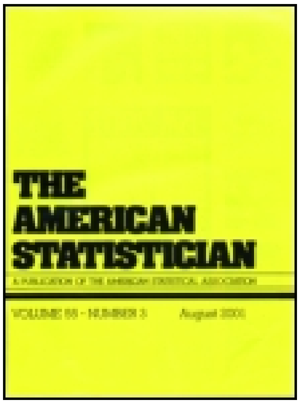
This article was downloaded by: [Linköping University Library]
On: 09 October 2013, At: 08:08
Publisher: Taylor & Francis
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



The American Statistician
Publication details, including instructions for authors and subscription information:
http://amstat.tandfonline.com/loi/utas20

Difficulties in Drawing Inferences With Finite-Mixture Models

Hwan Chung[a], Eric Loken[a] & Joseph L Schafer[a]

[a] Hwan Chung is Research Associate of The Methodology Center, Eric Loken is Assistant Professor of Human Development and Family Studies, and Joseph L. Schafer is Associate Professor, Department of Statistics and The Methodology Center, The Pennsylvania State University, S-159 Henderson, University Park, PA16802 . This research was supported by the National Institute on Drug Abuse Grant 1-P50-DA10075. The authors' names appear in alphabetical order.

Published online: 01 Jan 2012.

To cite this article: Hwan Chung, Eric Loken & Joseph L Schafer (2004) Difficulties in Drawing Inferences With Finite-Mixture Models, The American Statistician, 58:2, 152-158, DOI: 10.1198/0003130043286

To link to this article: http://dx.doi.org/10.1198/0003130043286

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the “Content”) contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at http://amstat.tandfonline.com/page/terms-and-conditions

Statistical Computing and Graphics

Difficulties in Drawing Inferences With Finite-Mixture Models: A Simple Example With a Simple Solution

Hwan CHUNG, Eric LOKEN, and Joseph L. SCHAFER

Downloaded by [Linköping University Library] at 08:08 09 October 2013

Likelihood functions from finite mixture models have many unusual features. Maximum likelihood (ML) estimates may behave poorly over repeated samples, and the abnormal shape of the likelihood often makes it difficult to assess the uncertainty in parameter estimates. Bayesian inference via Markov chain Monte Carlo (MCMC) can be a useful alternative to ML, but the component labels may switch during the MCMC run, making the output difficult to interpret. Two basic methods for handling the label-switching problem have been proposed: imposing constraints on the parameter space and cluster-based relabeling of the simulated parameters. We have found that label switching may also be reduced by supplying small amounts of prior information that are asymmetric with respect to the mixture components. Simply assigning one observation to each component a priori may effectively eliminate the problem. Using a very simple example—a univariate sample from a mixture of two exponentials—we evaluate the performance of likelihood and MCMC-based estimates and intervals over repeated sampling. Our simulations show that MCMC performs much better than ML if the label-switching problem is adequately addressed, and that asymmetric prior information performs as well as or better than the other proposed methods.

KEY WORDS: EM algorithm; Label switching; Markov chain Monte Carlo.

1. INTRODUCTION

Under a finite mixture model, the density of a random variable or vector y has the form

$$f(y;\boldsymbol{\theta}) = \pi_1 f_1(y;\boldsymbol{\lambda}_1) + \pi_2 f_2(y;\boldsymbol{\lambda}_2) + \cdots + \pi_k f_k(y;\boldsymbol{\lambda}_k), \quad (1)$$

where $\pi_1 + \pi_2 + \cdots + \pi_k = 1$. This is a weighted average of densities $f_1, \ldots, f_k$, with f_j present in proportion π_j, and $\boldsymbol{\lambda}_j$ represents a vector of unknown parameters in f_j. Finite mixtures have been applied to a wide variety of data in the physical, social and medical sciences. Overviews of mixture modeling were given by Titterington, Smith, and Makov (1985) and McLachlan and Peel (2000).

With many finite mixtures, maximum-likelihood (ML) estimates are easily calculated by an EM algorithm (Dempster, Laird, and Rubin 1977). Despite the relative ease of ML, the likelihood function has unusual features that make it difficult to obtain reliable inferences. One such feature is the *labeling issue,* the well-known fact that the likelihood is invariant under permutations of the class labels. Labeling does not necessarily create difficulty in ML estimation, because solutions converging to different permutations of a single mode are easily identified. However, other anomalies may arise in certain samples: regions where the likelihood is constant or nearly so; multiple local modes; and suprema on the boundary of the parameter space. These may seriously degrade the performance of ML estimates over repeated samples, making traditional methods for obtaining standard errors inadequate.

Given the difficulties associated with likelihood-based inference, some have adopted a Bayesian approach, simulating random draws of parameters from a posterior distribution using Markov chain Monte Carlo (MCMC) (Robert 1996). MCMC may produce estimates and credible regions for $\boldsymbol{\theta}$ without appealing to large-sample approximations. With MCMC, however, the labeling problem becomes more acute, because class labels may permute during the simulation run, requiring an intelligent strategy for summarizing and interpreting the output stream. Techniques recommended for overcoming this problem—imposing constraints on the parameters (Richardson and Green 1997) and clustering methods (Stephens 1997; Celeux 1998; Stephens 2000; Celeux, Hurn, and Robert 2000)—may perform well in some cases, but are not without difficulty.

This article explores problems in drawing inferences from samples of $n = 100$ observations from a mixture of two exponential distributions. We chose this example because it has already appeared in print (Celeux 1998; Celeux et al. 2000; Seidel, Mosler, and Alker 2000) to illustrate the performance of various techniques. This example is both simple and complex. Estimates are easy to compute, but the sample size is small and the component densities substantially overlap, making the estimation imprecise. We show that ML estimates behave erratically over repeated samples, and that conventional methods for constructing confidence intervals, including asymptotic normal approximations and the bootstrap, tend to perform poorly.

Bayesian methods fare better than ML if the label-switching problem is addressed. We review some of the currently recom-

Hwan Chung is Research Associate of The Methodology Center, Eric Loken is Assistant Professor of Human Development and Family Studies, and Joseph L. Schafer is Associate Professor, Department of Statistics and The Methodology Center, The Pennsylvania State University, S-159 Henderson, University Park, PA 16802 (E-mail: loken@psu.edu). This research was supported by the National Institute on Drug Abuse Grant 1-P50-DA10075. The authors' names appear in alphabetical order.

© 2004 American Statistical Association DOI: 10.1198/000313004043286

Downloaded by [Linköping University Library] at 08:08 09 October 2013

mended strategies, and we suggest a new technique that is both computationally and conceptually simple. Our method is to simply assign one or more observations to components of the mixture with probability one. For the mixture of two exponentials, this trivial modification of MCMC performs as well as, or better than, clustering or deterministic constraints.

2. LIKELIHOOD METHODS

Consider a sample $y_1, \ldots, y_n$ from a mixture of two exponentials with means $1/\lambda_1$ and $1/\lambda_2$,

$$f(y_i;\boldsymbol{\theta}) = \pi\lambda_1 \exp(-\lambda_1 y_i) + (1-\pi)\lambda_2 \exp(-\lambda_2 y_i), \quad (2)$$

$y_i > 0$, where $\boldsymbol{\theta} = (\pi, \lambda_1, \lambda_2)$. Under ordinary circumstances, the ML estimate for $\boldsymbol{\theta}$ solves the score equations, $\partial \log L/\partial\boldsymbol{\theta} = 0$, where $L = \prod_{i=1}^{n} f(y_i;\boldsymbol{\theta})$. In this case, an ML estimate is easily calculated by an EM algorithm. For the E-step, we compute the conditional probability that each y_i came from component 1, given provisional estimates $(\hat{\pi}, \hat{\lambda}_1, \hat{\lambda}_2)$ for the parameters,

$$\hat{\delta}_i = \frac{\hat{\pi}\hat{\lambda}_1 \exp(-\hat{\lambda}_1 y_i)}{\hat{\pi}\hat{\lambda}_1 \exp(-\hat{\lambda}_1 y_i) + (1-\hat{\pi})\hat{\lambda}_2 \exp(-\hat{\lambda}_2 y_i)}. \quad (3)$$

In the M-step, we update the parameter estimates by

$$\hat{\pi} = \frac{\sum_i \hat{\delta}_i}{n}, \quad \hat{\lambda}_1 = \frac{\sum_i \hat{\delta}_i}{\sum_i \hat{\delta}_i y_i}, \quad \hat{\lambda}_2 = \frac{\sum_i (1-\hat{\delta}_i)}{\sum_i (1-\hat{\delta}_i) y_i}. \quad (4)$$

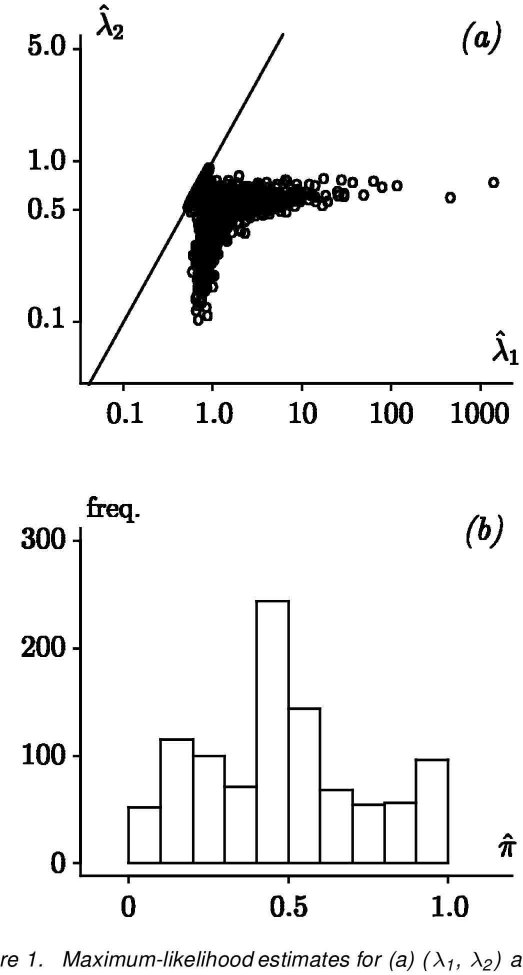


Figure 1. Maximum-likelihood estimates for (a) (λ_1, λ_2) and (b) π from 1,000 samples with π = 0.5, λ_1 = 1, λ_2 = 0.5.

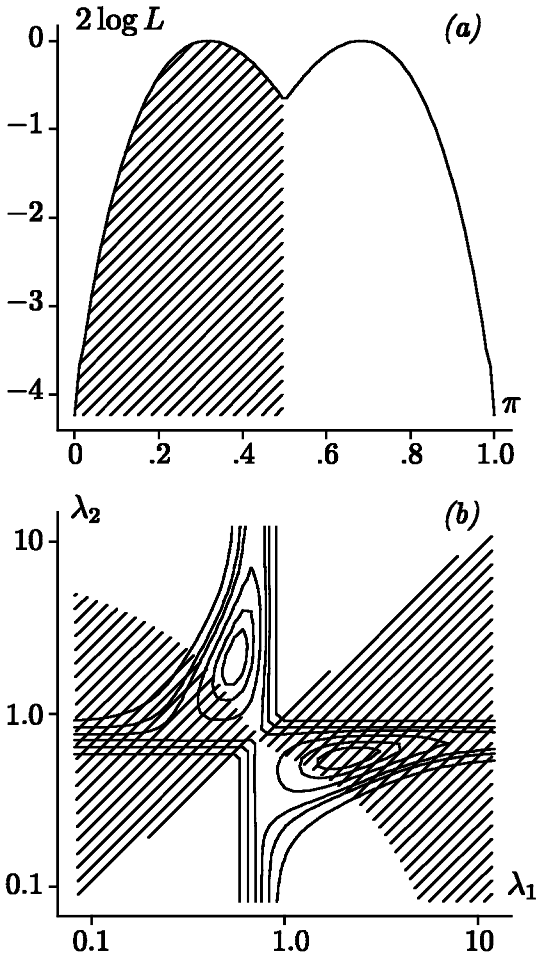


Figure 2. (a) Profile log-likelihood for π, with diagonal shading to indicate where $\hat{\lambda}_1$ > $\hat{\lambda}_2$; (b) profile log-likelihood for (λ_1, λ_2), with diagonal shading to indicate where $\hat{\pi}$ < 0.5.

Iterating these two steps produces a sequence of parameter estimates that converges reliably to a local or global maximum of L. EM algorithms for finite mixture models were reviewed by Titterington et al. (1985); Little and Rubin (1987); McLachlan and Krishnan (1997); and McLachlan and Peel (2000).

The simplicity of EM and the relative ease of ML estimation have made likelihood methods very popular for finite mixtures. However, the likelihood function has some features which can make likelihood-based inferences troublesome. To illustrate the unusual behavior of ML, we simulated data from (2) with $\pi = 0.5$, $\lambda_1 = 1.0$, and $\lambda_2 = 0.5$. We drew 1,000 samples and computed ML estimates $(\hat{\pi}, \hat{\lambda}_1, \hat{\lambda}_2)$ by an EM algorithm, relabeling the classes if necessary to ensure that $\hat{\lambda}_1 \geq \hat{\lambda}_2$. The sampling distribution of the estimates is displayed in Figure 1 with logarithmic axes for λ_1 and λ_2. Several features should be noted. For about 25% of the samples, $\hat{\lambda}_1 = \hat{\lambda}_2$ which causes indeterminacy in π. In many other samples, $\hat{\lambda}_1$ is unusually large; if any single observation y_i is close to zero, high likelihood can be achieved by assigning only that observation to f_1 and allowing $1/\lambda_1$ to approach zero. The strikingly nonnormal shape of the sampling distribution suggests that usual large-sample approximations will be inaccurate.

We investigated the shape of $\log L$ for one sample which gave $\hat{\pi} = 0.32$, $\hat{\lambda}_1 = 2.09$, and $\hat{\lambda}_2 = 0.56$. Figure 2 shows plots of the profile log-likelihood (a) with respect to π and (b) with re-

Downloaded by [Linköping University Library] at 08:08 09 October 2013

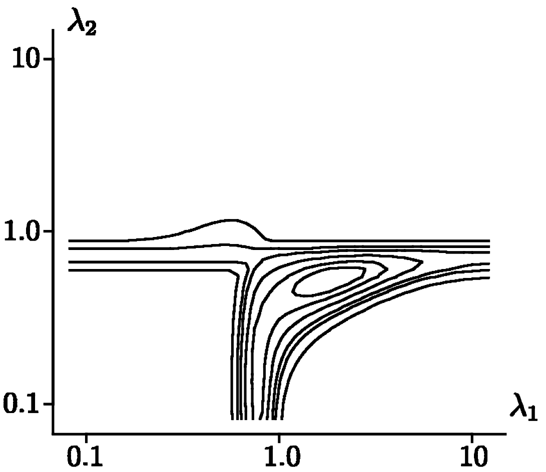


Figure 3. Modified profile log-likelihood for (λ_1, λ_2) after assigning observation 10 to component 2.

spect to (λ_1, λ_2). The profile is the value of the log-likelihood that can be achieved at fixed values for a subset of parameters, maximizing over the others. Many statisticians (ourselves included) prefer to remove nuisance parameters from a likelihood function by integration; we have chosen to display the profile log-likelihood because its contours delineate the approximate confidence regions that would result from inverting a standard likelihood-ratio test. In Figure 2(a), the profile log-likelihood for π has been shifted so that the maximum value is zero, with shading over the region where $\hat{\lambda}_1 > \hat{\lambda}_2$. This profile is bimodal and symmetric about $\pi = 0.5$, a consequence of the labeling issue. For a well-behaved, single-parameter profile, a dropoff in $2 \log L$ equal in magnitude to 3.84—the 95th percentile of χ^2_1—encloses an approximate 95% confidence interval for the parameter. In this problem, as we move to the right of $\hat{\pi} = 0.32$, the profile drops off only slightly before it starts to rise again as we approach the alternate mode. Any reasonable confidence region for π will cross $\pi = 0.5$, at which point the ordering of the $\hat{\lambda}_j$'s is reversed and the interpretation of π becomes unclear. Similar difficulties arise for the λ_j's. Figure 2(b) shows contours of the joint profile for λ_1 and λ_2, with shading over the region where $\hat{\pi} < 0.5$. The contour lines in this plot correspond to the "approximate" 25, 50, 75, 90, 95, and 99% confidence regions as determined by χ^2_2. As we move away from either mode, the regions of high confidence extend deep into areas of the parameter space where the λ_j's are apparently reversed and π has become $1 - \pi$.

Some have addressed this issue by imposing restrictions on the parameter space to enforce a unique labeling (Aitkin and Rubin 1985). In our example, one could require $\pi < 0.5$ or $\lambda_1 > \lambda_2$. Either constraint will get rid of the unwanted second mode without changing the shape of the likelihood at the mode that remains. A similar result, however, can be achieved in a very different way. Consider the sample that we used to generate the profile likelihood plots in Figure 2. Evaluating conditional probabilities (3) at the first mode, we found that the largest observation in this sample, $y_{10} = 7.72$, is assigned to component 2 with probability $1 - \hat{\delta}_{10} = 0.9999873$. Suppose that we simply assign this observation to component 2 with certainty. This is equivalent to introducing the additional factor $(1 - \delta_{10})/(1 - \pi)$ into the likelihood, so that the likelihood changes to

$$\begin{aligned} L^*_{(10,2)} &= \prod_{i \neq 10} \{\pi\lambda_1 \exp(-\lambda_1 y_i) + (1-\pi)\lambda_2 \exp(-\lambda_2 y_i)\} \\ &\quad \times \lambda_2 \exp(-\lambda_2 y_{10}) \\ &= L \times \frac{(1-\delta_{10})}{(1-\pi)}, \end{aligned}$$

where δ_{10} is the true value of (3) for y_{10}. We can maximize $L^*_{(10,2)}$ by a trivial modification of EM, setting $\hat{\delta}_{10} = 0$ at each M-step (4). A plot of the joint profile $2 \log L^*_{(10,2)}$ for λ_1 and λ_2 is shown in Figure 3. The first mode remains essentially unchanged at $(0.32, 2.09, 0.56)$, but the likelihood at the other mode has been sharply reduced to a point where $(0.68, 0.56, 2.09)$ is no longer a plausible estimate.

From an ML perspective, the log-likelihood shown in Figure 3 is essentially no different from that of Figure 2 in the vicinity of the major mode. When viewed as a log-posterior density, however, the new function is easier to summarize; the nuisance mode has very little probability content, suggesting that mode switching during simulation will no longer be a problem. We will explore this possibility in greater detail after reviewing Bayesian methods in the next section.

3. BAYESIAN METHODS

Bayesian analysis by Markov chain Monte Carlo (MCMC) has found widespread application over the last decade. The most popular MCMC method for finite mixtures is closely related to EM; it may be viewed either as a Gibbs sampler (Gelfand and Smith 1990) or a variant of data augmentation (Tanner and Wong 1987). For the mixture of two exponentials, it is convenient to assign independent beta and gamma prior distributions $\pi \sim \mathrm{B}(\alpha_1, \alpha_2)$, $\lambda_1 \sim \mathrm{G}(\beta_1, \beta_2)$, and $\lambda_2 \sim \mathrm{G}(\gamma_1, \gamma_2)$, whose densities are proportional to $\pi^{\alpha_1 - 1}(1-\pi)^{\alpha_2 - 1}$, $\lambda_1^{\beta_1 - 1} \exp(-\beta_2 \lambda_1)$, and $\lambda_2^{\gamma_1 - 1} \exp(-\gamma_2 \lambda_2)$, respectively. Let z_i be a latent indicator equal to one if y_i belongs to f_1 and zero if it belongs to f_2. For the first step, given provisional guesses $(\pi^*, \lambda_1^*, \lambda_2^*)$ for the parameters, we calculate

$$\delta_i^* = \frac{\pi^* \lambda_1^* \exp(-\lambda_1^* y_i)}{\pi^* \lambda_1^* \exp(-\lambda_1^* y_i) + (1-\pi^*)\lambda_2^* \exp(-\lambda_2^* y_i)},$$

and draw $z_i \sim \mathrm{Bernoulli}(\delta_i^*)$ independently for $i = 1, \ldots, n$. For the second step, we draw new random values for the parameters from

$$\begin{aligned} \pi^* &\sim \mathrm{B}(\alpha_1 + n_1, \alpha_2 + n_2), \\ \lambda_1^* &\sim \mathrm{G}(\beta_1 + n_1, \beta_2 + w_1), \\ \lambda_2^* &\sim \mathrm{G}(\gamma_1 + n_2, \gamma_2 + w_2), \end{aligned}$$

where $n_1 = \sum_i z_i$, $n_2 = n - n_1$, $w_1 = \sum_i z_i y_i$, and $w_2 = \sum_i y_i - w_1$. Repeating these two steps produces a sequence of $(\pi^*, \lambda_1^*, \lambda_2^*)$ values from a discrete-time, continuous-state-space Markov chain whose stationary distribution is the joint posterior for $(\pi^*, \lambda_1^*, \lambda_2^*)$ given the data.

From a purely computational standpoint, simulating draws from the posterior distribution by MCMC is no more difficult than ML estimation by EM. With MCMC, however, many new issues arise. Special attention must be given to monitoring convergence (Cowles and Carlin 1996; Brooks 1998) and prior distributions must be chosen with care. Perhaps the most troubling aspect of MCMC with finite mixtures is dubious interpretation of long-run average of the output stream because the component labels may switch during the MCMC run. This label-switching phenomenon is evident in many applications, particularly with smaller samples.

Several solutions to the label-switching problem have been proposed. Richardson and Green (1997) imposed constraints on the parameter space, eliminating $k! - 1$ redundant regions so that the labels may no longer be permuted. Possible constraints for the mixture of two exponentials include either $\pi < 0.5$ or $\lambda_1 < \lambda_2$. If the prior distribution is symmetric, then constraints may be applied after the MCMC run, relabeling the simulated parameters from each iteration as needed. Alternatively, the constraint could be applied within the P-step by rejecting and redrawing any parameters that fail to satisfy it; the resulting target posterior will be the same. Richardson and Green (1997) demonstrated that changing the constraints may greatly alter the shape of posterior distribution of the parameters, and they advise that MCMC output be postprocessed by a variety of different rules. Stephens (1997) noted that constraints may fail to eliminate label switching, particularly when the true values of the parameters lie near a constraint.

Another approach, proposed by Stephens (1997) and Celeux (1998), is to run MCMC without constraints, and then choose a permutation of the component labels at each iteration so that the relabeled parameters within each component lie closest to the centroids of the previous iterates. Cluster-based algorithms require examination of all $k!$ possible labelings at each iteration; this is not burdensome for small values of k but becomes impractical as the number of components grows. These algorithms also require training samples that are free of label switching to provide initial estimates for the centroids. For the mixture of two exponentials, Celeux (1998) recommended an initial run of 100 cycles before relabeling.

4. DATA-DEPENDENT PRIORS

A different strategy for the label-switching problem is to identify the components through an informative prior distribution. Accurate prior knowledge about the parameters of the component densities may drastically reduce the incidence of label switching and improve the properties of the resulting estimates. A major problem, of course, is that such knowledge may be unavailable or difficult to quantify. To be effective, a prior distribution must apply enough information to break the symmetry of the likelihood and dampen the posterior density over $k! - 1$ nuisance regions, yet remain diffuse enough to let the data speak for themselves.

A simple and effective solution for our example of a mixture of two exponentials might be to assign one or more observations to individual components with certainty. Recall that, for our sample of $n = 100$ observations, assigning the largest value y_{10} to the second component drastically reduced the likelihood at the nuisance mode. Suppose we alter our MCMC algorithm in a similar fashion, deterministically assigning y_{10} to class 2 by setting $z_{10} = 0$ at every cycle. This is equivalent to changing the posterior density to

$$\begin{aligned}\Pr^*(\boldsymbol{\theta} \mid y)_{(10,2)} &\propto \prod_{i \neq 10} f(y_i; \boldsymbol{\theta}) \times f_2(y_{10}; \lambda_2) \times \Pr(\boldsymbol{\theta}) \\ &\propto L \times \frac{1 - \delta_{10}}{1 - \pi} \times \Pr(\boldsymbol{\theta}),\end{aligned}$$

where $\Pr(\boldsymbol{\theta})$ is the original prior distribution for $\boldsymbol{\theta} = (\pi, \lambda_1, \lambda_2)$ described in Section 3. This simple modification may be viewed simply as identifying the model by defining the labels in the augmented-data likelihood (the joint distribution of the observed data and the missing component indicator labels).

From one standpoint, assigning y_{10} to class 2 seems innocuous; y_{10} must logically belong to one of the two components, so this merely identifies component 2 as the one to which y_{10} belongs. It is somewhat unsettling, however, that any one of the observations $y_1, \ldots, y_n$ could have been assigned to component 1 or component 2 in this manner, producing $2n$ different modified posterior likelihoods $\Pr^*(\boldsymbol{\theta} \mid y)_{(i,j)}$ whose shapes radically differ. The strategic choice of assigning the largest y_i to component 2—the observation that could be classified with the greatest certainty—produces the greatest impact in dampening the nuisance mode; choosing an observation near the sample median does little to break the symmetry.

If we assign two cases to components, the situation is different. We might, for example, assign the largest observation to component 2 and the smallest to component 1. This has the desired effect of identifying the components, however it also stipulates that these two observations belong in different components, which is a stronger assumption than when only one observation was classified. Nevertheless, such prior assumptions may be valuable in delivering improved posterior inference. As a strategy to address the label switching problem, preclassifying one or more cases requires only a trivial modification of the MCMC algorithm, and it does not require any monitoring of constraints or other post-processing techniques.

5. SIMULATION

This section formally compares the performance of ML and Bayesian methods over repeated samples. Our purpose is to investigate whether the Bayesian methods have better frequentist properties than ML for this example, and whether label switching can be effectively controlled using our preclassification technique. For ML, we performed the EM algorithm and generated bootstrap samples to compute confidence intervals. For Bayesian methods, we applied the relatively diffuse priors $\pi \sim \mathrm{B}(1, 1)$, $\lambda_1 \sim \mathrm{G}(0.5, 0.5)$, $\lambda_2 \sim \mathrm{G}(0.5, 0.5)$. Three methods for handling the label-switching problem were applied: imposing constraints on the parameter space, clustering parameter draws, and assigning one observation to each component.

In our simulation, we drew 1,000 samples of $n = 100$ observations each from the mixture of two exponentials with parameters $(\pi, \lambda_1, \lambda_2) = (0.5, 1.0, 0.5)$. For each sample, we calculated estimates and nominal 95% intervals for π, λ_1, and λ_2 by the following seven methods.

EM-HESS: Run EM until convergence, relabeling the classes if necessary to ensure that $\hat{\lambda}_1 \geq \hat{\lambda}_2$. At the mode, calculate and invert the Hessian of the loglikelihood to obtain normal-theory intervals for π, λ_1 and λ_2. (The Hessian

Downloaded by [Linköping University Library] at 08:08 09 October 2013

Downloaded by [Linköping University Library] at 08:08 09 October 2013

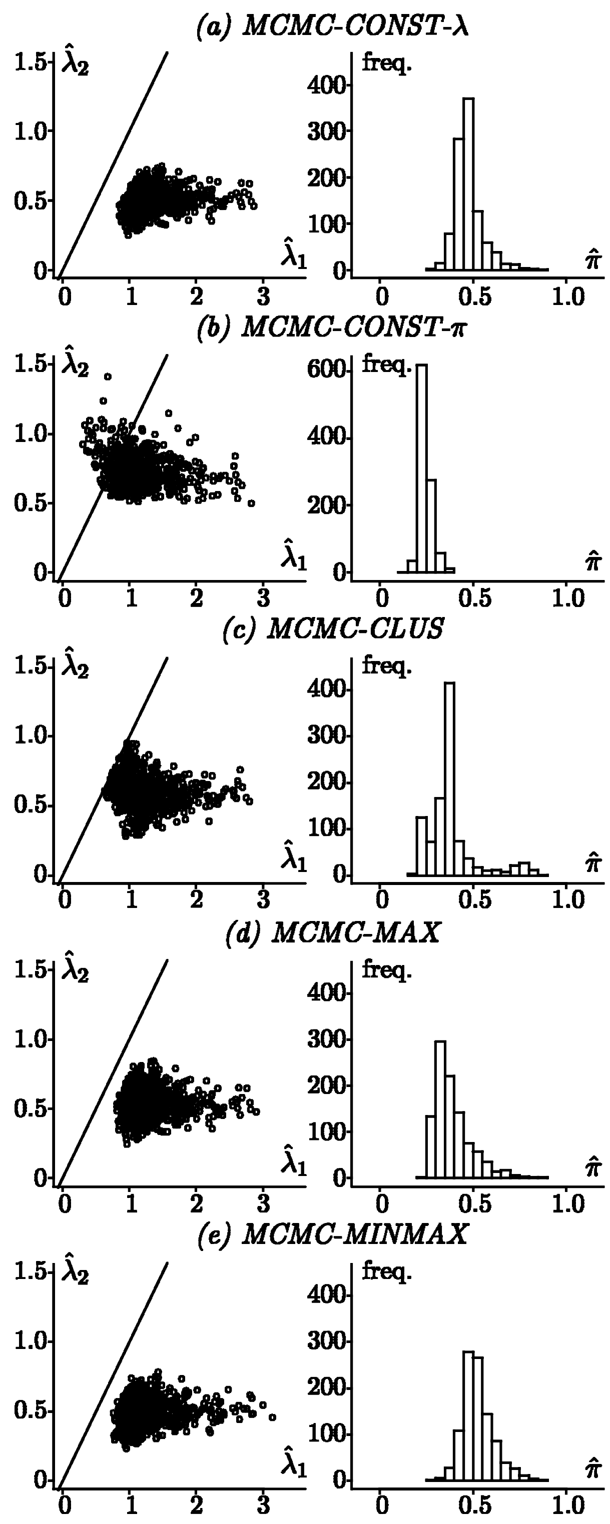


Figure 4. Distribution of point estimates for (λ_1, λ_2) and π over 1,000 samples obtained by five methods: (a) MCMC-CONST-λ; (b) MCMC-CONST-π; (c) MCMC-CLUS; (d) MCMC-MAX; and (e) MCMC-MINMAX.

could not be inverted for 246 samples having $\hat{\lambda}_1 = \hat{\lambda}_2$; no intervals were calculated for these.)

EM-BOOT: Run EM and relabel to ensure $\hat{\lambda}_1 \geq \hat{\lambda}_2$. Generate 1,000 bootstrap resamples and run EM on each one; use the 2.5th and 97.5th percentiles of the bootstrap distributions as interval endpoints.

MCMC-CONST-λ: Run MCMC for a long burn-in period plus 100,000 iterations, relabeling the output to satisfy the constraint $\lambda_1 \geq \lambda_2$. Discard the burn-in and average the series to obtain estimates; use the 2.5th and 97.5th percentiles as interval endpoints.

MCMC-CONST-π: Like MCMC-CONST-λ, but apply the constraint $\pi \leq 0.5$.

MCMC-CLUS: Like the two previous methods, but relabel the output by the method of Celeux (1998). Following Celeux's recommendation, use an initial training sample of 100 iterations to estimate the centroids.

MCMC-MAX: Assign the largest sample observation to class 2, and run MCMC without constraints.

MCMC-MINMAX: Assign the largest sample observation to class 2 and the smallest sample observation to class 1, and run MCMC without constraints.

The distribution of the ML estimates from EM over the 1,000 samples have already been shown in Figure 1(a)–(b); estimates from the five MCMC procedures are shown in Figure 4. Comparing the estimates of λ_1 and λ_2 in Figure 4 to those of Figure 1(a), we notice a dramatic difference in scale. Unlike ML, MCMC does not produce wildly extreme estimates for the λ_j's. Differences among the MCMC methods are apparent, but as far as the λ_j's are concerned, any of them represents an improvement over ML. The estimates for π have also become less variable, and a downward bias is now apparent with MCMC-CONST-π, MCMC-CLUS, and MCMC-MAX.

The average and root-mean squared error (RMSE) of the parameter estimates under each estimation method are shown in Table 1. MCMC-CONST-λ performs well, but MCMC-CONST-π does poorly. Under the latter, π is severely underestimated and λ_2 is overestimated; the labels for λ_1 and λ_2 are still switching. This example illustrates both the power and danger of constraints. Constraints are most effective when the true parameter values lie well in the interior of the constrained parameter space. When a parameter lies near a boundary imposed by a constraint, however, label switching may still occur, and estimates for that parameter may be badly biased.

In this example, Celeux's clustering method (MCMC-CLUS) does not perform as well as MCMC-CONST-λ. The problem with Celeux's method stems from occasional label switching in the training sample. When switching occurs, the initial centroids lie too close together, making the whole relabeling process unreliable. When using this method in a real application, one should

Table 1. Average (RMSE) of Point Estimates Over 1,000 Repetitions

		MCMC				
	ML†	*CONST-λ*	*CONST-π*	*CLUS*	*MAX*	*MINMAX*
λ_1	3.968 (47.34)	1.311 (0.447)	1.101 (0.355)	1.221 (0.405)	1.271 (0.427)	1.259 (0.437)
λ_2	0.516 (0.162)	0.510 (0.077)	0.720 (0.244)	0.600 (0.144)	0.549 (0.106)	0.494 (0.086)
π	0.485 (0.252)	0.477 (0.076)	0.243 (0.259)	0.380 (0.176)	0.391 (0.147)	0.525 (0.087)

† Used in EM-HESS and EM-BOOT

Table 2. Percent Coverage (average width) of Nominal 95% Interval Estimates Over 1,000 Repetitions

	EM		*MCMC*				
	HESS	*BOOT*	*CONST-λ*	*CONST-π*	*CLUS*	*MAX*	*MINMAX*
λ_1	93.10 (15.43)	66.10 (101.7)	99.50 (3.005)	100.0 (3.465)	99.50 (3.308)	99.60 (3.314)	99.50 (2.755)
λ_2	92.04 (0.782)	67.40 (0.393)	99.60 (0.558)	79.80 (0.758)	95.00 (0.711)	99.60 (0.548)	99.50 (0.557)
π	72.02 (1.830)	86.30 (0.531)	100.0 (0.928)	0.00 (0.467)	91.20 (0.812)	100.0 (0.877)	99.80 (0.838)

carefully examine the initial sample, discarding it and drawing a new one if label switching has occurred. If we had done this for each of our 1,000 repetitions, the performance of Celeux's method would have more closely resembled that of MCMC-CONST-λ.

Based on Table 1, we see that MCMC-MAX performs slightly better than MCMC-CONST-λ for λ_1 and slightly worse for λ_2, but it has a downward bias for π. A sample of $n = 100$ observations from this population contains so little information about π that the assignment of the largest observation to class 2 has a substantial impact. MCMC-MINMAX balances out the estimate for π and smoothes it toward 0.5; it performs well for this population, but will not necessarily improve matters if the true value of π is far from 0.5. For estimation of λ_1, λ_2, and π, one could argue that MCMC-CONST-λ performs best. Looking at Figure 4(a), however, we see that it tends to push the estimates away from the $\lambda_1 = \lambda_2$ boundary more than the other methods do. By removing the entire region $\lambda_2 > \lambda_1$ from the parameter space, this method tends to exaggerate differences between the components, biasing estimates for parameters such as $\lambda_1 - \lambda_2$ and λ_1/λ_2. From the standpoint of parameter estimation, there is not a clear winner among MCMC-CONST-λ, MCMC-MAX, or MCMC-MINMAX, but any of these methods are sensible and perform far better than ML.

The performance of interval estimates is summarized in Table 2, which shows the percentage of intervals that covered their targets and the average interval width. Narrow intervals are desirable provided that coverage remains at or above the nominal rate of 95%. EM-HESS has reasonably good coverage for the λ_j's, but the intervals are wide. The EM-HESS intervals for π are extremely wide, often straying outside of the parameter space, yet their coverage is poor. Suitable transformations (e.g., logit for π and log for the λ_j's) would keep the interval estimates inside the parameter space, but the flatness of the log-likelihood function creates difficulty on any scale. For this example, the simple bootstrap (EM-BOOT) is a disaster. Bootstrap intervals for λ_1 are highly unstable because the resampling distribution of the ML estimate suffers from the extreme skewness found in the actual sampling distribution.

Among the Bayesian methods, MCMC-CONST-λ, MCMC-MAX, and MCMC-MINMAX are conservative, exhibiting higher-than-nominal rates of coverage for all parameters. Under these three, the intervals for λ_1 and λ_2 are shorter than the likelihood-based intervals, yet their rates of coverage are higher. Their intervals for π nearly cover the range $[0, 1]$, suggesting that very little information about π can be gleaned from a sample of $n = 100$ from this population. The conservative nature of these intervals indicates that they can be improved upon, but there is certainly no false sense of precision associated with them.

6. DISCUSSION

Mixture models are an attractive tool for many areas of substantive research. However, they can be difficult to estimate and may present severe challenges for population inference. We have illustrated some of these challenges with a relatively simple example, a mixture of two exponential distributions. Although this example is certainly not representative of all finite mixtures, it nevertheless lends several important insights which we believe have general relevance.

First, it seems clear that likelihood methods should not be expected to perform well with smaller samples and component densities that substantially overlap. In these situations, the likelihood surface is often so irregular that standard asymptotics are hopelessly inaccurate. We also doubt the usefulness of the bootstrap in these settings, because the repeated-sampling behavior of the ML estimate is so erratic.

Second, our simulations have shown that good inferences are indeed possible through a Bayesian approach. Posterior means and credible sets perform surprisingly well in our example, provided that the MCMC label-switching problem is addressed.

Third, in Bayesian analyses of finite mixtures, we believe that our suggestion to preclassify one or more observations provides a simple and effective technique to reduce or eliminate label-switching. It requires only a minor modification to the MCMC algorithm, and yet is at least as effective as imposing constraints as demonstrated in our simulations. Both Chung (2003) and Loken (in press) have successfully applied the technique in latent class analysis (Goodman 1974), and generalizations to other mixture models seem promising.

We expect that our method may generalize better than the method of constraints, because it is simpler to assign a few observations to classes than to devise a set of constraints that will effectively eliminate label switching over a high-dimensional parameter space. Consider extending our example to a mixture of three exponentials. Imposing the constraints that $\lambda_1 > \lambda_2 > \lambda_3$ identifies a unique mode, but it may present some problems in MCMC. Judging from the simulation results in Figure 4, constraints on the λ may exaggerate differences among the parameters, and may still allow label switching in the π parameters. By contrast, we suspect that assigning the largest observation to component 1, and the smallest to component 3 might actu-

Downloaded by [Linköping University Library] at 08:08 09 October 2013

ally perform well in a three-component mixture of exponentials, although we still need to verify this.

Future work should also explore the degree of prior information implicit in preclassification. Classifying more than one observation to different classes implies a stronger prior, making a priori assumptions that the cases do not belong to the same component. As we have shown, however, a data dependent prior may be an efficient solution to the label switching problem, and thus facilitate Bayesian inference in finite-mixture models.

[Received November 2003. Revised February 2004.]

REFERENCES

Aitkin, M., and Rubin, D. B. (1985), "Estimation and Hypothesis Testing in Finite Mixture Models," *Journal of the Royal Statistical Society*, Series B, 47, 67–75.

Brooks, S. P. (1998), "Markov Chain Monte Carlo Method and its Application," *The Statistician*, 44, 69–100.

Celeux, G. (1998), "Bayesian Inference for Mixture: The Label-Switching Problem," in *Compstat98*, eds. R. Payne and P. Green, Heidelberg: Physica, pp. 227–232.

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.

Chung, H. (2003), "Latent-Class Modeling with Covariates," doctoral dissertation, The Pennsylvania State University, University Park.

Cowles, M. K., and Carlin, B. P. (1996), "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review," *Journal of the American Statistical Association*, 91, 883–904.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood Estimation From Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Series B, 39, 1–38.

Goodman, L. A. (1974), "Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models," *Biometrika*, 61, 215–231.

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

Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis with Missing Data*, New York: Wiley.

Loken, E. (in press), "Multi-Modality in Mixture Models and Latent Trait Models," in *Missing Data and Bayesian Methods in Practice: Contributions by Donald Rubin's Statistical Family*, eds. A. Gelman and X. Meng, New York: Wiley.

McLachlan, G., and Krishnan, T. (1997), *The EM Algorithm and Expectations*, New York: Wiley.

McLachlan, G., and Peel, D. (2000), *Finite Mixture Models*, New York: Wiley.

Richardson, S., and Green, P. J. (1997), "On Bayesian Analysis of Mixtures With an Unknown Number of Components," *Journal of the Royal Statistical Society*, Series B, 59, 731–792.

Robert, C. P. (1996), "Mixtures of Distributions: Inference and Estimation," in *Markov Chain Monte Carlo in Practice*, eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, London: Chapman & Hall, pp. 441–464.

Seidel, W., Mosler, K., and Alker, M. (2000), "Likelihood Ratio Tests Based on Subglobal Optimization: A Power Comparison in Exponential Mixture Models," *Statistical Papers*, 41, 85–98.

Stephens, M. (1997), "Bayesian Methods for Mixtures of Normal Distributions," Ph.D. dissertation, Oxford: Magdalen College.

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

Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation" (with discussion), *Journal of the American Statistical Association*, 82, 528–550.

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

Downloaded by [Linköping University Library] at 08:08 09 October 2013