Statistics and Computing **14**: 323–332, 2004
© 2004 Kluwer Academic Publishers. *Manufactured in The Netherlands.*

An empirical comparison of EM, SEM and MCMC performance for problematic Gaussian mixture likelihoods

JOSÉ G. DIAS* and MICHEL WEDEL†

**Department of Quantitative Methods, Instituto Superior de Ciências do Trabalho e da Empresa—ISCTE, Av. das Forças Armadas, Lisboa 1649–026, Portugal*
jose.dias@iscte.pt
†*The University of Michigan Business School, 701 Tappan Street, MI 48109 Ann Arbor, USA*

Received October 2002 and accepted April 2004

We compare EM, SEM, and MCMC algorithms to estimate the parameters of the Gaussian mixture model. We focus on problems in estimation arising from the likelihood function having a sharp ridge or saddle points. We use both synthetic and empirical data with those features. The comparison includes Bayesian approaches with different prior specifications and various procedures to deal with label switching. Although the solutions provided by these stochastic algorithms are more often degenerate, we conclude that SEM and MCMC may display faster convergence and improve the ability to locate the global maximum of the likelihood function.

Keywords: Gaussian mixture models, EM algorithm, SEM algorithm, MCMC, label switching, loss functions, conjugate prior, hierarchical prior

1. Introduction: Algorithms, model, data

The EM (Expectation-Maximization) algorithm, proposed by Dempster, Laird and Rubin (1977), has become popular to obtain maximum likelihood estimates (MLE), in particular for finite mixture distributions (McLachlan and Peel 2000). However, it suffers from slow convergence and may converge to local maxima or saddle points. Stochastic EM (SEM) and Markov chain Monte Carlo (MCMC) estimation procedures are viable alternatives, but may pose problems of label switching. This paper compares EM, SEM, and MCMC algorithms. Although some comparisons of these approaches are available (e.g., Sahu and Roberts (1999) compare EM and MCMC and Celeux *et al.* (1996) compare EM and SEM), our study examines the behavior of all of them simultaneously for likelihood surfaces having a sharp ridge or saddle point. We use synthetic and empirical data where the likelihood has this particular shape. The comparison includes Bayesian approaches with different prior specifications and different procedures to deal with label switching.

The Gaussian mixture model is formulated as follows. Let $\mathbf{y} = (y_1, \ldots, y_n)$ denote a sample of size n. Each data point is assumed to be a realization of the random variable Y with k-component mixture probability density function (p.d.f.) $f(y_i;\boldsymbol{\varphi}) = \sum_{j=1}^{k} \pi_j f_j(y_i;\boldsymbol{\theta}_j)$, where the mixing proportions π_j are nonnegative and sum to one, $\boldsymbol{\theta}_j = (\mu_j, \sigma_j^2)$ denotes the parameters of the conditional univariate Gaussian distribution of component j defined by $f_j(y_i;\boldsymbol{\theta}_j)$, and $\boldsymbol{\varphi} = \{\pi_1, \ldots, \pi_{k-1}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k\}$. In this paper, we focus on the case where k is fixed. Note that $\pi_k = 1 - \sum_{j=1}^{k-1} \pi_j$. The log-likelihood function is $\ell(\boldsymbol{\varphi};\mathbf{y}) = \sum_{i=1}^{n} \log f(y_i;\boldsymbol{\varphi})$. The likelihood of the finite mixture model is invariant under permutations of the k components. This is not a problem for a deterministic algorithm such as the EM, but it complicates the inference from sampling procedures such as MCMC, because the labels of components may be randomly switched during the iterative process.

Two datasets, one synthetic and one empirical, are used to illustrate that the shape of the log-likelihood surface of the finite mixture presents specific problems to the estimation procedures. For both datasets the log-likelihood function displays ridges or saddle points. The first dataset—a synthetic dataset with $n = 150$—is generated from a Gaussian mixture of three components ($k = 3$) defined by $\boldsymbol{\pi} = (1/3, 1/3, 1/3)$, $\boldsymbol{\mu} = (2, -1, 0)$, and $\boldsymbol{\sigma}^2 = (0.3, 0.4, 15.0)$. The second dataset is the GDP per capita (PPP US$) in 1998 from 174 countries (UNDP,

0960-3174 © 2004 Kluwer Academic Publishers

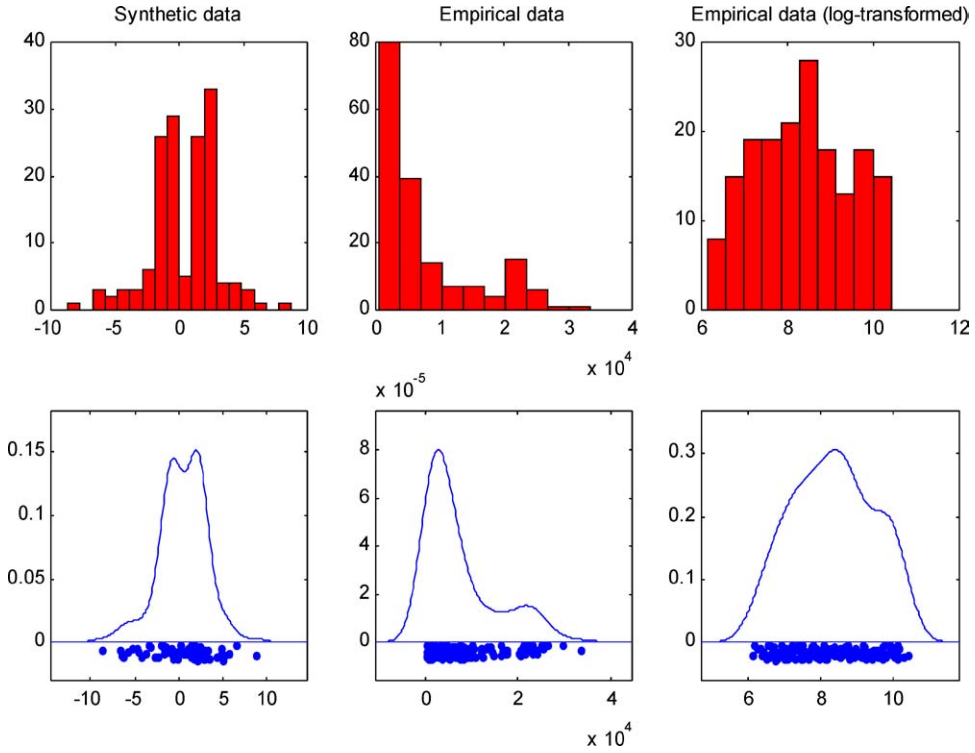

Fig. 1. *Histograms and Gaussian kernel density estimates of the synthetic and empirical data*

2000). Figure 1 presents the histograms and Gaussian kernel density estimates. From Fig. 1, we observe a strongly skewed distribution for the GDP data. As in the literature on income distribution, we log-transformed this dataset. Visual inspection of Fig. 1 suggests a two-component Gaussian mixture model ($k = 2$) for the transformed data.

Section 2 briefly introduces the algorithms. For the MCMC approach, different specifications of the priors and label-switching strategies are provided. Section 3 presents results for the synthetic and empirical datasets. Concluding remarks are made in Section 4.

2. The EM, SEM, and MCMC algorithms

2.1. *Introduction*

The algorithms that we investigate are based on data augmentation, i.e., the observed data ($\mathbf{y}$) is expanded to a new space ($\mathbf{y}$, $\mathbf{z}$), which includes the missing data ($\mathbf{z}$) (Dempster, Laird and Rubin 1977, Tanner and Wong 1987). The missing datum (z_{ij}) indicates whether component j generated observation i. The EM algorithm cycles between computing the expectation of $\mathbf{z}$ and maximization to obtain the MLE of φ ($\hat{\varphi}$). The SEM algorithm cycles between imputing $\mathbf{z}$ by drawing from its predictive distribution, and maximization to obtain $\hat{\varphi}$. The MCMC approach (or, its special case that involves data augmentation, the Gibbs sampler) iterates between simulating from the conditional distributions of $\mathbf{z}$ and φ. In the case of finite mixtures of Gaussian distributions, each of the algorithms takes a simple form with closed form solutions for the steps and/or draws from standard distributions.

2.2. *EM*

The EM algorithm iterates between the E-step, in which the expectation $\mathrm{E}(\mathbf{Z}^{(m)} \,|\, \mathbf{y}, \varphi^{(m)})$ is computed, and the M-step, in which the complete data log-likelihood $\ell(\varphi^{(m+1)}; \mathbf{y}, \mathbf{z}^{(m+1)})$ is maximized to obtain a revised estimate $\varphi^{(m+1)}$. For a description of the EM algorithm for estimating finite mixture models, see, e.g., Dempster, Laird and Rubin (1977), McLachlan and Peel (2000). Since $\ell(\varphi^{(m+1)}; \mathbf{y}) \geq \ell(\varphi^{(m)}; \mathbf{y})$, $m = 0, 1, \ldots,$ under suitable regularity conditions, $\varphi^{(m)}$ converges to a stationary point of $\ell(\varphi; \mathbf{y})$ (see Dempster, Laird and Rubin 1977, Wu 1983, McLachlan and Krishnan 1997). However, the EM algorithm has been observed to converge extremely slowly, because its convergence rate, governed by the fraction of missing information, is linear. In addition, it may converge to a local maximum or saddle point—although that problem can be handled by randomly perturbing the solution away from the saddle point. Therefore, the criterion of convergence of the EM algorithm is of particular interest. We are interested in three definitions of numerical convergence based on the log-likelihood function (Fletcher 1980, McLachlan and Peel 2000): 1. the absolute difference $|\ell(\varphi^{(m+1)}; \mathbf{y}) - \ell(\varphi^{(m)}; \mathbf{y})| \leq \varepsilon$; 2. the relative difference $|\ell(\varphi^{(m+1)}; \mathbf{y})/\ell(\varphi^{(m)}; \mathbf{y}) - 1| \leq \varepsilon$; and 3. the Aitken's absolute difference (McLachlan and Peel 2000)

$|\ell_A(\boldsymbol{\varphi}^{(m+1)}; \mathbf{y}) - \ell_A(\boldsymbol{\varphi}^{(m)}; \mathbf{y})| \leq \varepsilon$, where

$$\ell_A\big(\boldsymbol{\varphi}^{(m+1)}; \mathbf{y}\big) = \ell\big(\boldsymbol{\varphi}^{(m)}; \mathbf{y}\big) + \frac{\big[\ell\big(\boldsymbol{\varphi}^{(m+1)}; \mathbf{y}\big) - \ell\big(\boldsymbol{\varphi}^{(m)}; \mathbf{y}\big)\big]\big[\ell\big(\boldsymbol{\varphi}^{(m)}; \mathbf{y}\big) - \ell\big(\boldsymbol{\varphi}^{(m-1)}; \mathbf{y}\big)\big]}{\big[\ell\big(\boldsymbol{\varphi}^{(m)}; \mathbf{y}\big) - \ell\big(\boldsymbol{\varphi}^{(m-1)}; \mathbf{y}\big)\big] - \big[\ell\big(\boldsymbol{\varphi}^{(m+1)}; \mathbf{y}\big) - \ell\big(\boldsymbol{\varphi}^{(m)}; \mathbf{y}\big)\big]},$$

for some small value of the tolerance ε. For example, Wedel and DeSarbo (1995) use the absolute criterion and Vlassis and Likas (2002) use the relative criterion, where typically ε is set to values in the range of 10^{-4}–10^{-6}. In addition we investigate the influence of starting values. Two strategies are compared: starting with random centers (RC) based on McLachlan and Peel (2000, p. 55), or with a random partition of the data (RP), in which starting values for the posterior probabilities (α_{ij}) are drawn from the uniform distribution, scaled to sum to one.

2.3. *SEM*

The Stochastic EM (SEM) algorithm (Celeux and Diebolt 1985, Diebolt and Ip 1996) incorporates a stochastic step (S-step) which simulates a realization $\mathbf{z}^{(m)}$ of the missing data from its predictive density $p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\varphi}^{(m)})$ based on the current estimate $\boldsymbol{\varphi}^{(m)}$, which is then updated by maximizing the log-likelihood function of the complete data set $\mathbf{x}^{(m+1)} = (\mathbf{y}, \mathbf{z}^{(m+1)})$. For computational details see, e.g., Diebolt and Ip (1996). Convergence is assessed from plots of the log-likelihood against the iterates (see, e.g., Celeux, Chauveau and Diebolt 1996).

2.4. *MCMC*

In the Gibbs sampler, one iteratively generates the parameters and the missing data from $p(\boldsymbol{\varphi} \mid \mathbf{y}, \mathbf{z})$ and $p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\varphi})$ (Tanner and Wong 1987, Gelfand and Smith 1990, Diebolt and Robert 1994). For the parameters of mixtures of Gaussian distributions, full conditional distributions can usually be derived analytically. For a discussion of convergence criteria, see, e.g., Cowles and Carlin (1996). We will inspect plots of the draws of the parameters against the iterates to assess convergence (see, e.g., Stephens 2000).

2.4.1. Prior specification

We analyze the influence of the priors on the performance of the MCMC algorithm in estimating finite mixtures using three different settings: an independent prior, a conjugate prior, and a hierarchical prior.

Independent priors: Means and variances are assumed a priori independent, as was done by Escobar and West (1995):

$$p(\boldsymbol{\varphi}) = p(\boldsymbol{\pi}) \prod_j p(\mu_j) p\big(\sigma_j^2\big)$$

with $\boldsymbol{\pi} \sim \mathcal{D}(\delta, \delta, \ldots, \delta)$, $\mu_j \sim \mathcal{N}(\xi, \kappa^{-1})$, and $\sigma_j^2 \sim \mathcal{IG}(\alpha, \beta)$, where $\mathcal{D}$ is the Dirichlet distribution, $\mathcal{N}$ is the Gaussian distribution, $\mathcal{IG}$ is the inverse gamma distribution, and δ, ξ, κ, α, and β are constants. This prior is referred to in the sequel as IPRIOR.

Conjugate priors: Diebolt and Robert (1994) and Robert (1996) suggest a conjugate prior of the form:

$$p(\boldsymbol{\varphi}) = p(\boldsymbol{\pi}) \prod_j p\big(\mu_j \mid \sigma_j^2\big) p\big(\sigma_j^2\big)$$

with $\boldsymbol{\pi} \sim \mathcal{D}(\delta, \delta, \ldots, \delta)$, $\mu_j \mid \sigma_j^2 \sim \mathcal{N}(\xi, \sigma_j^2/\lambda)$, $\sigma_j^2 \sim \mathcal{IG}(\alpha, \beta)$, and δ, ξ, λ, α, and β are constants. This prior is referred to in the sequel as CPRIOR.

Hierarchical prior: A third option is the hierarchical structure of $p(\boldsymbol{\varphi})$ proposed by Richardson and Green (1997), where the vector $\boldsymbol{\varphi}$ is augmented with the hyperparameter (β):

$$p(\boldsymbol{\varphi}) = p(\boldsymbol{\pi}) p(\beta) \prod_j p(\mu_j) p\big(\sigma_j^2 \mid \beta\big)$$

with $\boldsymbol{\pi} \sim \mathcal{D}(\delta, \delta, \ldots, \delta)$, $\mu_j \sim \mathcal{N}(\xi, \kappa^{-1})$, $\sigma_j^2 \mid \beta \sim \mathcal{IG}(\alpha, \beta)$, $\beta \sim \Gamma(g, h)$, where $\Gamma(a, b)$ denotes the gamma distribution with mean ab and variance ab^2, β is a hyperparameter, and δ, ξ, κ, α, g, and h are constants. This prior is referred to as HPRIOR.

2.4.2. The label-switching problem

Because the likelihood is invariant under permutation of the k components, if there is no prior information that distinguishes these components, the posterior distribution will have $k!$ symmetric modes. During the MCMC sampling process a permutation of the components may occur, resulting in multimodal marginal distributions of the parameters. If this label switching happens, summary statistics of the marginal distributions will not give accurate estimates (see, e.g., Stephens 1997a). We present and compare different procedures available in the literature to deal with this problem. We refer to the strategy where none of these procedures is applied as NONE.

Identifiability constraints: One approach to minimize the label-switching effect is based on imposing identifiability constraints on the parameters (see, e.g., Dielbolt and Robert 1994, Roeder and Wasserman 1997, Richardson and Green 1997). In the context of univariate Gaussian mixture models such constraints can be one of the following: $\pi_1 < \pi_2 < \cdots < \pi_k$, $\mu_1 < \mu_2 < \cdots < \mu_k$, or $\sigma_1^2 < \sigma_2^2 < \cdots < \sigma_k^2$. These three strategies are referred to as C1, C2, and C3, respectively. However, it has been shown (see, e.g., Stephens 1997a, Celeux, Hurn and Robert 2000) that these constraints may distort the posterior distribution of parameters.

Celeux method: Celeux (1998) suggests post-processing the simulated MCMC chain by minimizing a function of the parameters (see also Celeux, Hurn and Robert 2000). Let $\varphi_s^{(m)}$ be the simulated value of the parameter φ_s at iteration m, with

$s = 1, \ldots, S$, and S the number of parameters across all components of the mixture $\{\pi_j, \mu_j, \sigma_j^2, j = 1, \ldots, k\}$. Let $\varphi_s^{(m)}$, $m = 1, \ldots, m^*$ be the set of initial values used to initialize a K-means-type algorithm. The initial reference center is defined as $\bar{\varphi}_s^{[0]} = \bar{\varphi}_s = \frac{1}{m^*} \sum_{m=1}^{m^*} \varphi_s^{(m)}$. For the mth draw, compute the distance from $\varphi_s^{(m^*+m)}$ for each $u = 1, \ldots, k!$ centers using the normalized squared distance; then, permute labels according to the initial order. This strategy is referred to as CELEUX. For more computational details, see Celeux (1998).

Stephens method: Stephens (1997b, 2000) suggests relabelling based on the minimization of a function of the posterior probabilities $\alpha_{ij}^{(m)} = \pi_j^{(m)} f_j(y_i; \boldsymbol{\theta}_j^{(m)})[\sum_{h=1}^{k} \pi_h^{(m)} f_h(y_i; \boldsymbol{\theta}_h^{(m)})]^{-1}$. It has the advantage of being normalization free, but demanding much computer memory. Therefore, Stephens (2000) proposed an on-line version of his algorithm. Let $\upsilon_m(\boldsymbol{\varphi}^{(m)})$ define a permutation of the parameters at stage m and let $\mathbf{Q}^{(m-1)} = (q_{ij}^{(m-1)})$ be the current estimate of α_{ij}. The algorithm is initialized with a small number of runs, say m^*: $\mathbf{Q}^{(0)} = (\frac{1}{m^*} \sum_{m=1}^{m^*} \alpha_{ij}^{(m)})$. Then, at mth iteration, choose υ_m to minimize the Kullback-Leibler divergence between the posterior probabilities $\alpha_{ij}\{\upsilon_m(\boldsymbol{\varphi}^{(m)})\}$ and the estimate of the posterior probabilities $\mathbf{Q}^{(m-1)}$, and subsequently compute $\mathbf{Q}^{(m)}$. This strategy of dealing with the label switching is referred to as STEPHENS. For computational details, we refer to Stephens (2000).

CHR method: Celeux, Hurn and Robert proposed to handle the label-switching problem by computing the Bayes estimator defined by $\hat{\boldsymbol{\varphi}}^* = \arg\min_{\hat{\varphi}} \mathrm{E}_{\varphi|\mathbf{y}} L(\boldsymbol{\varphi}, \hat{\boldsymbol{\varphi}})$, where $L(\boldsymbol{\varphi}, \hat{\boldsymbol{\varphi}})$ defines a loss function for which the label switching is immaterial (Celeux, Hurn and Robert 2000). We use a global loss function based on a symmetrized Kullback-Leibler distance (see Celeux, Hurn and Robert 2000; Hurn, Justel and Robert 2003) defined by $L(\boldsymbol{\varphi}, \hat{\boldsymbol{\varphi}}) = \int \{f(\mathbf{y}; \boldsymbol{\varphi}) \log \frac{f(\mathbf{y};\boldsymbol{\varphi})}{f(\mathbf{y};\hat{\boldsymbol{\varphi}})} + f(\mathbf{y}; \hat{\boldsymbol{\varphi}}) \log \frac{f(\mathbf{y};\hat{\boldsymbol{\varphi}})}{f(\mathbf{y};\boldsymbol{\varphi})}\}\, d\mathbf{y}$. This loss function possesses attractive properties such as being invariant under reparameterization (Hurn, Justel and Robert 2003). The estimation is based on the two-step approach of Rue (1995). The first step approximates $\mathrm{E}_{\varphi|\mathbf{y}} L(\boldsymbol{\varphi}, \hat{\boldsymbol{\varphi}})$ using MCMC ergodic means for a given $\hat{\boldsymbol{\varphi}}$; the second step computes $\hat{\boldsymbol{\varphi}}$ by minimizing the expected loss. The minimization problem for $\hat{\boldsymbol{\varphi}}$ is addressed using simulated annealing (Rue 1995, Frigessi and Rue 1997). This strategy is referred to as CHR. For computational details, we refer to Celeux, Hurn and Robert (2000).

3. Results

3.1. *Synthetic data*

We ran the *EM algorithm* 1000 times with a three-component solution until iteration $m = 1000$. Figure 2 plots the log-likelihood against the EM iterations. For clarity, only the first 100 runs are shown (RC starting values). From the figure we infer that the log-likelihood has a complex shape. The EM algorithm did not reach the maximum for 24.3% (53.9% for RP starting values) within 250 iterations and in 3.5% (7.9% for RP) of the solutions even after 1000 iterations (absolute convergence criterion with $\varepsilon = 10^{-6}$). The parameter estimates from the best EM solution, corresponding to a log-likelihood value of -328.65, are presented in Table 1, along with those for all other procedures being compared.

To better understand the behavior of the EM algorithm, we provide a representation of the log-likelihood surface. Since the parameter space is eight-dimensional, in Fig. 3 we plot the log-likelihood surface $\ell(\boldsymbol{\varphi}; \mathbf{y})$ in the neighborhood of the EM solution against μ_1 and μ_2, keeping the other parameters constant. Note that the log-likelihood is close to symmetric, since $\hat{\pi}_1 \simeq \hat{\pi}_2$, $\hat{\sigma}_1^2 \simeq \hat{\sigma}_2^2$ and it is invariant under permutations of the component labels. The shape of the log-likelihood explains the behavior of the EM algorithm. Surrounding the maximum there are various saddle regions, and irrespective the starting values are, the EM algorithm almost always visits a saddle region from which it is difficult to escape.

Given the shape of the likelihood, the impact of the stopping rules on the convergence of the EM algorithm is of interest. For each run for each criterion three states are possible: *true convergence*, i.e., the stopping rule holds and the difference between the log-likelihood at that iteration and the "true" maximum value of the log-likelihood (across all runs) is less than $\phi = 0.01$; *false convergence*, i.e.,the stopping rule holds, but the difference with

Table 1. *EM, SEM, and MCMC results for the synthetic data**

	Proportions			Means			Variances		
	1	2	3	1	2	3	1	2	3
EM	0.353 (0.06)	0.330 (0.06)	0.317 (−)	2.084 (0.10)	−0.884 (0.11)	−0.110 (0.61)	0.333 (0.11)	0.399 (0.16)	15.332 (5.69)
SEM	0.353 (0.05)	0.327 (0.06)	0.320 (−)	2.087 (0.10)	−0.872 (0.12)	−0.138 (0.60)	0.334 (0.11)	0.404 (0.20)	15.191 (5.37)
MCMC (NONE)									
IPRIOR	0.346 (0.05)	0.325 (0.06)	0.329 (0.08)	2.091 (0.10)	−0.875 (0.13)	−0.107 (0.62)	0.340 (0.08)	0.433 (0.12)	16.006 (3.78)
CPRIOR	0.342 (0.05)	0.321 (0.06)	0.338 (0.08)	2.095 (0.10)	−0.874 (0.12)	−0.097 (0.60)	0.328 (0.08)	0.413 (0.12)	15.291 (3.42)
HPRIOR	0.343 (0.05)	0.319 (0.05)	0.338 (0.07)	2.091 (0.11)	−0.878 (0.13)	−0.101 (0.59)	0.368 (0.08)	0.447 (0.12)	14.357 (3.02)

*Asymptotic standard errors for the EM and SEM estimates and posterior standard errors for the MCMC estimates, no label-switching strategy applied.

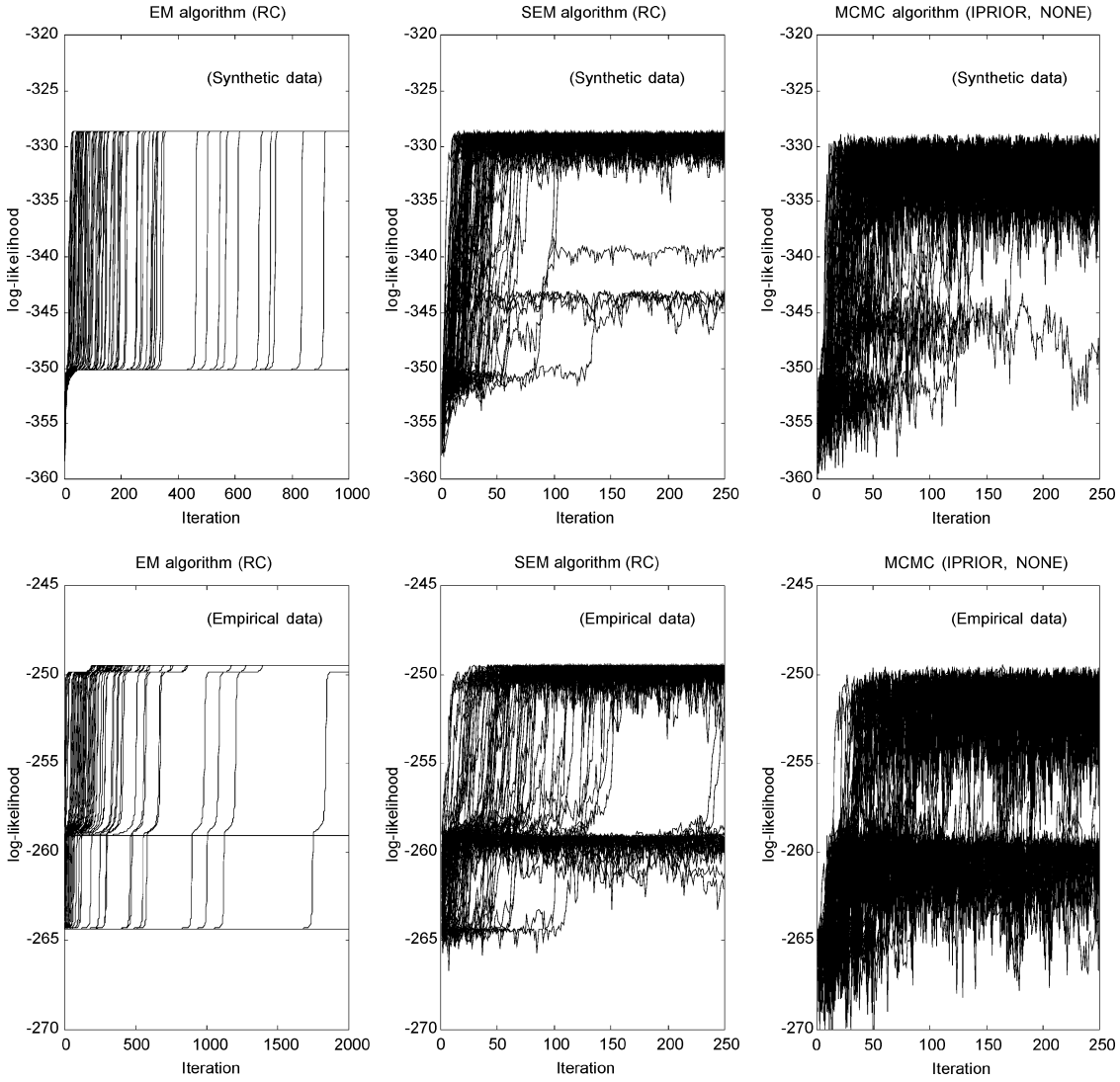

Fig. 2. *EM, SEM, and MCMC iterative process for synthetic and empirical data (100 runs)*

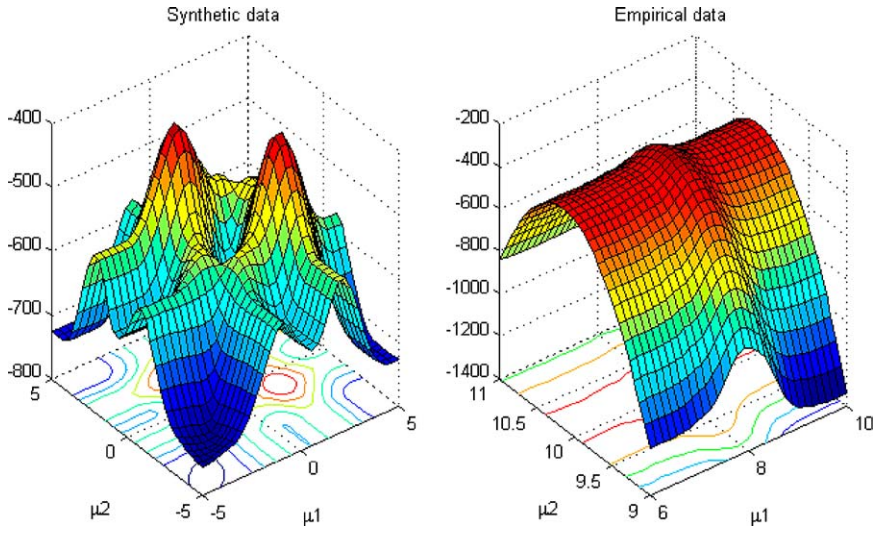

Fig. 3. *Log-likelihood surface against μ_1 and μ_2 in the neighborhood of the EM solution, for the synthetic and empirical data*

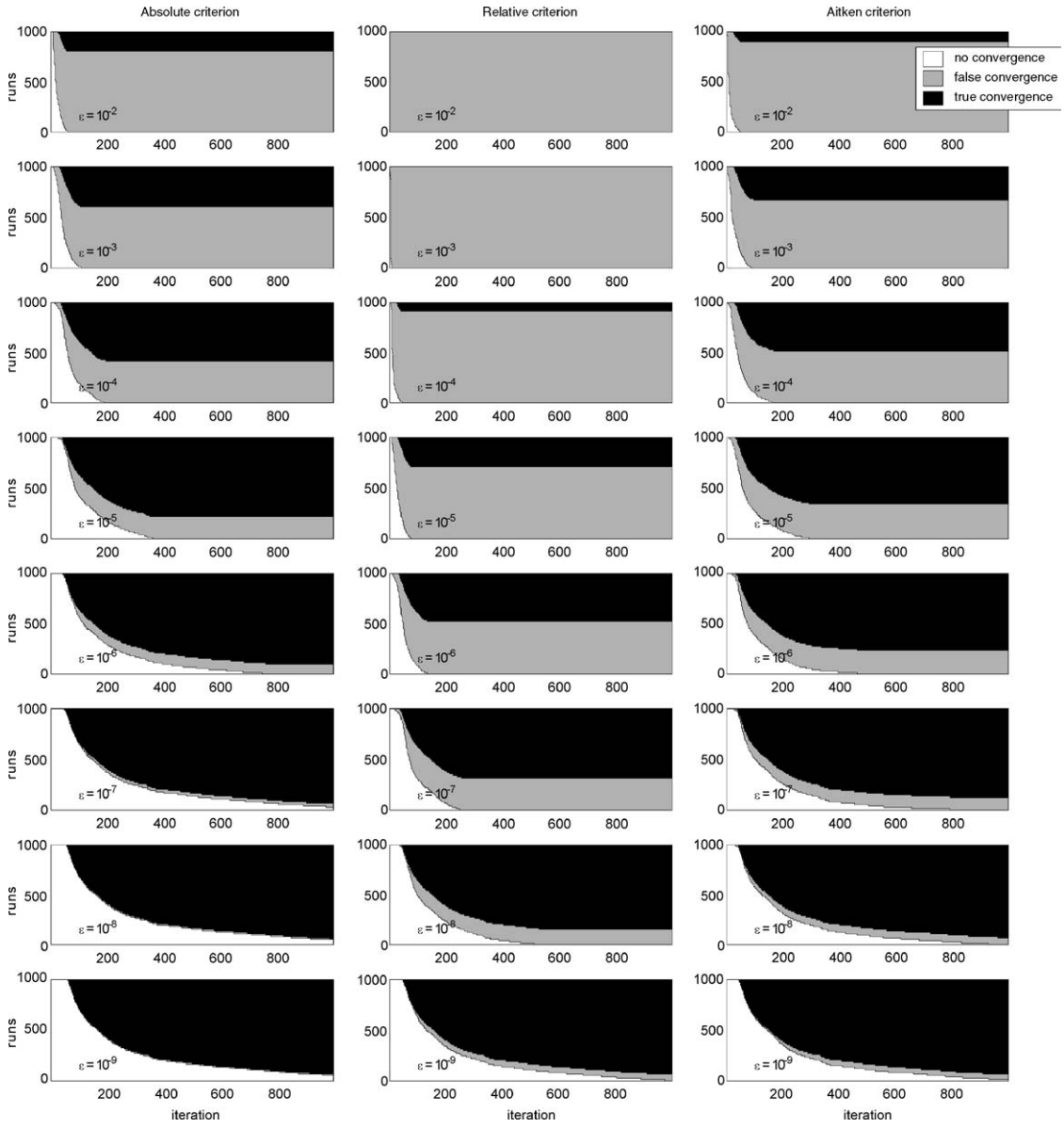

Fig. 4. *Convergence of the EM algorithm for the synthetic dataset, based on absolute, relative and Aitken criteria*

the "true" maximum value of the log-likelihood (across all runs) is larger than ϕ; *no convergence*, i.e., the stopping criterion does not hold at that iteration. In our analyses, ε varies from 10^{-2} to 10^{-9}. Figure 4 presents the results for the absolute, relative, and Aitken's absolute criteria with random centers (RC) as starting values. For $m = 1000$ and $\varepsilon \geq 10^{-6}$, all runs converged, but not always to the maximum value of the likelihood. Reducing the tolerance from $\varepsilon = 10^{-2}$ to $\varepsilon = 10^{-9}$ reduces the proportion of false solutions, increases the proportion of correct solutions, and increases the proportion of non-converged solutions. The relative criterion and the Aitken's absolute criterion underperform the absolute criterion in this dataset, but the results are dependent on the log-likelihood values. These results demand for caution when choosing a stopping criterion for the EM algorithm. Depending on the stopping rule and tolerance level one may only locate a local optimum or saddle point of the log-likelihood, and falsely report that as the maximum. We do find that using random centers (RC) for each convergence criterion decreases the proportion of false solutions (for a given number of iterations).

The *SEM algorithm* was run 1000 times. For 100 runs, the iteration process until $m = 250$ is presented in Fig. 2. We report results for runs with randomly chosen centers (RC) as initial values. In this case, 175 solutions (94 for RP) outperform the best EM solution in terms of the log-likelihood value. However, all these solutions with log-likelihood above -325.00 are not identified. If at some iteration no (or just one) observations are assigned to one of the components, the algorithm breaks down and some of the parameters become non-identified (we call this

Table 2. *Results for MCMC different label-switching strategies (IPRIOR) for the synthetic data*

	Proportions			Means			Variances		
	1	2	3	1	2	3	1	2	3
NONE	0.346	0.325	0.329	2.091	−0.875	−0.107	0.340	0.433	16.006
C1	0.398	0.332	0.270	0.576	0.499	0.034	4.950	3.610	8.220
C2	0.346	0.321	0.333	2.091	−0.915	−0.068	0.359	2.424	13.996
C3	0.335	0.336	0.329	1.340	−0.124	−0.107	0.320	0.453	16.006
CELEUX	0.346	0.325	0.329	2.091	−0.875	−0.107	0.340	0.433	16.006
STEPHENS	0.345	0.326	0.329	2.091	−0.875	−0.103	0.341	0.431	15.961
CHR	0.347	0.327	0.326	2.089	−0.875	−0.087	0.344	0.439	15.266

a degenerate solution). For the remaining 825 runs (RP: 906), 93.1% (RP: 97.8%) are in the neighborhood of the MLE in no more than 250 iterations, which clearly shows superior convergence performance as compared to the EM algorithm. Thus, the SEM is faster and displays better convergence properties, but is less stable than EM, since a proportion of the solutions is degenerate. The RP strategy increases the proportion of degenerate solutions, i.e., solutions that have component means that are closer, while some of the components may be empty. The best SEM solution has a maximum log-likelihood value of −328.67. The estimates are given in Table 1. The best SEM estimates and their standard errors are quite close to those of the best EM solution.

The *MCMC algorithm*[1] was run 1000 times. Because MCMC estimates are function of the prior (IPRIOR, CPRIOR, and HPRIOR) and label-switching strategies (NONE, C1, C2, C3, CELEUX, STEPHENS, and CHR), we investigate all combinations of priors and label-switching strategies, but present only the main results. The iteration process for 100 runs until iteration 250 using independent prior (IPRIOR) and no further processing of the draws (NONE) is presented in Fig. 2. From 1000 runs, 65 yielded degenerate solutions. For the remaining 935 runs, 934 (99.9%) runs reach the neighborhood of the MLE in no more than 250 iterations, which is clearly better than the EM and SEM algorithms. In terms of the proportion of degenerate solutions, MCMC appears more stable than SEM, but less stable than EM. Inspection of the iteration histories shows that the hierarchical prior (HPRIOR) with the constants chosen as suggested by Richardson and Green (1997) tends to suffer less from that instability than the other prior specifications. The reason may be that here the prior β is no longer set to a fixed number, but is a draw from its conditional distribution that takes the variances into account. This influences the posterior especially whenever the likelihood information on a component is poor as occurs when most observations have close to zero posterior probability for that component.

The MCMC chains were run with 25000 iterations, with a burn-in of 5000. The chains converged to the posterior distribution well before $m = 5000$. The parameter estimates (for NONE) are provided in Table 1. First, the effect of the prior specification is negligible, all three priors yielding very similar posterior means and standard deviations. These are close to the estimates and asymptotic standard errors of EM and SEM, which is to be expected since they converge asymptotically (see, e.g., Gelman *et al.* 1995). However, there are important effects of the label-switching strategies on the parameter estimates (Table 2). The identifiability constraints negatively affect parameter recovery, in particular when the component sizes or variances are constrained (strategies C1 and C3, respectively). Here, the posterior means of the μ_j are far off and their posterior standard deviations are large, while in some cases the same holds for the σ_j^2, when the μ_j are subject to identifiability constraints (strategy C2).

To shed further light on parameter recovery, the empirical cumulative distribution function is compared with the predicted cumulative distribution function obtained with each of these estimation procedures. Let $\tilde{F}(x)$ and $\hat{F}(x)$ represent the empirical and predictive cumulative distribution functions, respectively. We compute the maximum (vertical) distance between the two distributions, the Kolmogorov-Smirnov statistic $D_1 = \max|\tilde{F}(x) - \hat{F}(x)|$, and the area between the two distributions $D_2 = \int|\tilde{F}(x) - \hat{F}(x)|\,dx$. For perfect fit, we have $D_1 = D_2 = 0$. The results are presented in Table 3, where we present next to EM and SEM, those of MCMC with different label-switching strategies, but for the IPRIOR only, since those are best overall (although the differences between prior specifications are small). We observe that EM, SEM, and MCMC provide similar results. For all the priors the absence of any label-switching procedure (NONE) outperforms procedures in which identifiability constraints are imposed (C1, C2, and C3) for this dataset. Across different prior specifications, Stephens and CHR relabelling procedures are the most effective.

3.2. *Empirical data*

For the empirical dataset, we ran the *EM algorithm* 1000 times with $m = 2000$, for two components. Figure 2 presents the iteration process for the first 100 runs. In the empirical dataset the log-likelihood also presents a problematic shape, which can be inferred from the convergence of the algorithm. For some runs convergence to the maximum is fast, but in 695 cases (272 for RP) the algorithm did not converge after 250 iterations, and in

Table 3. *Distances between c.d.f.'s for synthetic and empirical data*

			MCMC (IPRIOR)						
	EM	SEM	NONE	C1	C2	C3	CELEUX	STEPHENS	CHR
Synthetic data									
D_1	0.033	0.033	0.033	0.087	0.093	0.180	0.033	0.033	0.033
D_2	0.010	0.010	0.009	0.034	0.024	0.084	0.009	0.009	0.009
Empirical data									
D_1	0.046	0.046	0.046	0.046	0.046	0.046	0.046	0.046	0.046
D_2	0.015	0.016	0.015	0.015	0.015	0.015	0.015	0.015	0.015

67 cases (8 for RP) after 2000 iterations (absolute criterion with $\varepsilon = 10^{-6}$). In this example, the use of RP starting values provides better EM performance. For this dataset too we observed that the EM algorithm was less stable. To obtain 1000 non-degenerate runs, we had to run the algorithm 1142 times, because in 142 runs the algorithm reached the boundary of the likelihood surface (note that the variance of the second component is close to zero), which did not happen with RP starting values. Figure 3 presents the log-likelihood against values of μ_1 and μ_2, fixing the other parameters at their MLE's, which reveals that a ridge in the likelihood causes these convergence problems. The best EM solution, corresponding to the log-likelihood value of -249.45, is presented in Table 4. For each of the three stopping rules, both starting values strategies, and $\varepsilon = 10^{-2}$ and $\varepsilon = 10^{-3}$ almost none of the runs stopped at the maximum of the log-likelihood surface. For $\varepsilon \geq 10^{-7}$ almost all runs converged, but not always to the maximum of the likelihood (as identified across all runs). Reducing the tolerance from $\varepsilon = 10^{-2}$ to $\varepsilon = 10^{-9}$ reduces the proportion of false solutions, increases the proportion of correct solutions, and increases the proportion of non-convergence. The relative criterion and the Aitken's absolute criterion do worse than the absolute criterion in this dataset too. These results confirm those for the synthetic data: the reported EM solution strongly depends on the initial values, the strategy to generate them, the stopping rule, and its tolerance level, while problems of local maxima are exacerbated by incorrect choice of that rule and its tolerance.

As before, we run the *SEM algorithm* 1000 times. Figure 2 presents 100 runs up to $m = 250$. With this empirical dataset, we obtain 70 non-identified solutions (RP: 69), with a log-likelihood value above -249.45. For the remaining 930 runs, 73.2% (RP: 72.5%) reach the neighborhood of the MLE in no more than 250 iterations. Thus, on the empirical data too SEM is faster and displays better convergence properties, but is again more instable than EM. The best (identified) SEM solution has a log-likelihood value of -249.46. The SEM estimates are given in Table 4. The estimates and their asymptotic standard errors are very close to those of EM.

The *MCMC algorithm*[2] was run 1000 times using independent prior (IPRIOR) and no further processing of the draws (NONE). Figure 2 gives the log-likelihood against the iterations up to $m = 250$. From 1000 runs, 65 were degenerate. From the remaining 935 runs, 707 (75.6%) reach the neighborhood of the MLE in no more than 250 iterations, so that MCMC outperforms EM too, and outperforms SEM slightly in terms of convergence properties. Although MCMC is faster than the EM algorithm, it is less stable.

Then, the MCMC chain was run for 25000 iterations, from which the first 5000 were excluded. Convergence is fast, well before iteration $m = 5000$. Although we investigate all combinations of priors and label-switching strategies, we present the estimates obtained with different prior specifications (IPRIOR, CPRIOR, and HPRIOR) without a label-switching strategy (NONE) in Table 4. It seems that MCMC gives

Table 4. *EM, SEM, and MCMC (NONE) results for the empirical data*[*]

	Proportions		Means		Variances	
	1	2	1	2	1	2
EM	0.899 (0.03)	0.101 (−)	8.148 (0.08)	10.027 (0.02)	1.002 (0.16)	0.006 (0.004)
SEM	0.897 (0.03)	0.103 (−)	8.142 (0.08)	10.027 (0.02)	0.994 (0.16)	0.007 (0.004)
MCMC (NONE)						
IPRIOR	0.871 (0.04)	0.129 (0.04)	8.099 (0.10)	9.992 (0.05)	0.957 (0.11)	0.023 (0.02)
CPRIOR	0.879 (0.04)	0.121 (0.04)	8.116 (0.10)	10.004 (0.05)	0.969 (0.11)	0.017 (0.01)
HPRIOR	0.834 (0.04)	0.166 (0.04)	8.028 (0.10)	9.931 (0.08)	0.851 (0.10)	0.062 (0.04)

[*]Asymptotic standard errors for the EM and SEM estimates and posterior standard errors for the MCMC estimates, no label-switching strategy applied.

Table 5. *Results for MCMC different label-switching strategies (IPRIOR) for the empirical data*

	Proportions		Means		Variances	
	1	2	1	2	1	2
NONE	0.871	0.129	8.099	9.992	0.957	0.023
C1	0.871	0.129	8.099	9.992	0.957	0.023
C2	0.871	0.129	8.099	9.992	0.957	0.023
C3	0.871	0.129	8.099	9.992	0.957	0.023
CELEUX	0.871	0.129	8.099	9.992	0.957	0.023
STEPHENS	0.874	0.126	8.107	9.994	0.965	0.021
CHR	0.883	0.117	8.115	10.002	0.977	0.019

larger estimates of the size of the smallest component. There are some small differences between different prior specifications (which are the same across label switching strategies). The posterior means of the μ_j for the conjugate prior are closest to the MLE's, those for the independent prior and hierarchical prior seem to be slightly biased downwards. Relative to the MLE, the posterior means of the σ_j^2 are lower for all prior specifications, but, somewhat less so for the conjugate prior. The posterior standard deviations are somewhat larger than the asymptotic SE's in most cases. The results for different label-switching strategies are virtually the same, which is caused by the fact that the components are well separated and label switching did not occur (Table 5).[3] The D_1 and D_2 statistics show that all priors yield very similar predictive c.d.f.'s. The procedures to prevent label-switching have no influence on the result (Table 3).

4. Conclusion

The mixture log-likelihood presents well known problems to parameter estimation with iterative procedures such as EM, SEM, and MCMC. These include slow convergence, degenerate solutions, label switching, and convergence to local optima. We have investigated such problems in detail for a synthetic and an empirical dataset, where the log-likelihoods present particularly problematic shapes, involving ridges and/or saddle points.

We find that EM converges slowly and often fails to converge to the global maximum of the likelihood surface. EM convergence is very dependent upon the type of starting values, stopping rule used, and its tolerance level, which is due to flat regions on the log-likelihood surface from which it can only escape by chance and/or if the tolerance level is sufficiently low.

SEM was shown to be an attractive alternative to exploring those complex likelihood surfaces, since it exhibits much faster and more reliable convergence. The simulation step enables this algorithm to escape from saddle points in the likelihood. However, SEM tends to be less stable, a proportion of the solutions being degenerate due to allocation of single observations to a mixture component. Although SEM can be affected by label switching, this did not occur in our two examples.

MCMC convergence properties appear superior to EM and somewhat better than SEM. Thus, we add to the theoretical results of Sahu and Roberts (1999), who derive that the approximate rates of convergence of these algorithms are the same. However, they also show that under some conditions convergence of MCMC is faster than that of EM, and our examples illustrate that this may be the case especially in mixture models where the likelihood surface displays ridges and/or saddle points. MCMC does suffer from degenerate solutions, but less so than SEM. In both synthetic and empirical data applications, in spite of the complex shape of the likelihood, label switching did not present much of a problem. Label switching strategies that impose identifiability constraints on the parameters were in fact found to deteriorate the solutions, while those based on clustering techniques (Celeux 1998, Stephens 2000) and minimization of loss functions (Celeux, Hurn and Robert 2000) perform well. The effect of different prior specifications on convergence and parameter recovery was small. This may be caused by the fact that all specifications involve proper non-informative priors (Sahu and Roberts 1999).

Thus, although problems have been reported in estimating mixture models with MCMC methods, our results reveal that problems of label switching may not be severe, especially if there are properly addressed, for which Stephens' method (Stephens 2000) and CHR's method are preferred. Our results corroborate the conclusion by Celeux, Hurn and Robert (2000) that simpler methods such as Celeux's method (clustering) are a good approximation to (and less time-consuming than) the more elegant approach based on loss functions. We conclude that MCMC is preferable over EM and SEM in recovering the parameters of mixture models, in particular if the shape of the likelihood surface is problematic, exhibiting ridges, flat regions and/or saddle points as was the case in both our synthetic and empirical data. However, one should also take into account the cost of implementing these algorithms. The relative merits then become less pronounced, especially whenever one has to handle the label-switching problem.

Our paper focused on mixtures of univariate Gaussian distributions. For high dimensional data (e.g., mixture of multivariate Gaussian distributions), the advantages of MCMC algorithms for problematic likelihoods remain to be investigated. Future research could extend our findings to mixtures for high dimensional data, other MCMC algorithms beyond the Gibbs sampler that may ensure adequate mixing, mixtures of other distributions, and more complex mixture models, such as mixtures of regressions.

Acknowledgment

José G. Dias' research was supported by Fundação para a Ciência e Tecnologia Grant no. SFRH/BD/890/2000 (Portugal) and conducted mainly at the University of Groningen (Faculty of Economics and Population Research Centre). The authors are thankful to the Editor and anonymous referees for valuable comments which improved the paper significantly.

Notes

1. For IPRIOR and CPRIOR, we set $\alpha = \beta = 0.001$. For CPRIOR, we set $\lambda = 0.001$. For HPRIOR values are based on a partially empirical Bayes approach (Richardson and Green 1997): ξ corresponds to the midpoint of the observed data range, R corresponds to the length of the observed interval, $\kappa = R^{-2}$, $\alpha = 2$, $g = 0.2$, $h = 100g/\alpha R^2$, and $\delta = 1$. From the values chosen for ξ and κ, a prior for μ results which is fairly flat over the observed range of data. The choice of $\alpha = 2$ with a relatively flat hyperprior on the hyperparameter β expresses the belief that the σ_j^2 are similar (Richardson and Green 1997). The Dirichlet distribution with $\delta = 1$ gives a uniform prior over the space $\sum_{j=1}^{k} \pi_j = 1$. These values define proper but vague priors.
2. We have: $k = 2$, $\xi = 2.1463$, $\kappa = 0.0543$ and $h = 0.5427$.
3. We thank anonymous reviewers for pointing out that this may indicate lack of convergence of the MCMC sampler; formally it has converged only if it has visited all the k! modes in the posterior distribution, but that may in this case be of no importance for statistical inference.

References

Celeux G. 1998. Bayesian inference for mixture: The label switching problem. In: Payne R. and Green P. (Eds.), COMPSTAT 98. Physica-Verlag, Heidelberg, pp. 227–232.

Celeux G. and Diebolt J. 1985. The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. Computational Statistics Quarterly 2: 73–82.

Celeux G., Chauveau D., and Diebolt J. 1996. Stochastic versions of the EM algorithm: An experimental study in the mixture case. Journal of Statistical Computation and Simulations 55: 287–314.

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.

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 from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society B 39: 1–38.

Diebolt J. and Ip E.H.S. 1996. Stochastic EM: Method and application. In: Gilks W.R., Richardson S.T., and Spiegelhalter D.J. (Eds.), Markov Chain Monte Carlo in Practice. Chapman & Hall, London, pp. 259–273.

Diebolt J. and Robert C.P. 1994. Estimation of finite mixture distributions through Bayesian sampling. Journal of the Royal Statistical Society B 56(2): 363–375.

Escobar M.D. and West M. 1995. Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association 90: 577–588.

Fletcher R. 1980. Practical Methods of Optimization. Vol. 1 Unconstrained Optimization. John Wiley & Sons, Chichester.

Frigessi A. and Rue H. 1997. Bayesian image classification with Baddeley's delta loss. Journal of Computational and Graphical Statistics 6(1): 55–73.

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.

Gelman A., Carlin J.B., Stern H.S., and Rubin D.B. 1995. Bayesian Data Analysis. Chapman & Hall, Boca Raton.

Hurn M., Justel A., and Robert C.P. 2003. Estimating mixtures of regressions. Journal of Computational and Graphical Statistics 12(1): 55–79.

McLachlan G.J. and Krishnan T. 1997. The EM Algorithm and Extensions. John Wiley & Sons, New York.

McLachlan G.J. and Peel D. 2000. Finite Mixture Models. John Wiley & Sons, New York.

Richardson S. and Green P.J. 1997. On Bayesian analysis of mixtures with an unknown number of components (with discussion). Journal of the Royal Statistical Society B 59(4): 731–792.

Robert C.P. 1996. Mixtures of distributions: Inference and estimation. In: Gilks W.R., Richardson S., and Spiegelhalter D.J. (Eds.), Markov Chain Monte Carlo in Practice. Chapman & Hall, London, pp. 441–464.

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

Rue H. 1995. New loss functions in Bayesian imaging. Journal of the American Statistical Association 90(431): 900–908.

Sahu S.K. and Roberts G.O. 1999. On convergence of the EM algorithm and the Gibbs sampler. Statistics and Computing 9: 55–64.

Stephens M. 1997a. Discussion on 'On Bayesian analysis of mixtures with an unknown number of components (with discussion)'. Journal of Royal Statistical Society B 59(4): 768–769.

Stephens M. 1997b. Bayesian methods for mixtures of normal distributions, DPhil Thesis Oxford. University of Oxford.

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

Tanner M.A. and Wong W.H. 1987. The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association 82: 528–540.

UNDP. 2000. Human Development Report 2000. Human Rights and Human Development United Nations Development Programme, Oxford University Press, New York.

Vlassis N. and Likas A. 2002. A greedy EM algorithm for Gaussian mixture learning. Neural Processing Letters 15: 77–87.

Wedel M. and DeSarbo W.S. 1995. A mixture likelihood approach for generalized linear models. Journal of Classification 12: 21–55.

Wu C.F.J. 1983. On the convergence properties of the EM algorithm. Annals of Statistics 11: 95–103.