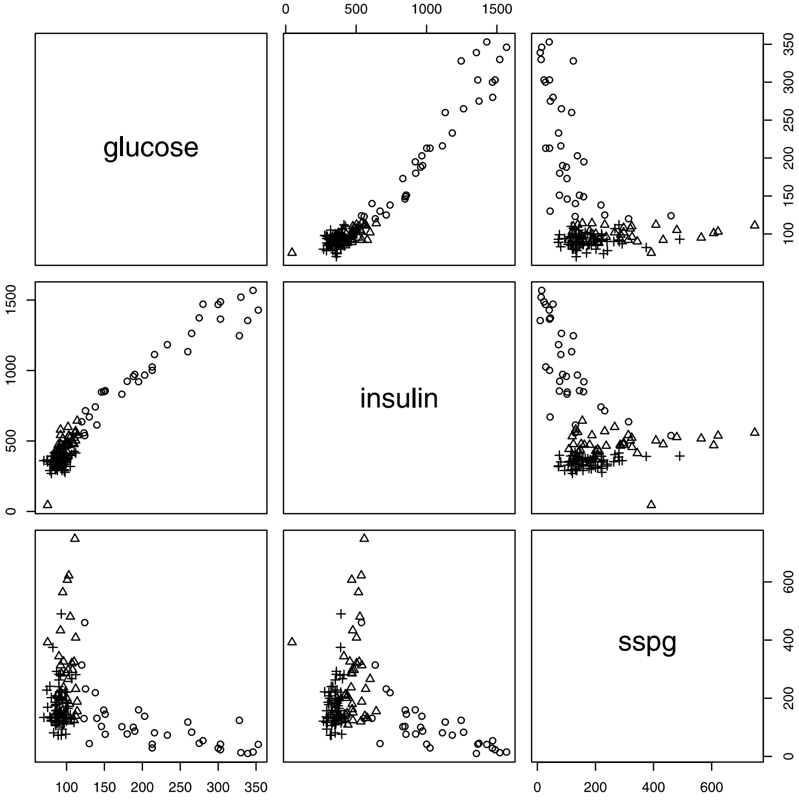

Figure 4. Pairwise plots of glucose, insulin, and SSPG for the diabetes dataset. Triangles $\triangle$ denote diagnosed chemical diabetes patients, crosses $+$ denote diagnosed normal patients, and circles $\bigcirc$ denote diagnosed overt diabetes patients.

Table 3. IMIS Log-Integrated Likelihood Estimates for the Diabetes Dataset. Each estimate is based on a single importance sampling run of 100,000 with 101 components, where the components were chosen using IMIS with smaller runs of 10,000 samples. To obtain actual log integrated llikelihood estimates, subtract 2,400 from each number in the table.

Clusters	*IMIS estimate*	*IMIS estimated CV*	*Bridge sampling estimates*
2	−49.04	0.020	(−48.55, −46.93, −44.51)
3	−15.09	0.051	(−27.58, −26.28, −25.66)
4	−14.09	0.082	(−27.36, −25.83, −24.41)
5	−14.68	0.100	(−23.74, −22.85, −21.17)

to 5 using values of K ranging from 51 (for $G = 2$) to 101 (for $G = 5$) adaptive components and $T = 100{,}000$. The four plots of Figure 5 show the IMIS-estimated values of $\log(I)$ (the solid line and left vertical axis) and the estimates of coefficients of variation (the dashed line and the right vertical axis) for the four different models under consideration as the number of fitted modes (and therefore the number of components in the IMIS sampling distribution) increases. We see results similar to those from the one-dimensional galaxy example, as the two- and three-component model estimates appear to be very stable, whereas the four and

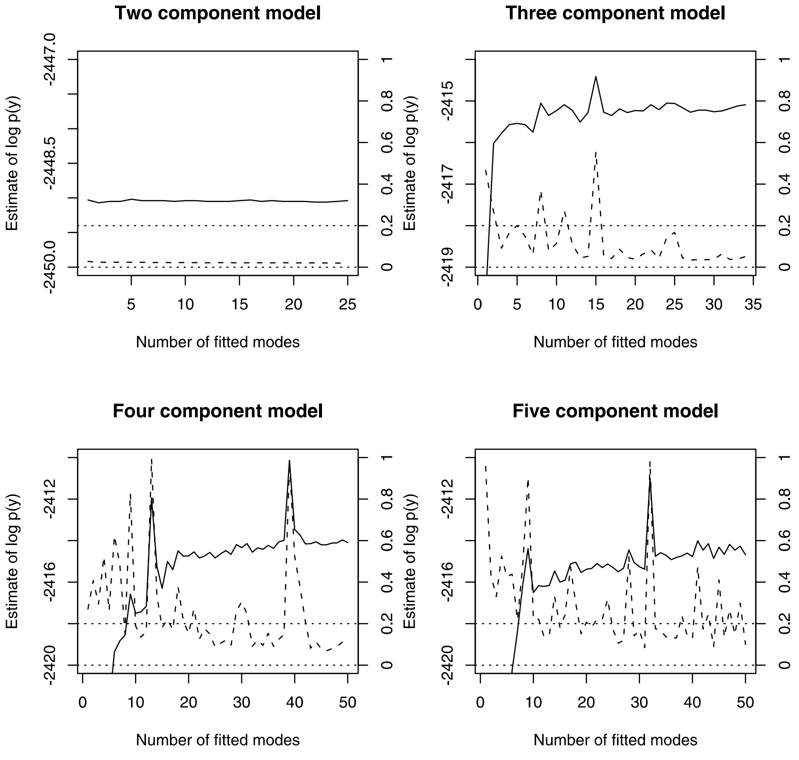

Figure 5. IMIS log$(\hat{I})$ (solid line, left axis labels) trace plots with estimated coefficients of variation (dashed line, right axis labels). The desired coefficient of variation range denoting relative stability is shown by dotted lines.

five component model model estimates are acceptable. Note that large variability in the $\log(\hat{I})$ estimates (the solid line) are usually accompanied by large values of the coefficient of variation (the dashed line). This is encouaraging, as it indicates that the IMIS algorithm provides reasonable estimates of the uncertainty of the $\log(\hat{I})$ values.

Table 3 also includes results for a parallel MCMC algorithm which does not include the EMC cross-over moves (which makes it essentially a parallel tempering algorithm with bridge sampling used to estimate the normalizing constant). The parallel tempering algorithm was significantly more difficult to tune and took much longer (approximately 10 hours of computation for the three-group problem as compared to hours for IMIS). The parallel tempering runs used Metropolis updates for the parameters that yielded reasonable acceptance rates across chains (although this was difficult to monitor). Because of difficulties with initializing the optimal bridge sampling estimator, we used the geometric bridge sampling function of Meng and Wong (1996) . The use of this bridge-sampling function also allowed us to trace the value of the $\hat{I}_{\mathrm{BS}}$ estimate over the course of the MCMC simulations.

Our table includes three separate values for the MCMC algorithm, one obtained for each of three possible lengths of "burn-in" for the 31 parallel chains of 500,000 iterations (thinned by taking every 10th value). The first value in the triad corresponds to an estimate with 5,000 iterations allocated to burn-in and 45,000 allocated to estimation of $\hat{I}$. The second value in the triad corresponds to 15,000 burn-in iterations and 35,000 for estimation and the third value corresponds to 40,000 burn-in and 10,000 for estimation. We can see a possible disadvantage of using the MCMC approach in that different allocations of iterations to burn-in and estimation give differing values of $\hat{I}$ with no estimate of the coefficient of variation to guide which value is "best."

Figures 6 and 7 shows the $\hat{I}_{\mathrm{BS}}$ paths for our three choices for the amount of burn-in for the two simplest models considered, $G = 2, 3$. The difficulty of monitoring and tuning the MCMC algorithm for this more complex problem suggests an advantage of the IMIS algorithm for higher dimensional problems. The large difference in estimated $\hat{I}$'s between our method and the bridge sampling method is alarming. We ran a much longer run for the three-group model (1 million iterations thinned to 100,000) and for certain burn-in values (up to 70,000) we got $\hat{I}$ values on the order of $-2{,}408$ using bridge sampling, which is well above that which is apparently well estimated by the IMIS algorithm. Because of the difficulty of tuning the bridge sampling algorithm for this small example, we excluded the method from the simulation study in the following section.

3.3 SIMULATION STUDY

We conducted a simulation study using trivariate normal densities similar to the one for the diabetes dataset. We simulated 100 random datasets and used the IMIS algorithm to estimate the integrated likelihood for 2, 3, 4, and 5 components. We restricted the algorithm to 50,000 iterations at each adaptive level and allowed for the addition of 25 total IMIS sampling functions to the final mixture estimation.

The IMIS algorithm proved to be extremely stable, as shown by Table 4. The coefficients

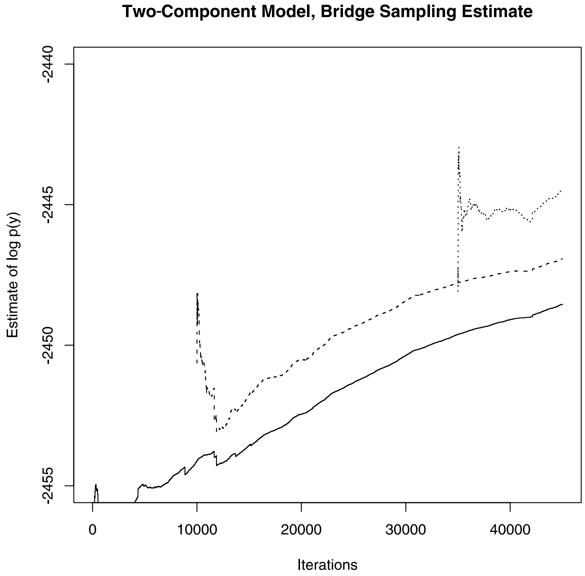

Figure 6. $log(\hat{I}_{BS})$ trace plots for the two-component model using the bridge sampling estimator. The solid line represents the estimate after 5,000 burn-in iterations, the dashed line represents the estimate after 15,000 burn-in iterations, and the dotted line represents the estimate after 35,000 burn-in iterations.

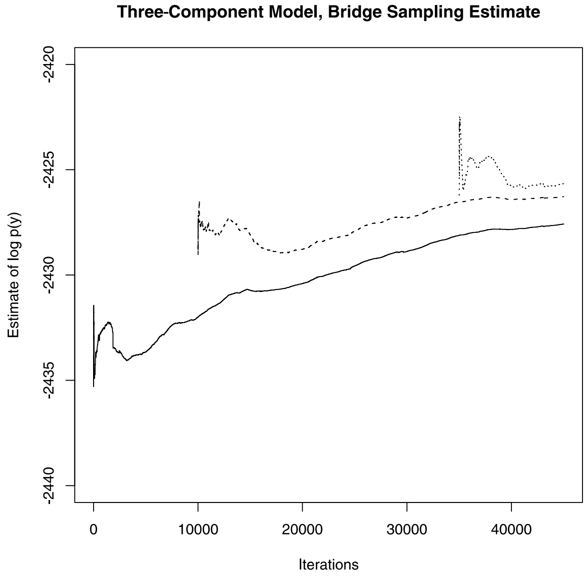

Figure 7. $log(\hat{I}_{BS})$ trace plots for three component model for the bridge sampling estimator. The solid line represents the estimate after 5,000 burn-in iterations, the dashed line represents the estimate after 15,000 burn-in iterations, and the dotted line represents the estimate after 35,000 burn-in iterations.

Table 4. Summary of Coefficients of Variation from IMIS Simulation Study Based on the Diabetes Dataset. Summary statistics for the coefficients of variation when using the IMIS algorithm to estimate *I* for 100 datasets under finite mixture models with two to five components. The results were obtained with 25,000 iterations per adaptive step and 25 final components in the adaptive mixture distribution.

	$G = 2$	$G = 3$	$G = 4$	$G = 5$
Minimum	0.008	0.012	0.040	0.066
1st Quartile	0.009	0.022	0.092	0.131
Median	0.009	0.031	0.14	0.198
Mean	0.009	0.042	0.188	0.272
3rd Quartile	0.009	0.046	0.217	0.344
Maximum	0.017	0.328	0.865	0.989

of variation were very low for the two- and three-component models. Even for the five-component model (a model with 52 parameters to integrate over), we see that the IMIS estimate of the integrated likelihood had a coefficient of variation less than 0.35 in three-quarters of the experiments.

Because calculation of normalizing constants for multivariate normal mixtures of this type is difficult for even sophisticated integration methods, we compare the results for the IMIS method to those obtained using an analytical model selection criterion, the BIC. Because the data were simulated from a normal mixture distribution, we should see a rough correspondence (although not an exact match) between the results for the IMIS integrated likelihoods and the BIC as calculated using the `mclust` software package.

Table 5 gives the comparison results for the 100 experiments. In 84% of the experiments, the IMIS estimate of the integrated likelihoods and the BIC values based on the maximum likelihood estimates both chose the correct model (the model with three mixture components). However, this does not take into consideration the IMIS estimates of the integrated likelihood that were in doubt because of high coefficients of variation. Table 5 also shows that the IMIS estimate of the integrated likelihoods and the BIC values agree on a three-component model for 53 of 60 (89%) of the experiments where the estimated coefficient of variation was less than 0.3.

We remark that the examples for which the BIC and IMIS estimate of the integrated likelihoods disagree do not indicate necessarily that the IMIS has misestimated the integrated likelihood, as the priors used in the experiments to calculate the integrated likelihoods are different from those that are approximated via the BIC.

4. DISCUSSION

We have proposed a general approach for calculating integrated likelihoods for finite mixture models via an adaptive mixture importance sampling, called IMIS. We used two types of sampling functions in the adaptive mixture sampling distribution, one concentrated on particular posterior modes and one that was relatively spread out over the parameter space. The resulting algorithm is relatively easy to implement compared to competing MCMC

Table 5. Concordance of Model Selection via IMIS $\hat{I}$ Estimates With Model Selection via the BIC. Two-way table describing the concordance of results for all 100 experiments. The true model was a three-component trivariate Gaussian mixture.

	Best model via $\hat{I}$ from IMIS									
	All experiments					*Only experiments with CV $<$ 0.3*				
Best model via Mclust and BIC	*2*	*3*	*4*	*5*	*Total*	*2*	*3*	*4*	*5*	*Total*
G = 2	0	2	0	0	2	0	1	0	0	1
G = 3	7	84	4	2	97	3	53	2	0	58
G = 4	0	1	0	0	1	0	1	0	0	1
Total	7	87	4	2	100	3	55	2	0	60

methods and runs more quickly than a standard MCMC implementation. Because each iteration is independent, one can obtain reasonable standard error estimates for the estimated integrated likelihood values, especially as the number of adaptive components increases. Monitoring $\hat{I}$ as the algorithm adds components also gives the user the opportunity to adjust the settings of the algorithm in an attempt to improve performance (e.g., by adjusting δ_K, K, and T). One can also stop and restart the method at any point by outputting the stored $\hat{z}$ matrices (a similar advantage enjoyed by the MCMC methods).

The bridge sampling and evolutionary Monte Carlo methods employed in this article could be improved upon. In order to make a fair comparison of our general IMIS method with an MCMC-based method, we used simple Metropolis proposal densities. Other proposal approaches and tuning parameters could have been used that may have increased the efficiency of the bridge sampling estimator (see Gilks and Wild 1992 or Neal 2003 for examples). However, the fact that other more sophisticated (and computationally expensive) adjustments must be made in order to get better estimates from MCMC-based methods only makes the nearly automatic IMIS algorithm more attractive as a general method. Similarly, the methods of Gelman and Meng (1998) could provide better estimates, but only at the cost of greater computational complexity and without removing the fundamental dependence in the MCMC sampled values. For further comparison of MCMC and importance sampling methods, see Stephens and Donnelly (2000).

Although implemented here only for mixture models, the method could be extended to the calculation of Bayes factors for other types of latent or missing data models. We think that the method could be useful for some other models for which the EM algorithm can be used to obtain maximum likelihood or posterior mode parameter estimates. More generally, it seems that the basic idea of IMIS, namely incrementally adding components to a mixture importance sampling function to cover areas of substantial contribution to the integral not well covered by the current function, could be applied to many importance sampling problems. Of course, there are issues of implementation to be addressed for each application.

Other approaches have been proposed for approximating the integrated likelihood for mixture models by using EM algorithm output. One approach is the Cheeseman-Stutz (1995) estimator

$$\hat{I}_{\mathrm{CS}} = p(y|\hat{z})\frac{p(\hat{z})}{h(\hat{z})},$$

where $h(\hat{z}) = p(\hat{z}|y, \hat{\tau})$. This is related to a simple case of an importance sampling function on the component labels. If one were to sample T times from $h(z)$, the importance sampling estimator would be

$$\hat{I}_{\mathrm{IS}} = \frac{1}{T}\sum_{t=1}^{T} p(y|z^t)\frac{p(z)}{p(z|x, \hat{\tau})} \equiv \frac{1}{T}\sum_{t=1}^{T} v(z^t).$$

So the Cheeseman-Stutz (1995) estimator is equivalent to taking one summand of the importance sampling estimator above at $\hat{z}$, that is,

$$\hat{I}_{\mathrm{CS}} = v(\hat{z}).$$

This estimator will not be unbiased in general, as the expectation over z is being taken inside the function $v(z)$, rather than outside, and the latter is needed to ensure unbiasedness. Note that the published derivation of the Cheeseman-Stutz estimator was based on a Laplace approximation that is not valid in general for mixture models. Biernacki, Celeux, and Govaert (2000) proposed the integrated classification likelihood (ICL), which is similar to the Cheeseman-Stutz estimator in that it is based on a single value of z, but instead replaces the z in $v(z)$ with $\hat{z}_M$, the most likely labeling of the components given the data and $\hat{\tau}$, and then integrates the resulting completed likelihood over τ. Thus Biernacki et al. (2000) reported

$$p(y|\hat{z}_M) = \int p(y|\hat{z}_M, \tau)p(\tau|\hat{z}_M)d\tau.$$

Biernacki et al. did not suggest that this is an approximation to the integrated likelihood, but instead argued that it is useful in its own right. It is worth noting that $p(y|\hat{z}_M)$ is a component of $v(\hat{z}_M)$ with importance sampling function $h(z) = p(z)$ (i.e., $p(y|\hat{z}_M)$ is one potential summand of $\hat{I}_{\mathrm{MC}}$ where the z^t are sampled from $p(z)$).

The methods proposed in this article are closest in spirit to the adaptive importance sampling methods of Raghavan and Cox (1998). They proposed a method for calculating the optimal δ for defensive mixture importance sampling in the context of estimating several integrands. They used a complex minimization and reweighting scheme to match the asymptotic variances of several importance sampling estimators based on the randomly sampled values. Our method also shares similarities with the adaptive importance sampling method of West (1993). We do not collapse the mixture importance sampling components as West advocated because we felt the reduction in sampling complexity was not worth the additional computational expense.

Our methods are also related to the nonparametric importance sampling estimator of Zhang (1996), in that we choose an adaptive importance sampling function. Adaptive methods for estimating an intractable $p(z)$ were described by Escobar (1995), Givens and Raftery

(1996), and Oh and Berger (1993). Our approach here is to use a mixture to approximate $p(z|y)$, rather than $p(z)$.

Owen and Zhou (2000) discussed the use of control variates to improve the performance of defensive mixture importance sampling for integration. The control variate method provided impressive gains in efficiency for their examples. It should be possible to improve results by using their method at certain places in the IMIS algorithm. It might prove useful because of the large number of adaptive components used, as their method uses the components of a mixture importance sampling function as control variates to reduce both the potential for underestimation of integrands and to guard against high variance of estimates.

There are various other potential ways to improve our method, albeit at the cost of increased complexity. Based on our experience to date, we have used $\delta = 0.5$ (i.e., taking half of the samples from the prior at all stages) and $\delta_k = \frac{0.5}{K-1}$ for all other components. Hesterberg (1995) suggested values of δ between 0.1 and 0.5. A possible improvement on the current method would be to estimate δ from previous importance sampling runs (Raghavan and Cox 1998).

Another parameter of the adaptive IS method that one needs to consider is T, the number of samples drawn at each step of the algorithm. We take T to be 10,000 and 100,000 in the two examples. There is a trade-off between choosing large values for K and T. Increasing K for fixed T gives accurate results in a reasonable amount of time. One could suggest, however, a schedule of T_K such that the number of samples at each iteration varied for different numbers of adaptive components. We have roughly implemented this by increasing T to $10T$ for the final estimate of I in the implementation for this article.

ACKNOWLEDGMENTS

The research of Steele was supported by a grant from the Natural Sciences and Engineering Council of Canada, the research of Steele and Raftery was supported by Office of Naval Research Grants N00014-96-1-0192 and N00014-96-1-0330, and Raftery's research was also supported by NIH Grant 1R01CA094212-01 and ONR Grant N00014-01-10745. Emond's research was supported by National Institute of Health Grant 1R29CA77607. The authors are grateful to David Bryant, Manisha Desai and Matthew Stephens for helpful discussions, to Faming Liang for providing help with the EMC method, and to the reviewers for their helpful comments.

[Received December 2003. Revised November 2005.]

REFERENCES

Atwood, L. D., Wilson, A. F., Elston, R. C., and Bailey-Wilson, L. E. (1992), "Computational Aspects of Fitting a Mixture of Two Normal Distributions Using Maximum Likelihood," *Communications in Statistics, Part B—Simulation and Computation*, 21, 769–781.

Biernacki, C., Celeux, G., and Govaert, G. (2000), "Assessing a Mixture Model for Clustering With the Integrated Complete Likelihood," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 719–725.

Celeux, G. (1997), Discussion of "Bayesian Analysis of Mixtures With an Unknown Number of Components" by Richardson and Green *Journal of the Royal Statistical Society*, Series B, 59, 775–776.

Celeux, G., Hurn, M., and Robert, C. P. (2000), "Computational and Inferential Difficulties With Mixture Posterior Distributions," *Journal of the American Statistical Association*, 95, 957–970.

Cheeseman, P., and Stutz, J. (1995), "Bayesian Classification (AutoClass): Theory and Results," in *Advances in*

Knowledge Discovery and Data Mining, eds. U. Fayyad, G. Piatesky-Shapiro, P. Smyth, and R. Uthurasamy, Menlo Park, CA: AAAI Press, pp. 153–180.

Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. (2000), *Monte Carlo Methods in Bayesian Computation*, Berlin: Springer-Verlag Inc.

Chib, S. (1995), "Marginal Likelihood From the Gibbs Output," *Journal of the American Statistical Association*, 90, 1313–1321.

Chib, S., and Jeliazkov, I. (2001), "Marginal Likelihood from the Metropolis-Hastings Output," *Journal of the American Statistical Association*, 96, 270–281.

Desai, M. (2000), "Mixture Models for Genetic Changes in Cancer Cells," unpublished Ph.D. thesis, University of Washington, Department of Biostatistics.

Desai, M., and Emond, M. (2004), "A New Mixture Model Approach to Analyzing Allelic-Loss Data Using Bayes Factors," *BMC Bioinformatics*, 5, 1–12.

Escobar, M. D. (1995), "Nonparametric Bayesian Methods in Hierarchical Models," *Journal of Statistical Planning and Inference*, 43, 97–106.

Evans, M., and Swartz, T. (1995), "Methods for Approximating Integrals in Statistics with Special Emphasis on Bayesian Integration Problems," *Statistical Science*, 10, 254–272.

Fraley, C., and Raftery, A. E. (1998), "How Many Clusters? Which Clustering Method?—Answers via Model-Based Cluster Analysis," *The Computer Journal*, 41, 578–588.

Gelman, A., and Meng, X.-L. (1998), "Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling," *Statistical Science*, 13, 163–185.

Geyer, C. J. (1991), "Reweighting Monte Carlo Mixtures," Technical Report 518, School of Statistics, University of Minnesota.

Gilks, W. R., and Wild, P. (1992), "Adaptive Rejection Sampling for Gibbs Sampling," *Applied Statistics*, 41, 337–348.

Givens, G. H., and Raftery, A. E. (1996), "Local Adaptive Importance Sampling for Multivariate Densities with Strong Nonlinear Relationships," *Journal of the American Statistical Association*, 91, 132–141.

Grunwald, G. K., Raftery, A. E., and Guttorp, P. (1993), "Time Series of Continuous Proportions," *Journal of the Royal Statistical Society*, Series B, 55, 103–116.

Hammersley, J., and Handscomb, D. (1964), *Monte Carlo Methods*, New York: Wiley.

Hesterberg, T. (1995), "Weighted Average Importance Sampling and Defensive Mixture Distributions," *Technometrics*, 37, 185–194.

Jeffreys, W. H. (1961), *Theory of Probability* (3rd ed.), Clarendon, TX: Clarendon Press.

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Kass, R. E., and Wasserman, L. (1995), "A Reference Bayesian Test for Nested Hypotheses and its Relationship to the Schwarz Criterion," *Journal of the American Statistical Association*, 90, 928–934.

Lewis, S. M., and Raftery, A. E. (1997), "Estimating Bayes Factors via Posterior Simulation with the Laplace-Metropolis Estimator," *Journal of the American Statistical Association*, 92, 648–655.

Liang, F., and Wong, W. H. (2001), "Real-Parameter Evolutionary Monte Carlo With Applications to Bayesian Mixture Models," *Journal of the American Statistical Association*, 96, 653–666.

Lindsay, B. G. (1995), *Mixture Models: Theory, Geometry and Applications*. Hayward, CA: Institute of Mathematical Statistics.

Liu, J. S., Chen, R., and Wong, W. H. (1998), "Rejection Control and Sequential Importance Sampling," *Journal of the American Statistical Association*, 93, 1022–1031.

MacEachern, S. N., Clyde, M., and Liu, J. S. (1999), "Sequential Importance Sampling for Nonparametric Bayes Models: The Next Generation," *The Canadian Journal of Statistics*, 27, 251–267.

Meng, X.-L., and Wong, W. H. (1996), "Simulating Ratios of Normalizing Constants via a Simple Identity: A Theoretical Exploration," *Statistica Sinica*, 6, 831–860.

Neal, R. M. (1998), "Erroneous Results in "Marginal Likelihood from the Gibbs Output." Available online at *http://www.cs.utoronto.ca/~radford*.

——— (2003), "Slice Sampling," *The Annals of Statistics*, 31, 705–767.

Oh, M.-S., and Berger, J. O. (1993), "Integration of Multimodal Functions by Monte Carlo Importance Sampling," *Journal of the American Statistical Association*, 88, 450–456.

Owen, A., and Zhou, Y. (2000), "Safe and Effective Importance Sampling," *Journal of the American Statistical*

Association, 95, 135–143.

Postman, M., Huchra, J., and Geller, M. (1986), "Probes of Large-Scale Structure in the Corona Borealis Region," *The Astronomical Journal*, 92, 1238–47.

Raftery, A. E. (1995), "Bayesian Model Selection in Social Research" (with discussion), *Sociological Methodology*, 25, 111–193.

——— (1996a), "Approximate Bayes Factors and Accounting for Model Uncertainty in Generalised Linear Models," *Biometrika*, 83, 251–266.

——— (1996b), "Hypothesis Testing and Model Selection," in *Markov Chain Monte Carlo in Practice*, eds. W. R. Gilks, D. J. Spiegelhalter, and S. Richardson, London: Chapman & Hall, pp. 163–188.

Raghavan, N., and Cox, D. D. (1998), "Adaptive Mixture Importance Sampling," *Journal of Statistical Computation and Simulation*, 60, 237–259.

Reaven, G. M., and Miller, R. G. (1979), "An Attempt to Define the Nature of Chemical Diabetes Using a Multidimensional Analysis," *Diabetologia*, 16, 17–24.

Richardson, S., and Green, P. J. (1997), "On Bayesian Analysis of Mixtures with an Unknown Number of Components," *Journal of the Royal Statistical Society*, Ser. B, 59, 731–758.

Roeder, K. (1990), "Density Estimation With Confidence Sets Exemplified by Superclusters and Voids in the Galaxies," *Journal of the American Statistical Association*, 85, 617–624.

Rozenkranz, S. L., and Raftery, A. E. (1994), "Covariate Selection in Hierarchical Models of Hospital Admission Counts: A Bayes Factor Approach," Technical Report 268, Department of Statistics, University of Washington.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Steele, R. (2002), "Practical Importance Sampling Methods for Finite Mixture Models and Multiple Imputation," Ph.D. thesis, University of Washington.

Stephens, M. (1997), "Discussion of "On the Bayesian Analysis of Mixtures with an Unknown Number of Components" by Richardson and Green, *Journal of the Royal Statistical Society*, Series B, 59, 768–769.

Stephens, M. (2000a), "Bayesian Analysis of Mixture Models with an Unknown Number of Components—An Alternative to Reversible Jump Methods," *The Annals of Statistics*, 28, 40–74.

——— (2000b), "Dealing With Label Switching in Mixture Models," *Journal of the Royal Statistical Society*, Series B, 62, 795–809.

Stephens, M., and Donnelly, P. (2000), "Inference in Molecular Population Genetics," *Journal of the Royal Statistical Society*, Series B, 62, 605–655.

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

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

Wei, G. C. G., and Tanner, M. A. (1990), "A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms," *Journal of the American Statistical Association*, 85, 699–704.

West, M. (1993), "Approximating Posterior Distributions by Mixtures," *Journal of the Royal Statistical Society*, Series B, 55, 409–422.

Zhang, P. (1996), "Nonparametric Importance Sampling," *Journal of the American Statistical Association*, 91, 1245–1253.