Computing Normalizing Constants for Finite Mixture Models via Incremental Mixture Importance Sampling (IMIS) [1]

Russell J. Steele
McGill University

Adrian E. Raftery
University of Washington

Mary J. Emond
University of Washington

Technical Report 436
Department of Statistics
University of Washington
July 30, 2003

[1]Russell J. Steele is an Assistant Professor of Mathematics and Statistics, McGill University, 805 Sherbrooke O., Montreal, PQ, Canada H3A 2K6 Email: 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. Email: raftery@stat.washington.edu, Web: www.stat.washington.edu/raftery; and Mary J. Emond is Research Assistant Professor of Biostatistics, University of Washington, Box 357237, Seattle, WA 98195-7237; email: emond@u.washington.edu.
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 Manisha Desai and Matthew Stephens for helpful discussions and to Faming Liang for providing help with the EMC method.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
30 JUL 2003		**00-07-2003 to 00-07-2003**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Computing Normalizing Constants for Finite Mixtue Models via Incremental Mixture Importance Sampling (IMIS)	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Washington,Department of Statistics,Box 354322,Seattle,WA,98195-4322**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**32**	

Abstract

We propose 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 important 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 it 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 molecular biology, astronomy, and medical research.
Key Words: Allelotype; Bayes factor; Bayesian model averaging; Beta-binomial distribution; Defensive mixture importance sampling; Gibbs sampling; Label-switching; Markov chain Monte Carlo; Multimodality.

Contents

List of Tables

List of Figures

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

$$I \equiv \mathrm{pr}(\mathbf{x}) = \int f(\mathbf{x}|\tau)p(\tau)d\tau, \tag{1}$$

where $\mathbf{x}$ denotes the observed data, $f(\mathbf{x}|\tau)$ 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.

Since 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, et al. (2000) include 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 are 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 *et al*, 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 approx-

imations 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.

The original justification of the BIC was in terms of the Laplace method and the BIC provides a good approximation to the integrated likelihood in regular models for a unit information prior on the parameters (Kass and Wasserman 1995; Raftery 1995). This justification does not hold for mixture models, although the BIC does provide a consistent estimate of the number of components in the mixture (Leroux 1992; Keribin 2000), leads to density estimates that are consistent for the true density (Roeder and Wasserman 1997), and has given good results in a range of applications (e.g. Dasgupta and Raftery 1998; Fraley and Raftery 1998, 2002).

Quadrature methods can be used but they begin to break down for problems with more than about four parameters (Evans and Swartz 1995), and the number of parameters in mixture models quickly surpasses this as the number of groups and/or the dimension of the data increase. When testing or comparing mixture models, one is typically considering at least some models that have substantial numbers of parameters.

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.

Reversible jump and marked point process MCMC methods can be used to estimate Bayes factors and posterior model probabilities for mixture models (Richardson and Green 1997; Stephens 2000a), but they are also prone to label-switching problems, and convergence can be even more of a problem than for regular MCMC. For example, in the rejoinder to the discussion of their paper, Richardson and Green (1997) mentioned that diagnostics indicated that their method may not have converged even after 500,000 iterations for the one-dimensional mixture model of the galaxy data they analyzed. The difficulty of implementing reversible jump MCMC efficiently for mixture models seems to increase with the dimension of the data.

Our goal in this paper 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 method consists first of reformulating the model in terms of the unobserved group memberships, z, as is done, for example, for estimation via the EM algorithm, and then estimating the integrated likelihood using importance sampling on z. Analytic or quadrature methods are used for integration over τ; this is a "complete-data" problem and is often straightforward.

The success of any importance sampling method depends critically on the importance sampling function. Here we develop importance sampling functions 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.

Our goals for integrated likelihood estimators are reasonably accurate estimates of the integral, realistic estimates of the accuracy with which the integral is estimated, straightforward monitoring of convergence for estimates, practicality (in terms of programming time and computational time), and conceptual simplicity. Simulation studies suggest that our strategy meets all these goals.

The resulting strategy is easy to implement, involving only two simple dependent multinomial sampling schemes. The samples on which the estimate is based are independent, so there are no problems of standard error estimation due to dependence between samples, as with MCMC. Our approach includes a way of dealing with the label-switching problem that can plague other approaches, and also is self-monitoring.

An 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.

In Section 2, we review mixture models, present our importance sampling based estimators, and show how they are applied in the context of multimodality due to the label–switching problem. Section 3 contains simulation results motivated by an application in molecular biology and applications to problems in astronomy and diabetes research. In Section 4, we discuss 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 of a random sample from a G–component mixture distribution. The corresponding likelihood is

$$\prod_{i=1}^{n} \sum_{j=1}^{G} \pi_j f_j(y_i|\theta_j) \equiv \prod_{i=1}^{n} p(y_i|\theta, \pi, G), \tag{2}$$

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:

$$I(y) \equiv \int_{\theta} \prod_{i=1}^{n} p(y_i|\theta, \pi) p(\theta, \pi) d\theta d\pi, \tag{3}$$

where $p(\theta, \pi)$ is the prior density for (θ, π). Analytic integration of (3) 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 i^{th} 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

$$\begin{aligned} I(y) &= \int_{\tau} \prod_{i=1}^{n} \sum_{z_i} p(y_i|z_i, \theta, \pi) p(z_i|\theta, \pi) p(\theta, \pi) d\tau & (4) \\ &= \sum_{z} \int_{\tau} \prod_{i=1}^{n} \prod_{j=1}^{G} (\pi_j f_j(y_i|\theta))^{z_{ij}} p(\theta, \pi) d\tau. & (5) \end{aligned}$$

In (4), the summation is over the G possible values of z_i, and in (5) 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, since the inner integral of (5) (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:

$$I = \sum_z p(y|z)p(z). \tag{6}$$

Here, $p(y|z) = \int_\tau \prod_{i=1}^n p(y_i|z_i,\theta)p(\theta)d\theta$, 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 (2001) treat cases where numerical methods are needed for integration with respect to (θ,π).

2.2 Defensive Mixture Importance Sampling

Because of the size of the space of possible allocations of observations to components, sampling based methods are required to calculate the integral (6). A simple Monte Carlo approach to integration would sample the z from their marginal distribution $p(z|G)$. A Dirichlet prior on λ 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

$$\hat{I}_{MC} = \frac{1}{T}\sum_{t=1}^T p(y|z^t,G). \tag{7}$$

where the $z^t, t = 1...T$ are sampled from $p(z|G)$. 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 highly unstable $\hat{I}_{MC}$ 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

$$\hat{I}_{IS} = \frac{1}{T}\sum_{t=1}^T p(y|z^t,G)w(z^t) = \frac{1}{T}\sum_{t=1}^T p(y|z^t,G)\frac{p(z^t|G)}{g(z^t)}. \tag{8}$$

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

$$\hat{I}_{IS} = \frac{1}{T}\sum_{t=1}^T p(y|z^t,G)\frac{p(z^t|G)}{p(z^t|y,G)} = p(y|G), \tag{9}$$

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

Wei and Tanner (1990) suggest that $p(z|\hat{\tau}, y, G)$ would make a good substitute for $p(z|y, G)$ in an imputation context (where, in their case, z represented missing observations). 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 $p(z|y, G)$. An overly concentrated importance sampling function can also cause 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 $p(z|G)$ and $p(y|z, G)$, 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 was the first to suggest using the Monte Carlo sampling function, $p(z|G)$, as a component of the mixture importance sampling function $\delta p(z|G) + (1 - \delta)g(z)$, giving the following importance sampling estimator

$$\hat{I}_{DM} = \frac{1}{T}\sum_{t=1}^{T} p(y|z^t, G)\frac{p(z^t|G)}{\delta p(z|G) + (1-\delta)g(z)} = \frac{1}{T}\sum_{t=1}^{T} p(y|z^t, G)w^*(z^t) \tag{10}$$

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 $\hat{I}_{MC}$ estimator, while $\delta = 0$ gives the $\hat{I}_{IS}$ estimator with importance sampling function $g(z)$. Hesterberg notes that a K-component mixture could also be used

$$h(z) = \sum_{k=1}^{K-1} \delta_k g_k(z) + \delta_K p(z|G) \tag{11}$$

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, $p(z|\hat{\tau}, y, G)$. Sampling from $p(z|\hat{\tau}, y, G)$ 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 $p(z|\hat{\tau}, y, G)$ 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 $p(z|\hat{\tau}, y, G)$ would result in either an underestimation of $p(y|G)$ (because certain parts of the space would rarely be visited) or estimates of $p(y|G)$ with high empirical variance (because of the larger importance sampling weights associated with sampling rare component labelings under a particular $\hat{\tau}$). Another difficulty with sampling from $p(z|\hat{\tau}, y, G)$ is that $p(z|\hat{\tau}, y, G)$ 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 $p(z|\hat{\tau}, y, G)$, we suggest two other importance sampling functions based on $p(z|\hat{\tau}, y, G)$ in the following subsections.

2.3.1 A Label-Switching Dependent Product of Multinomials

First, we introduce the following label-switching version of $p(z|\hat{\tau}, y, G)$. In most problems, there will be observations that have values of $p(z_i|\hat{\tau}, y, G)$ 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 $p(z|\hat{\tau}_s, y, G)$, 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 $\hat{z}_1$ be the $n \times G$ matrix of $p(z_{ij}|\hat{\tau}_1, y, G)$ 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

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

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

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

where k_1 is the group to which observation 1 was assigned and $l_1 = \text{argmax}_j \hat{z}_{1j}$, i.e. the $\hat{z}_1$ 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 $\hat{z}_1$,; 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 $\text{argmax}_j \hat{z}_{2j} \neq k_1$:

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

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

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

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:

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

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 $\hat{z}_1$ matrix. A specific example of this dependent sampling method makes the algorithm a bit more clear.

Example 1. Consider the following example $\hat{z}$ matrix, for $n = 4$ observations and $G = 3$ components:

$$\hat{z}_1 = \begin{bmatrix} 0.0 & 1.0 & 0.0 \\ 0.99 & 0.005 & 0.005 \\ 0.60 & 0.30 & 0.10 \\ 0.25 & 0.40 & 0.35 \end{bmatrix}$$

The algorithm assigns the first observation to groups uniformly. Assume that observation 1 has been assigned to component 3. That means that observation 2 will be assigned in the following manner:

$$\begin{aligned} \Pr(z_{21} = 1) &= \Pr(z_{22} = 1) = \frac{0.995}{2} = 0.4475 \\ \Pr(z_{23} = 1) &= 0.005 \end{aligned}$$

Now say that observation 2 has been assigned to component 2. Then, assign observations 3 and 4

according to the following probabilities:

$$\begin{aligned} \Pr(z_{31} = 1) &= 0.10 \qquad \Pr(z_{41} = 1) = 0.35 \\ \Pr(z_{32} = 1) &= 0.60 \qquad \Pr(z_{42} = 1) = 0.25 \\ \Pr(z_{33} = 1) &= 0.30 \qquad \Pr(z_{43} = 1) = 0.40 \end{aligned}$$

Note how the sampling will be equivalent to sampling from a permuted $\hat{z}$ matrix for observations 3 and 4.

Now consider a second case. Let observation 1 again be assigned to group 3, but now assume that observation 2 is also assigned to group 3. Then the algorithm specifies the following sampling probabilities for observation 3:

$$\begin{aligned} \Pr(z_{31} = 1) &= \Pr(z_{32} = 1) = \frac{0.70}{2} = 0.35 \\ \Pr(z_{33} = 1) &= 0.30. \end{aligned}$$

And then, if observation 3 is assigned to group 1, then one uses the following probabilities for observation 4:

$$\begin{aligned} \Pr(z_{41} = 1) &= 0.25 \\ \Pr(z_{42} = 1) &= 0.35 \\ \Pr(z_{43} = 1) &= 0.40. \end{aligned}$$

□

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 $Pr(z = z^*) = Pr(z_1 = z_1^*)Pr(z_2 = z_2^* | z_1 = z_1^*) \cdots Pr(z_n = z_n^* | z_{n-1} = z_{n-1}^*, ..., z_1 = z_1^*)$, where each probability is determined by the algorithm and requires only recording the probability used at the time of sampling. It may be viewed as a sequential importance sampling function (Liu et al. 1998; MacEachern et al. 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, $p(z|G)$, 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 amongst observations which have high posterior probability of belonging to the same group.

As before, let $\hat{z}_1$ be the matrix of $p(z_{ij}|\hat{\tau}_1, y, G)$ 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 non-empty group $r, (r = 1, .., G)$, sample η_r from a Dirichlet distribution with parameter vector $\alpha_r = (\alpha_{r1}, ..., \alpha_{rG})$. Now, for each group, re-assign observations to groups according to their group-specific η_r.

Example 2. The $\hat{z}_1$ matrix from the previous example yields the following initial groupings of observations:

Group 1	:	Observations 2,3
Group 2	:	Observations 1,4
Group 3	:	Empty

Draw η_1 from a Dirichlet $(\alpha_{11}, ..., \alpha_{1G})$ and η_2 from a Dirichlet $(\alpha_{21}, ..., \alpha_{2G})$. Next assign observations according to the following:

$$\begin{aligned} \Pr(z_{1j} = 1) &= \Pr(z_{4j} = 1) = \eta_{2j} \\ \Pr(z_{2j} = 1) &= \Pr(z_{3j} = 1) = \eta_{1j} \end{aligned}$$

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 $p(z|\hat{\tau}, y, G)$ 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

$$g_2(z) = \prod_{r=1}^{G} \frac{\Gamma(\sum_j \alpha_{rj})}{\Gamma(\sum_j n_{rj} + \alpha_{rj})} \frac{\prod_j \Gamma(n_{rj} + \alpha_{rj})}{\prod_j \Gamma(\alpha_{rj})} \tag{12}$$

where n_{rj} is the number of observations from the r^{th} group assigned to the j^{th} group. $\triangle$

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

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

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 $\hat{\tau}_j$'s, where each $\hat{\tau}_j$ corresponds to a local posterior mode.

Implementation requires a method for choosing the $\hat{\tau}_j$'s. A missed area of high posterior probability is indicated by a high importance sampling weight. 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 weight at the previous iteration, denoted by z^*, and has the form $p(z|\tau^*, y, G)$, where τ^* is the mode of $p(\tau|y, z^*, G)$.

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 $\hat{\tau}$. We call the resulting algorithm Incremental Mixture Importance Sampling (IMIS). Here is a summary of it:

1. Sample a set of T label assignments $z^{(0)}$ from the mixture distribution

$$g^{(0)}(z) = \delta_1 g_1(z) + \delta_2 g_2(z) + \delta_3 p(z|G), \tag{13}$$

   where $p(z|G)$ is the induced prior distribution on z, $g_1(z)$ is the label-switching product of multinomials described in Section 2.3.1, and $g_2(z)$ is the product of Dirichlet-Multinomials described in Section 2.3.2, where both of the non-prior functions are based on the original $\hat{z}$ matrix of conditional probabilities $p(z|\hat{\tau}^{(0)}, y, G)$, obtained, for instance, via the EM algorithm. For now, set $\delta_1 = \delta_2 = .25$ and $\delta_3 = .5$.

2. Use the sampled z^t's to calculate an initial estimate of I:

$$\hat{I}^0 = \sum_{t=1}^{T} p(y|z^t, G) \frac{p(z)}{g^{(0)}(z)}$$

and store the z^t that yields the largest value of $p(y|z^t, G)\frac{p(z^t|G)}{g^{(0)}(z^t)}$ (call this z_0^*).

3. Find the posterior mode of $p(\tau|y, z_0^*, G)$, $\hat{\tau}_1$.

4. Create a new $\hat{z}$ matrix ($\hat{z}_1$) based on $\hat{\tau}_1$, i.e. using $p(z_{ij}|\hat{\tau}_1, y, G)$.

5. Add two new components to the previous importance sampling function to create a new mixture importance sampling function:

$$g^{(1)}(z) = \delta_1 g_1(z) + \delta_2 g_2(z) + \delta_3 g_3(z) + \delta_4 g_4(z) + \delta_5 p(z)$$

where $g_1(z)$ and $g_2(z)$ are the same as before and $g_3(z)$ and $g_4(z)$ are the label-switching product multinomial and the product of Dirichlet multinomials based on $\hat{z}_1$. Set $\delta_i = \frac{1-\delta_5}{4}$ for $i = 1, .., 4$ and $\delta_5 = .5$.

6. Return to step (1) and sample from the new mixture distribution $g^{(1)}(z)$, repeating steps (1)-(6), adding two components at a time based on newly generated $\hat{z}_k$'s, until $\hat{I}^k$ becomes relatively stable.

7. After obtaining a K component adaptive mixture by finding $(K-1)/2$ $\hat{\tau}_j$'s, one can take a larger sample from this final mixture importance sampling function (say with $T_K = 10T$).

3 Examples

We now present three applications of the adaptive importance sampling method. The first set of examples and simulations focuses on a simple binomial mixture problem from cancer research, establishing that the method works in a situation where results can be compared to a known value of I. The second example is a common one-dimensional Gaussian mixture example showing that the method works well compared to a Markov Chain Monte Carlo method. Finally, we give a higher-dimensional example.

Table 1: Likelihood modes for Data Set 3

$-\log(p(y\|\hat{\tau}, G))$	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\pi}$	$p(y, \tilde{z})$
175.74	.169	.276	.155	2.48
175.78	.220	.151	.5	0.0013
175.91	.185	NA	1.0	11.3

Note: $p(y, \tilde{z})$ is the value of $p(y, z|G)$ evaluated at z values corresponding to the assignment of each observation to the group for which it has maximum $p(z|\hat{\tau}, y, G)$.

3.1 Tumor Site Data

First, we examine the use of the method to calculate integrated likelihood values for a mixture of binomial distributions where the data are in the form of y_i successes in n_i trials. We consider two small tumor-site datasets (Shibagaki, Shimada, Wagata, Ikenaga, Imamura, and Ishizaki 1994; Barrett, Galipeau, Sanchez, Emond, and Reid 1996) where the scientific goal is to determine if there are one or two components in the mixture distribution (Newton, Gould, Reznikoff, and Haag 1998). Data sets 1 and 2 are subsets from the two allelotype data sets respectively. Data set 3 consists of simulated data from a single binomial component with mean 0.22 and n_i's the same as the n_i's in Data set 2. We will use uniform prior distributions for the binomial parameters (μ_g) and the mixing parameters, λ.

The likelihood surfaces for the parameters τ range from highly peaked (in Data Set 1) to very flat, with multiple modes (Data Set 3). Table 1 shows three local modes of the likelihood surface for Data Set 3, along with the values of $p(y, z|G)$ evaluated at z values corresponding to the assignment of each observation to the group for which it has maximum $p(z|\hat{\tau}, y, G)$. Note that the $\hat{\tau}$ that corresponds to the largest value of the likelihood does not correspond to the z that gives the largest value of $p(y, z|G)$. This suggests that an adaptive method that searches for good $\hat{\tau}$'s to use for several $\hat{z}$ sampling matrices may be useful even for simple problems.

Table 2 shows the results using various δ values for two-component importance sampling functions. The first set of results corresponds to using the label-switching importance sampling function in conjunction with the prior as part of a defensive mixture distribution. The choice of δ affects the quality of the $\hat{I}$ estimates. The second set of results corresponds to using the product of Dirichlet-multinomials as the non-prior component of a defensive mixture. Note that using a δ value between 0 and 1 does not greatly affect the accuracy of the results, but setting $\delta = 0$ or 1 results in estimates with high observed and estimated variance.

Table 2: Results for Data Sets 1-3 for Various δ Values Using Two Component Defensive Mixtures for the Two Choices for $g(z)$.

Data set 1: markedly peaked likelihood: True log(Ī) = −43.59						
	Product of Multinomials			Prod. of Dir.-Multinomials		
δ	log($\bar{\hat{I}}$)	CV($\hat{I}$)	$\overline{\text{CV}(\hat{I})}$	log($\bar{\hat{I}}$)	CV($\hat{I}$)	$\overline{\text{CV}(\hat{I})}$
0	-43.59	0.233	0.036	-43.60	0.040	0.041
0.25	-43.58	0.031	0.031	-43.59	0.040	0.048
0.50	-43.59	0.028	0.027	-43.58	0.052	0.058
0.75	-43.59	0.028	0.032	-43.58	0.072	0.080
1.00	-43.63	0.248	0.291	-43.63	0.248	0.291
Data set 2: peaked likelihood : True log(Ī) = −44.55						
	Product of Multinomials			Prod. of Dir.-Multinomials		
δ	log($\bar{\hat{I}}$)	CV($\hat{I}$)	$\overline{\text{CV}(\hat{I})}$	log($\bar{\hat{I}}$)	CV($\hat{I}$)	$\overline{\text{CV}(\hat{I})}$
0.00	-44.60	0.250	0.071	-44.54	0.038	0.038
0.25	-44.55	0.045	0.040	-44.56	0.040	0.043
0.50	-44.55	0.042	0.038	-44.55	0.042	0.051
0.75	-44.55	0.044	0.040	-44.55	0.073	0.068
1.00	-44.61	0.276	0.238	-44.61	0.276	0.238
Data set 3: flat, multimodal likelihood : True log(Ī) = −38.39						
	Product of Multinomials			Prod. of Dir.Multinomials		
δ	log($\bar{\hat{I}}$)	CV($\hat{I}$)	$\overline{\text{CV}(\hat{I})}$	log($\bar{\hat{I}}$)	CV($\hat{I}$)	$\overline{\text{CV}(\hat{I})}$
0.00	-38.51	0.437	0.130	-38.39	0.025	0.025
0.25	-38.39	0.032	0.035	-38.39	0.023	0.022
0.50	-38.39	0.021	0.024	-38.39	0.022	0.021
0.75	-38.38	0.019	0.020	-38.39	0.018	0.020
1.00	-38.39	0.022	0.022	-38.39	0.022	0.022

Notes: Results based on 100 trials of $T = 1,000$ samples from each of the two importance sampling functions. $\bar{\hat{I}}$ is the average of the $\hat{I}$ estimates, CV($\hat{I}$) is the coefficient of variation (CV) of the 100 $\hat{I}$ values, $\bar{CV}(\hat{I})$ is the average of the 100 CV estimates.

Table 3: IMIS Estimates of the Log Integrated Likelihood Values for the Two-Component Model.

Data Set	$\log(\bar{\hat{I}})$	$CV(\hat{I})$	$\overline{CV(\hat{I})}$	Truth
1	-43.59	0.010	0.02	-43.59
2	-44.55	0.010	0.02	-44.55
3	-38.39	0.010	0.009	-38.39
Data Set	$\log(\bar{\hat{I}})$	$CV(\hat{I})$	$\overline{CV(\hat{I})}$	EMC Long run
4	-470.63	0.003	0.018	-470.63
5	-486.80	0.034	0.030	-486.77
6	-386.71	0.027	0.027	-386.72

Notes: Results are based on 100 trials using the adaptive importance sampling method with $K = 11$ components and $T = 10,000$ iterations at each step. The numbers are based on a final run of 100,000.

We then examined 3 artificial data sets based on the tumor site data. Data set 4 includes 12 replicates of data set 1 and data set 5 includes 12 replicates of data set 2. Data set 6 includes the binomial proportion 8/40 repeated 204 times. Data set 6 serves as a benchmark standard that any reasonable method should be able to calculate easily. Table 3 compares results using adaptive methods to the truth (for Data sets 1, 2, and 3) and to results obtained using the EMC approach (for Data sets 4, 5, and 6) described below. A fairly small number of adaptive components (11 components for the table) approximate the desired $p(z|y, G)$ well, producing estimates with very small mean-squared error for all of the data sets.

3.2 Galaxy Data

The next example extends the methodology beyond the realm of testing one component against two component mixtures. We try to approximate the integrated likelihood for data involving velocities of galaxies (Postman, Huchra, and Geller 1986), discussed in Roeder (1990). The number of possible groups of galaxies is the question of interest for this data set. 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 fit 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 $u_i = 0, 0.05, ..., 1.0$ is different for each chain. Their primary contribution was to suggest an evolutionary Monte Carlo approach for

swapping parts of the sample vector τ amongst chains. The annealed MCMC approach allowed the chain to visit the multiple modes of the likelihood. Neal (1998) pointed out that Chib's Gibbs sampling approach (1995) for the same galaxy data set 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, viewed as a 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, i.e.

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

which is similar to the prior used by Chib and by Liang and Wong, with the difference that they did not use a conjugate prior, but instead assumed prior independence of the mean and variance parameters.

Table 4 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 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 4 has no associated standard error. The Monte Carlo estimate based on sampling from the prior on τ is taken over 10 trials of one million Monte Carlo iterations. The IMIS estimate is based on 10 trials of the method. The EMC "short" runs are 10 trials of 125,000 iterations each, thinned to a sample size of 2500 to be used with the bridge sampling calculations. The short runs are an attempt to match the amount of computational time required for a reasonable run of the adaptive importance sampling method. 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.

Table 4 shows that IMIS provides accurate estimates of $\log(\hat{I})$. The coefficients of variation are low and all runs produce $\hat{I}$ estimates within 0.5 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 3 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

Table 4: Comparison of Log Integrated Likelihood Estimates for the Galaxy Data

Clusters	IMIS	EMC (short)	MC estimate using $p(\tau)$	EMC (long)
2	-2.96 (0.001)	-2.99 (0.08)	-2.89 (0.76)	-2.02
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

Notes: Values for IMIS were obtained using a 51-component mixture.
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.
To obtain actual integrated likelihood estimates, subtract 230 from each value in the table.

answer when averaged over the 10 trials, but the variability across those 10 trials is large. Also, sampling from the prior works less well 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 get 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 components used here. Figure 2 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 5 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 every 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 is included here only for comparison.

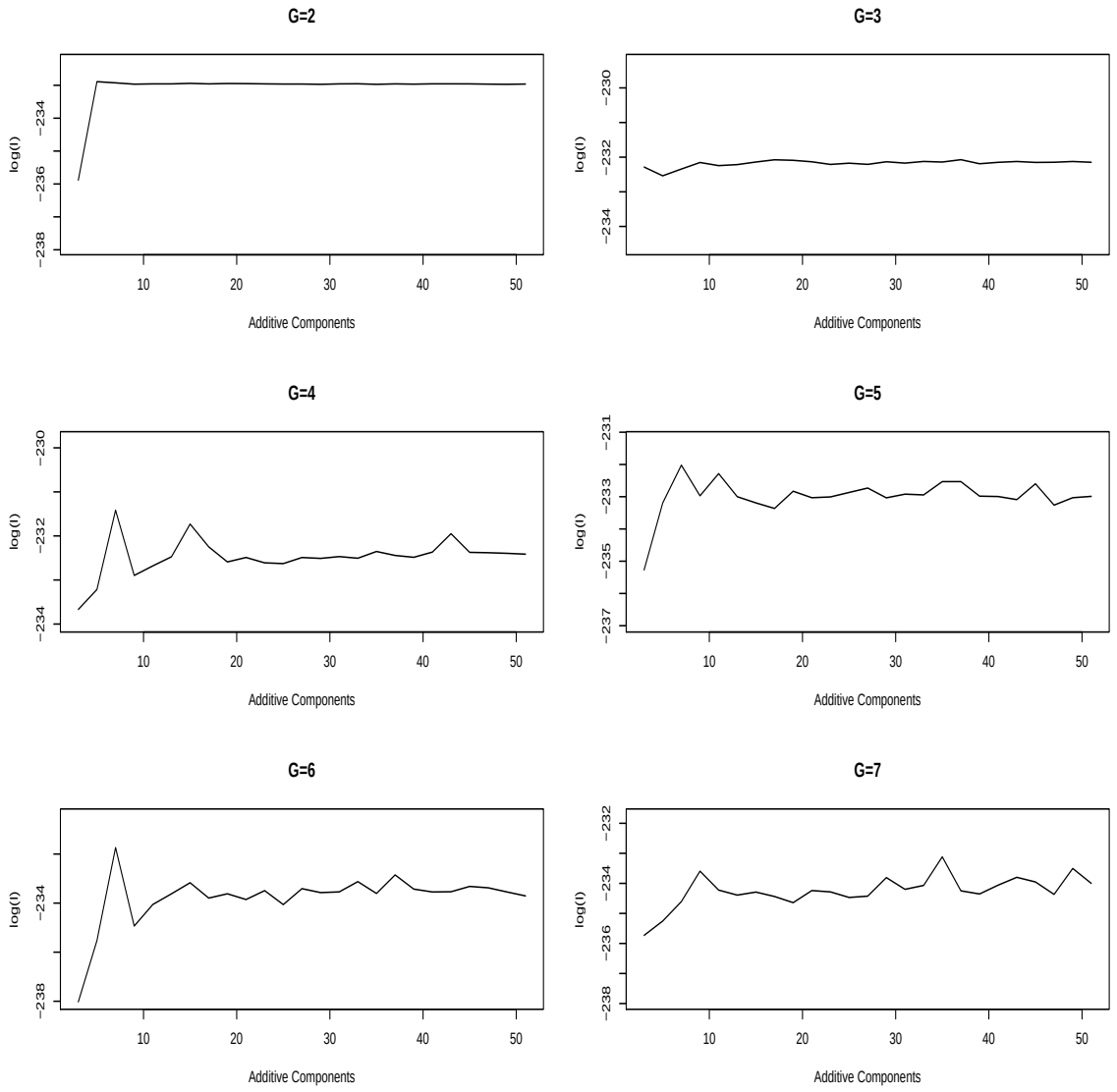

Figure 1: IMIS $\log(\hat{I})$ Trace Plots for Each Number of Components for the Galaxy Data. All runs used a maximum of 51 components based on 25 $\hat{\tau}_j$'s.

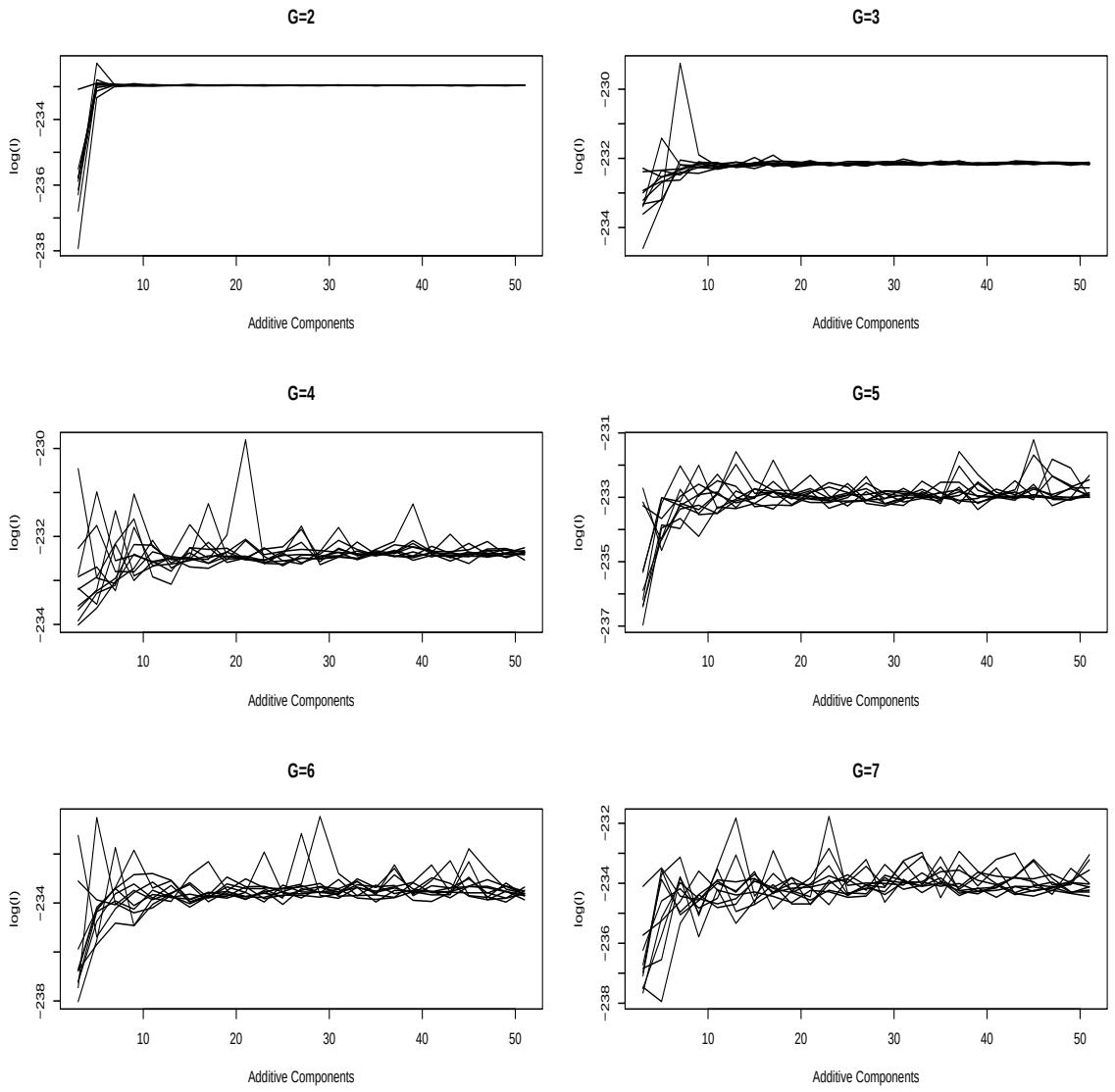

Figure 2: Trace Plots for 10 Runs of the IMIS Method for the Galaxy Data. All runs used a maximum of 51 components based on 25 $\hat{\tau}_j$'s.

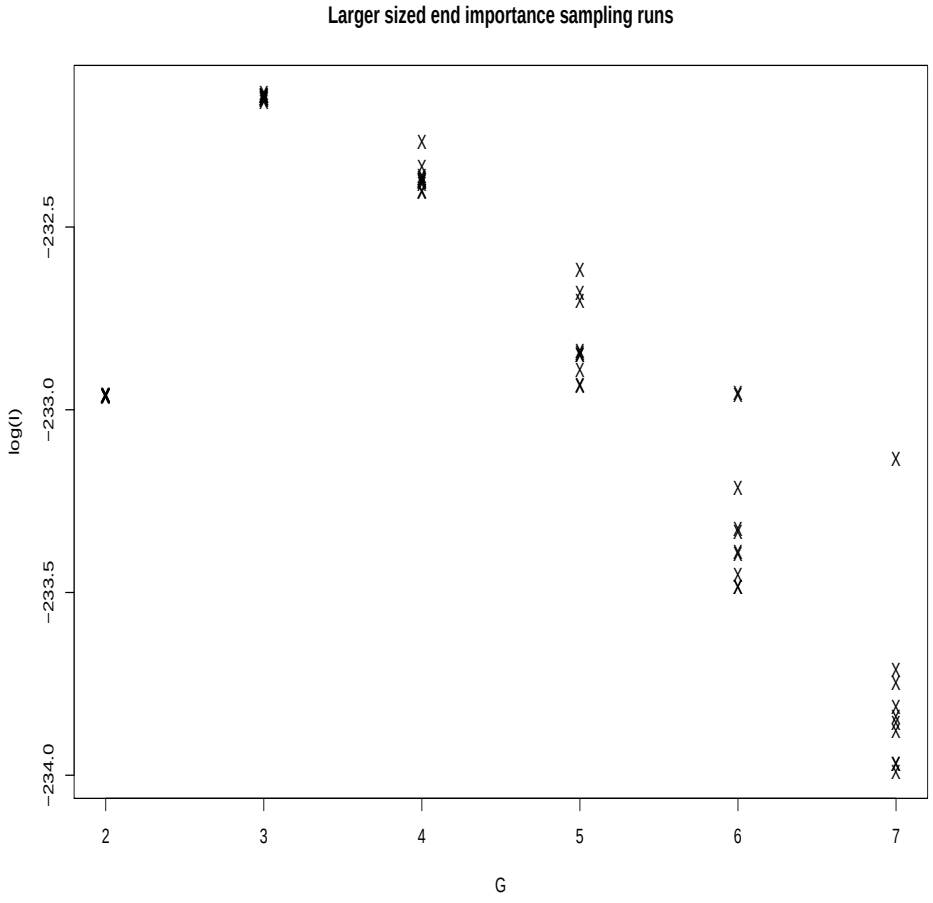

Figure 3: IMIS $\log(\hat{I})$ Final Estimates for the Galaxy Data for 2 to 7 Components. All points are based on 100,000 simulated values from a 51-component mixture importance sampling function.

Table 5: CPU Times for the Galaxy Data

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

Note: Table cells indicate seconds of CPU time required for running each method once to estimated the log integratated likelihood. The methods were implemented as described in the note to Table 4, so that the various runs estimates with relatively equal precision.

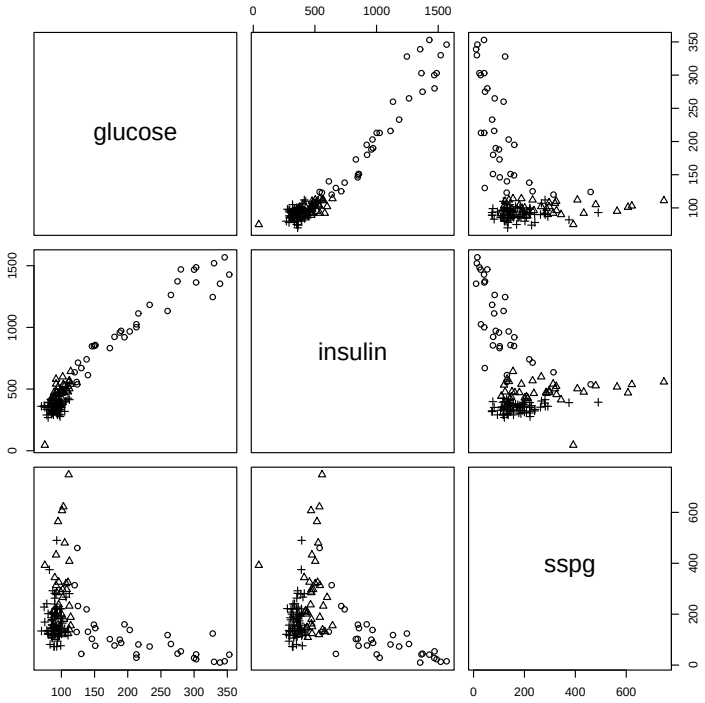

Figure 4: Pairwise Plots of Glucose, Insulin, and SSPG for the Diabetes Dataset. Triangles $\triangle$ denote diagnosed chemical diabetes patients, crosses + denote diagnosed normal patients, and circles o denote diagnosed overt diabetes patients.

3.3 Diabetes Data

Finally we consider a higher-dimensional example from the medical literature (Reaven and Miller 1979). The data set consists of blood measures of insulin, glucose, and insulin resistance levels (SSPG) for 145 diabetes patients; the pairs plot is shown in Figure 4. Fraley and Raftery (1998) analyzed the data set 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 $6 + 12 + 1 = 19$ 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 6 shows the estimated log integrated likelihoods for G from 2 to 5 using $K = 101$ adaptive components and $T = 10,000$ (with a final run of 100,000 to decrease the variability in the final estimate). The coefficients of variation are acceptable and there is not reasonably strong evidence

Table 6: IMIS Log-Integrated Likelihood Estimates for the Diabetes Dataset

Clusters	IMIS Estimate	Estimated CV
2	-2425.78	0.007
3	-2403.93	0.058
4	-2404.25	0.155
5	-2403.99	0.208

Notes: 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.

for more than three components.

4 Discussion

We have proposed a general approach for calculating integrated likelihoods for finite mixture models via mixture importance sampling, called IMIS. This samples the labels after analytic (or approximate) integration of the parameters of the finite mixture rather than trying to integrate numerically over the parameters using the observed data likelihood. We used two types of sampling function in the adaptive mixture. The first type of function is a new label-switching product of dependent multinomial distributions. The second type of function is a product of Dirichlet-multinomials that builds a weak dependency amongst observations that are likely to be from the same component. Both functions are based on the $\hat{z}$ matrix of $p(z_{ij}|\hat{\tau}, y, G)$. We build up the mixture importance sampling function incrementally, adding components one at a time to cover areas of high posterior density that have been largely missed.

The resulting algorithm is relatively easy to implement compared to competing MCMC 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. It is possible to monitor the estimates of $\hat{I}$ (and their estimated variability) as components are added. As the importance sampling function begins to stabilize, one can determine whether the variability across different values of $\hat{I}$ is within the limits needed for interpreting Bayes factors. 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 (for example, 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). 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

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

where $h(\hat{z}) = p(\hat{z}|y, \hat{\tau}, G)$. 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

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

So the Cheeseman-Stutz estimator is equivalent to taking one summand of the importance sampling estimator above at $\hat{z}$, i.e.

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

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 et al. (2000) have 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. report

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

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

Wei and Tanner (1990) first suggested the use of $p(z|\hat{\tau}, y)$ as an importance sampling function in the context of multiple imputation for missing data problems. They used it for posterior density estimation rather than for integrated likelihoods as we do here. As an importance sampling function, we found that $p(z|\hat{\tau}, y, G)$ does not work well by itself as it tends to be too concentrated and to miss mass corresponding to local modes of the likelihood surface, and only can sample from one mode corresponding to the original labeling of the components.

The methods proposed in this paper 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 IS method of West (1993). We don't collapse the mixture importance sampling components as West advocates because we felt 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 $p(z)$ are described by Escobar (1995), Givens and Raftery (1996), and Oh and Berger (1993). Our approach here is to use a mixture to approximate $p(z|y, G)$, 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. An obvious 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 in the examples. There is a tradeoff 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 $10 * T$ for the final estimate of I in the implementation for this paper.

References

Atwood, L. D., A. F. Wilson, R. C. Elston, and L. E. Bailey-Wilson (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.

Barrett, M. T., P. C. Galipeau, C. A. Sanchez, M. J. Emond, and B. R. Reid (1996). Determination of the frequency of loss of heterozygosity in esophageal adenocarcinoma by cell sorting, whole genome amplification and microsatellite polymorphisms. *Oncogene 12*, 1873–8.

Biernacki, C., G. Celeux, and G. Govaert (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). Contribution to the discussion of Richardson and Green, 1997: on Bayesian analysis of mixtures with an unknown number of components (with discussion). *J Royal Stat Soc B 59*, 775–776.

Celeux, G., M. Hurn, and C. P. Robert (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association 95*(451), 957–970.

Cheeseman, P. and J. Stutz (1995). Bayesian classification (AutoClass): Theory and results. In U.Fayyad, G. Piatesky-Shapiro, P. Smyth, and R. Uthurasamy (Eds.), *Advances in knowledge discovery and data mining*, pp. 153–180. Menlo Park, Calif.: AAAI Press.

Chen, M.-H., Q.-m. Shao, and J. G. Ibrahim (2000). *Monte Carlo methods in Bayesian computation*. Springer-Verlag Inc.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association 90*, 1313–1321.

Chib, S. and I. Jeliazkov (2001). Marginal likelihood from the metropolis-hastings output. *Journal of the American Statistical Association 96*(453), 270–281.

Dasgupta, A. and A. E. Raftery (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association 93*, 294–302.

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

Desai, M. and M. Emond (2001). A new importance sampling method to compute bayes factors for mixure models with application to allelic-loss data. Technical Report 173, Department of Biostatistics, University of Washington.

Escobar, M. D. (1995). Nonparametric Bayesian methods in hierarchical models. *Journal of Statistical Planning and Inference 43*, 97–106.

Evans, M. and T. Swartz (1995). Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statistical Science 10*, 254–272.

Fraley, C. and A. E. Raftery (1998). How many clusters? Which clustering method? - Answers via model-based cluster analysis. *The Computer Journal 41*, 578–588.

Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association 97*(458), 611–631.

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

Givens, G. H. and A. E. Raftery (1996). Local adaptive importance sampling for multivariate densities with strong nonlinear relationships. *Journal of the American Statistical Association 91*, 132–141.

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

Hammersley, J. and D. Handscomb (1964). *Monte Carlo Methods*. John Wiley and Sons.

Hesterberg, T. (1995). Weighted average importance sampling and defensive mixture distributions. *Technometrics 37*, 185–194.

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

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

Kass, R. E. and L. Wasserman (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association 90*, 928–934.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhya, Series A, Indian Journal of Statistics 62*(1), 49–66.

Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics 20*, 1350–1360.

Lewis, S. M. and A. E. Raftery (1997). Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator. *Journal of the American Statistical Association 92*, 648–655.

Liang, F. and W. H. Wong (2001). Real-parameter evolutionary Monte Carlo with applications to Bayesian mixture models. *Journal of the American Statistical Association 96*(454), 653–666.

Lindsay, B. G. (1995). *Mixture models: Theory, geometry and applications*. Institute of Mathematical Statistics.

Liu, J. S., R. Chen, and W. H. Wong (1998). Rejection control and sequential importance sampling. *Journal of the American Statistical Association 93*, 1022–1031.

MacEachern, S. N., M. Clyde, and J. S. Liu (1999). Sequential importance sampling for nonparametric Bayes models: The next generation. *The Canadian Journal of Statistics 27*, 251–267.

Meng, X.-L. and W. H. Wong (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". *http://www.cs.utoronto.ca/~radford*.

Newton, M. A., M. N. Gould, C. A. Reznikoff, and J. D. Haag (1998). On the statistical analysis of allelic-loss data. *Statistics in Medicine 17*, 1425–45.

Oh, M.-S. and J. O. Berger (1993). Integration of multimodal functions by Monte Carlo importance sampling. *Journal of the American Statistical Association 88*, 450–456.

Owen, A. and Y. Zhou (2000). Safe and effective importance sampling. *Journal of the American Statistical Association 95*(449), 135–143.

Postman, M., J. Huchra, and M. Geller (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.

Raftery, A. E. (1996a). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika 83*, 251–266.

Raftery, A. E. (1996b). Hypothesis testing and model selection. In *Markov chain Monte Carlo in practice,* edited by W.R. Gilks, D.J. Spiegelhalter and S. Richardson, Chapter 10, pp. 163–188. London: Chapman & Hall.

Raghavan, N. and D. D. Cox (1998). Adaptive mixture importance sampling. *Journal of Statistical Computation and Simulation 60*, 237–259.

Reaven, G. M. and R. G. Miller (1979). An attempt to define the nature of chemical diabetes using a multidimensional analysis. *Diabetologia 16*, 17–24.

Richardson, S. and P. J. Green (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society - Series B 59*, 731–792.

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.

Roeder, K. and L. Wasserman (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association 92*, 894–902.

Rozenkranz, S. L. and A. E. Raftery (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.

Shibagaki, I., Y. Shimada, T. Wagata, M. Ikenaga, M. Imamura, and K. Ishizaki (1994). Allelotype analysis of esophageal squamous cell carcinoma. *Cancer Res 54*, 2996–3000.

Stephens, M. (1997). Contribution to the discussion of Richardson and Green, 1997: On the Bayesian analysis of mixtures with an unknown number of components (with discussion). *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*(1), 40–74.

Stephens, M. (2000b). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, Series B, Methodological 62*(4), 795–809.

Stephens, M. and P. Donnelly (2000). Inference in molecular population genetics. *Journal of the Royal Statistical Society, Series B, Methodological 62*(4), 605–655.

Tierney, L. and J. B. Kadane (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association 81*, 82–86.

Titterington, D. M., A. F. M. Smith, and U. E. Makov (1985). *Statistical analysis of finite mixture distributions*. John Wiley and Sons.

Wei, G. C. G. and M. A. Tanner (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, Methodological 55*, 409–422.

Zhang, P. (1996). Nonparametric importance sampling. *Journal of the American Statistical Association 91*, 1245–1253.

Appendix

The first binomial dataset consists of allelic loss data for the first 17 markers in Table 1 from Barret, et al. (1996). The second binomial dataset consists of selected allelic loss data from markers in Table 1 of Shibagaki, et al. (1994). The third dataset is a random sample of size 17 from binomial distributions with $p_i = .22$ and n_i's equal to those in dataset 2.

Data set 1		Data set 2		Data set 3	
y_i	n_i	y_i	n_i	y_i	n_i
3	15	4	26	1	22
11	17	10	19	2	26
7	17	3	19	2	23
4	17	6	33	2	21
3	18	10	22	2	19
5	15	7	23	3	19
4	15	0	13	3	17
5	15	2	20	5	28
3	19	4	19	4	22
6	16	2	27	4	20
12	15	1	17	5	25
5	18	2	21	3	15
3	19	8	22	7	33
1	18	6	18	5	18
3	19	7	28	4	13
5	19	4	25	6	19
3	21	3	15	10	27