An Empirical Comparison of EM Initialization Methods and Model Choice Criteria for Mixtures of Skew-Normal Distributions*

José Raimundo Gomes Pereira Celso Rômulo Barbosa Cabral
Leyne Abuim de Vasconcelos Marques
José Mir Justino da Costa

Departamento de Estatística – Universidade Federal do Amazonas – Brazil

Abstract

We investigate, via simulation study, the performance of the EM algorithm for maximum likelihood estimation in finite mixtures of skew-normal distributions with component specific parameters. The study takes into account the initialization method, the number of iterations needed to attain a fixed stopping rule and the ability of some classical model choice criteria to estimate the correct number of mixture components. The results show that the algorithm produces quite reasonable estimates when using the method of moments to obtain the starting points and that, combining them with the AIC, BIC, ICL or EDC criteria, represents a good alternative to estimate the number of components of the mixture. Exceptions occur in the estimation of the skewness parameters, notably when the sample size is relatively small, and in some classical problematic cases, as when the mixture components are poorly separated.
Key Words: EM algorithm; Skew-normal distribution; Finite mixture of distributions.

1 Introduction

We define the (finite) mixture of the densities $f_1, \ldots, f_k$ as the density $g(x) = \sum_{i=1}^{k} p_i f_i(x)$, where $p_1, \ldots, p_k$ are unknown positive numbers, named *mixing weights*, satisfying $\sum_{i=1}^{k} p_i = 1$. We name f_i the $i-th$ *component of the mixture*.

Finite mixtures have been widely used as a powerful tool to model heterogeneous data and to approximate complicated probability densities, presenting multimodality, skewness and heavy tails. These models have been applied in several areas like genetics, image processing, medicine and economics. For comprehensive surveys, see McLachlan & Peel (2000) and Frühwirth-Schnatter (2006).

Maximum likelihood estimation in finite mixtures is a research area with several challenging aspects. There are nontrivial issues, like lack of identifiability and saddle regions surrounding the possible local maxima of the likelihood. Another problem is that the likelihood is possibly unbounded, which happens when the components are normal densities, for example.

*The authors acknowledge the partial financial support from CAPES, CNPq and FAPEAM. *Email addresses*: jrpereira@ufam.edu.br (José Raimundo Gomes Pereira), celsoromulo@gmail.com (Celso Rômulo Barbosa Cabral), leyneabuim@gmail.com (Leyne Abuim de Vasconcelos Marques), zemirufam@gmail.com (José Mir Justino da Costa)

There is a lot of literature involving mixtures of normal distributions, some references can be found in the above-mentioned books. In this work we consider mixtures of *skew-normal (SN) distributions*, as defined by Azzalini (1985). This distribution is an extension of the normal distribution that accommodates asymmetry.

The standard algorithm for maximum likelihood estimation in finite mixtures is the *Expectation Maximization (EM)* of Dempster *et al.* (1977), see also McLachlan & Krishnan (2008). It is well known that it has slow convergence and that its performance is strongly dependent on the stopping rule and starting points. For normal mixtures, several authors have computationally investigated the performance of the EM algorithm by taking into account initial values (Karlis & Xekalaki, 2003; Biernacki *et al.*, 2003) and asymptotic properties (Nityasuddhi & Böhning, 2003). See also Dias & Wedel (2004) for a broader simulation experiment comparing EM with SEM and MCMC algorithms.

Although there are some purposes to overcome the unboundedness problem in the normal mixture case, involving constrained optimization and alternative algorithms – see Hathaway (1985), Ingrassia (2004), and Yao (2010), for example, it is interesting to investigate the performance of the (unrestricted) EM algorithm in the presence of skewness in the component distributions, since algorithms of this kind have been presented in recent works as Lin *et al.* (2007a), Lin *et al.* (2007b), Basso *et al.* (2009), Lin (2009a), Lin (2009b) and Lin & Lin (2009).

The main goal of this work is to study the performance of the estimates produced by the EM algorithm, taking into account two different purposes to obtain initial values – method of moments and a random initialization method, the number of iterations needed to attain a fixed stopping rule and the ability of some classical model choice criteria (AIC, BIC, ICL and EDC) to estimate the correct number of mixture components. For each initialization method, the consistency of the estimates is analyzed by computing their bias and mean squared errors over repeated generated samples and the mean number of iterations is observed. Fixing the method of initialization, repeated samples of finite mixtures of SN distributions are generated with different number of components and the percentage of times some given criterion chooses the correct number of components is observed. The work is restricted to the univariate case.

The remainder of the paper is organized as follows. In Sections 2, 3 and 4 for the sake of completeness, we give a brief sketch of the SN distribution, of the skew-normal mixture model and of estimation via the EM algorithm, respectively. The simulation study through the analysis of the initialization methods and number of iterations is presented in Section 5. The study through the analysis of the model choice criteria is presented in Section 6. Finally, in order to compare results obtained by the several initialization and model choice methods, two real data sets are analyzed in Section 7.

2 The Skew-normal Distribution

The skew-normal distribution, introduced by Azzalini (1985), is given by the density

$$\mathrm{SN}(y|\mu,\sigma^2,\lambda) = 2\mathrm{N}(y|\mu,\sigma^2)\Phi\left(\lambda\frac{y-\mu}{\sigma}\right),$$

where $\mathrm{N}(\cdot|\mu,\sigma^2)$ denotes the univariate normal density with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ and $\Phi(\cdot)$ is the distribution function of the standard normal distribution. In this definition, μ, $\lambda \in \mathbb{R}$ and σ^2 are parameters regulating location, skewness and scale, respectively. For a random variable Y with this distribution, we use the notation $Y \sim \mathrm{SN}(\mu,\sigma^2,\lambda)$.

For a SN random variable Y, a convenient stochastic representation is given next. It can be used to simulate realizations of Y and to implement the EM-type algorithm that will be shown soon. The proof follows easily from Henze (1986).

Lemma 1. *A random variable $Y \sim SN(\mu, \sigma^2, \lambda)$ has a stochastic representation given by*

$$Y = \mu + \sigma\delta T + \sigma(1-\delta^2)^{1/2} T_1,$$

where $\delta = \lambda/\sqrt{1+\lambda^2}$, $T = |T_0|$, T_0 and T_1 are independent standard normal random variables and $|\cdot|$ denotes absolute value.

In what follows, in order to reduce computational difficulties related to the implementation of the algorithms used for estimation, we use the parametrization

$$\Gamma = (1-\delta^2)\sigma^2 \qquad \text{and} \qquad \Delta = \sigma\delta,$$

which was first suggested by Bayes & Branco (2007). Note that $(\lambda, \sigma^2) \to (\Delta, \Gamma)$ is a one to one mapping. To recover λ and σ^2, we use

$$\lambda = \Delta/\sqrt{\Gamma} \quad \text{and} \quad \sigma^2 = \Delta^2 + \Gamma.$$

Then, it follows easily from Lemma 1 that

$$Y|T = t \sim N(\mu + \Delta t, \Gamma) \quad \text{and} \quad T \sim HN(0,1), \tag{1}$$

where $HN(0,1)$ denotes the half-normal distribution with parameters 0 and 1.

The expectation, variance and skewness coefficient of $Y \sim \mathrm{SN}(\mu, \sigma^2, \lambda)$ are respectively given by

$$E(Y) = \mu + \sigma\Delta\sqrt{2/\pi}, \quad Var[Y] = \sigma^2\left(1 - \frac{2}{\pi}\delta^2\right), \quad \gamma(Y) = \frac{\kappa\delta^3}{(1-\frac{2}{\pi}\delta^2)^{3/2}}, \tag{2}$$

where $\kappa = \frac{4-\pi}{2}(\frac{2}{\pi})^{3/2}$, see Azzalini (2005, Lemma 2).

3 The FM-SN Model

In this section we give a brief introduction to the finite mixture of SN distributions model, hereafter FM-SN model. More details can be found in Basso *et al.* (2009), where a more general class of distributions is considered, and references herein.

The FM-SN model is defined by considering a random sample $\mathbf{y} = (y_1, \ldots, y_n)^\top$ from a mixture of SN densities given by

$$g(y_j|\Theta) = \sum_{i=1}^{k} p_i \mathrm{SN}(y_j|\theta_i), \qquad j = 1, \ldots, n, \tag{3}$$

where $p_i \geq 0$, $i = 1, \ldots, k$ are the mixing probabilities, $\sum_{i=1}^k p_i = 1$, $\theta_i = (\mu_i, \sigma_i^2, \lambda_i)^\top$ is the specific vector of parameters for the component i and $\Theta = ((p_1, \ldots, p_k)^\top, \theta_1^\top, \ldots, \theta_k^\top)^\top$ is the vector with all parameters.

For each j consider a latent classification random variable Z_j taking values in $\{1, \ldots, k\}$, such that

$$y_j|Z_j = i \sim \mathrm{SN}(\theta_i), \quad P(Z_j = i) = p_i, \quad i = 1, \ldots, k; \quad j = 1, \ldots, n.$$

Then it is straightforward to prove, integrating out Z_j, that y_j has density (3). If we combine this result with (1), we have the following stochastic representation for the FM-SN model

$$\begin{aligned} &y_j|T_j = t_j, Z_j = i \sim \mathrm{N}(\mu_i + \Delta_i t_j, \Gamma_i), \\ &T_j \sim \mathrm{HN}(0,1), \\ &P(Z_j = i) = p_i, \quad i = 1, \ldots, k; \quad j = 1, \ldots, n, \end{aligned}$$

where

$$\Gamma_i = (1-\delta_i^2)\sigma_i^2, \qquad \Delta_i = \sigma_i\delta_i, \qquad \delta_i = \lambda_i/\sqrt{1+\lambda_i^2}, \qquad i = 1, \ldots, k. \tag{4}$$

4 Estimation

4.1 An EM-type Algorithm

In this section, for the sake of completeness, we present an EM-type algorithm for estimation of the parameters of a FM-SN distribution. This algorithm was presented before in Basso *et al.* (2009), where a wider class of distributions that includes the FM-SN one is considered.

Here we obtain estimates of the parameters of the FM-SN model using a faster extension of EM called the ECM algorithm (Meng & Rubin, 1993). When applying it to the FM-SN model, we obtain a simple set of closed form expressions to update a current estimate of the vector Θ, as we will see below. It is important to emphasize that this procedure differs from the algorithm presented by Lin *et al.* (2007b), because in the former case the updating equations for the component skewness parameter have a closed form.

In what follows we consider the parametrization (4). Let

$$\hat{\Theta}^{(m)} = ((\hat{p}_1^{(m)},\ldots,\hat{p}_k^{(m)})^\top, (\hat{\theta}_1^{(m)})^\top,\ldots,(\hat{\theta}_k^{(m)})^\top)^\top$$

be the current estimate (at the mth iteration of the algorithm) of Θ, where $\hat{\theta}_i^{(m)} = (\hat{\mu}_i^{(m)}, \hat{\Delta}_i^{(m)}, \hat{\Gamma}_i^{(m)})^\top$. The E-step of the algorithm consists in evaluate the expected complete data function – known as the *Q−function*, defined as

$$Q(\Theta|\hat{\Theta}^{(m)}) = E[\ell_c(\Theta)|\mathbf{y}, \hat{\Theta}^{(m)}],$$

where $\ell_c(\Theta)$ is the *complete-data log-likelihood function*, given by

$$\ell_c(\Theta) = c + \sum_{j=1}^{n}\sum_{i=1}^{k} z_{ij}\left(\log p_i - \frac{1}{2}\log\Gamma_i - \frac{1}{2\Gamma_i}(y_j - \mu_i - \Delta_i t_j)^2\right),$$

where z_{ij} is the indicator function of the set $(Z_j = i)$ and c is a constant that is independent of Θ.

The M-step consists in maximizing the Q-function over Θ. As the M-step turns out to be analytically intractable, we use, alternatively, the ECM algorithm, which is an extension that essentially replaces it with a sequence of conditional maximization (CM) steps.

The following scheme is used to obtain an updated value $\hat{\Theta}^{(m+1)}$. We can find more details about the conditional expectations involved in the computation of the Q-function and the related maximization steps in Basso *et al.* (2009). Here, ϕ denotes the standard normal density

E-step: Given a current estimate $\hat{\Theta}^{(m)}$, compute $\hat{z}_{ij}$, $\hat{s}_{1ij}$ and $\hat{s}_{2ij}$, for $i = 1,\ldots,n$ and $j = 1,\ldots,k$, where

$$\begin{aligned}
\hat{z}_{ij}^{(m)} &= \frac{\hat{p}_i^{(m)}\mathrm{SN}(y_j|\hat{\theta}_i^{(m)})}{\sum_{i=1}^{k}\hat{p}_i^{(m)}\mathrm{SN}(y_j|\hat{\theta}_i^{(m)})}, \qquad (5)\\
\hat{s}_{1ij}^{(m)} &= \hat{z}_{ij}^{(m)}\left[\hat{\mu}_{T_{ij}}^{(m)} + \frac{\phi\left(\hat{\mu}_{T_{ij}}^{(m)}/\hat{\sigma}_{T_i}^{(m)}\right)}{\Phi\left(\hat{\mu}_{T_{ij}}^{(m)}/\hat{\sigma}_{T_i}^{(m)}\right)}\hat{\sigma}_{T_i}^{(m)}\right],\\
\hat{s}_{2ij}^{(m)} &= \hat{z}_{ij}^{(m)}\left[(\hat{\mu}_{T_{ij}}^{(m)})^2 + (\hat{\sigma}_{T_i}^{(m)})^2 + \frac{\phi\left(\hat{\mu}_{T_{ij}}^{(m)}/\hat{\sigma}_{T_i}^{(m)}\right)}{\Phi\left(\hat{\mu}_{T_{ij}}^{(m)}/\hat{\sigma}_{T_i}^{(m)}\right)}\hat{\mu}_{T_{ij}}^{(m)}\,\hat{\sigma}_{T_i}^{(m)}\right],\\
\hat{\mu}_{T_{ij}}^{(m)} &= \frac{\hat{\Delta}_i^{(m)}}{\hat{\Gamma}_i^{(m)} + (\hat{\Delta}_i^{(m)})^2}(y_j - \hat{\mu}_i^{(m)}),\\
\hat{\sigma}_{T_i}^{(m)} &= \left(\frac{\hat{\Gamma}_i^{(m)}}{\hat{\Gamma}_i^{(m)} + (\hat{\Delta}_i^{(m)})^2}\right)^{1/2}.
\end{aligned}$$

CM-steps: Update $\hat{\Theta}^{(m)}$ by maximizing $Q(\Theta|\hat{\Theta}^{(m)})$ over Θ, which leads to the following closed form expressions:

$$\begin{aligned}
\hat{p}_i^{(m+1)} &= n^{-1}\sum_{j=1}^{n}\hat{z}_{ij}^{(m)} \ , \\
\hat{\mu}_i^{(m+1)} &= \frac{\sum_{j=1}^{n}(y_j\hat{z}_{ij}^{(m)} - \hat{\Delta}_i^{(m)}\hat{s}_{1ij}^{(m)})}{\sum_{j=1}^{n}\hat{z}_{ij}^{(m)}} \ , \\
\hat{\Gamma}_i^{(m+1)} &= \frac{\sum_{j=1}^{n}(\hat{z}_{ij}^{(m)}(y_j-\hat{\mu}_i^{(m+1)})^2 - 2(y_j-\hat{\mu}_i^{(m+1)})\hat{\Delta}_i^{(m)}\hat{s}_{1ij}^{(m)} + (\hat{\Delta}_i^{(m)})^2\hat{s}_{2ij}^{(m)})}{\sum_{j=1}^{n}\hat{z}_{ij}^{(m)}} \ , \\
\hat{\Delta}_i^{(m+1)} &= \frac{\sum_{j=1}^{n}(y_j-\hat{\mu}_i^{(m+1)})\hat{s}_{1ij}^{(m)}}{\sum_{j=1}^{n}\hat{s}_{2ij}^{(m)}} \ .
\end{aligned}$$

The algorithm iterates between the E and CM steps until a suitable convergence rule is satisfied. Some classical purposes are (i) if $||\hat{\Theta}^{(m+1)} - \hat{\Theta}^{(m)}||$ is sufficiently small or (ii) if $|\ell(\hat{\Theta}^{(m+1)}) - \ell(\hat{\Theta}^{(m)})|$ is small enough or (iii) if $|\ell(\hat{\Theta}^{(m+1)})/\ell(\hat{\Theta}^{(m)}) - 1|$ is small enough, where $\ell(\Theta)$ the actual log-likelihood. In this work we use the purpose (ii).

4.2 Problems with Estimation in Finite Mixtures

4.2.1 Unbounded Likelihood

An important issue is the existence of the maximum likelihood estimator for finite mixture models. It is well known that the likelihood of normal mixtures can be unbounded – see Frühwirth-Schnatter (2006, Chapter 6), for example. FM-SN models also have this feature: for instance, let y_i, $i = 1,\ldots,n$, be a random sample from a mixture of two SN densities, $\text{SN}(\mu,\sigma_1^2,\lambda_1)$ with weight $p \in (0,1)$ and $\text{SN}(\mu,\sigma_2^2,\lambda_2)$. Fixing p, σ_1^2, λ_1 and λ_2, we have that the likelihood function $L(\mu,\sigma_2^2)$ evaluated at $\mu = y_1$ satisfies

$$L(y_1,\sigma_2^2) = g(\sigma_2^2)\prod_{i=2}^{n}(c_i + h_i(\sigma_2^2)) > g(\sigma_2^2)\prod_{i=2}^{n}c_i,$$

where

$$\begin{aligned}
g(\sigma_2^2) &= \left(p(2\pi\sigma_1^2)^{-1/2} + (1-p)(2\pi\sigma_2^2)^{-1/2}\right), \\
c_i &= p\text{SN}(y_i|y_1,\sigma_1^2,\lambda_1), \\
h_i(\sigma_2^2) &= (1-p)\text{SN}(y_i|y_1,\sigma_2^2,\lambda_2), \qquad i = 2,\ldots,n.
\end{aligned}$$

The strict inequality is valid because c_i and $h_i(\sigma_2^2)$ are positive. Now, we can choose a sequence of positive numbers $(\sigma_2^2)_m$, $m \in \mathbb{N}$, such that $\lim_{m\to\infty}(\sigma_2^2)_m = 0$. Then, because $\prod_{i=2}^{n}c_i > 0$ and $\lim_{m\to\infty} g((\sigma_2^2)_m) = +\infty$, we have that $\lim_{m\to\infty} L(y_1,(\sigma_2^2)_m) = +\infty$.

Thus, following Nityasuddhi & Böhning (2003), we mention the estimates produced by the algorithm of section 4.1 as "EM estimates", that is, some sort of solution of the score equation, instead of "maximum likelihood estimates".

4.2.2 Non-identifiability

Another nontrivial issue is the lack of identifiability. Strictly speaking, finite mixtures are always non-identifiable because an arbitrary permutation of the labels of the component parameters lead to the same finite mixture distribution. In the finite mixture context, a more flexible concept of identifiability is used and is defined as follows: the class of FM-SN models will be identifiable if distinct component densities correspond to distinct mixtures. More specifically, if we have two representations for the same mixture distribution, say

$$\sum_{i=1}^{k} p_i' \mathrm{SN}(\cdot|\mu_i', (\sigma_i')^2, \lambda_i') = \sum_{i=1}^{k} p_i \mathrm{SN}(\cdot|\mu_i, \sigma_i^2, \lambda_i),$$

then

$$p_i' = p_{\rho(i)}, \quad \mu_i' = \mu_{\rho(i)}, \quad (\sigma_i')^2 = \sigma_{\rho(i)}^2, \quad \lambda_i' = \lambda_{\rho(i)}$$

for some permutation ρ of the indexes $1, \ldots, k$ (Titterington *et al.*, 1985, Chapter 3).

The normal mixture model identifiability was first verified by Yakowitz & Spragins (1968), but it is interesting to note that subsequent discussions in the related literature concerning mixtures of Student-t distributions – see, for example, Peel & McLachlan (2000), Shoham (2002), Shoham *et al.* (2003) and Lin *et al.* (2004) – do not presented a formal proof of its identifiability.

The above commentaries may cause some caution when using maximum likelihood estimation in the finite mixture case. One way to circumvent the unboundedness problem is the constrained optimization of the likelihood, imposing conditions on the component variances in order to obtain global maximization. See Hathaway (1985), Ingrassia (2004), Ingrassia & Rocci (2007) and Greselin & Ingrassia (2009), for example. The non-identifiability problem is not a major one if we are interested only in the likelihood values, which are robust to label switching. This is the case, for example, when density estimation is the main goal.

In our study we investigate, as in Nityasuddhi & Böhning (2003), only the performance of the EM algorithm when are considered component specific parameters (that is, unrestricted) of the mixture.

5 A Simulation Study of Initial Values for the EM Algorithm

5.1 A Study of Consistency Through the Analysis of Bias and MSE

The EM algorithm is an efficient standard tool for maximum likelihood estimation in finite mixtures of distributions. It is well known that its performance is strongly dependent on the choice of the criterion of convergence and starting points. In this work, we do not consider the stopping rule issue. A detailed discussion about this theme can be founded in McLachlan & Peel (2000). Our fixed rule is to stop the process at stage m when

$$|\ell(\hat{\Theta}^{(m+1)}) - \ell(\hat{\Theta}^{(m)})| < c,$$

where we fixed $c = 10^{-5}$.

The choice of starting values for the EM algorithm plays a big role in parameter estimation. In the mixture context, its importance is amplified because, as noted by Dias & Wedel (2004), there are various saddle regions surrounding the possible local maxima of the likelihood function, and the EM algorithm can be trapped in some of these subsets of the parameter space.

In this work, we make a simulation study in order to compare some methods to obtain starting points for the algorithm proposed in section 4.1. Previous related work, considering normal mixtures, can be found in Karlis & Xekalaki (2003) and Biernacki *et al.* (2003). An interesting question is to investigate, in a similar fashion, the performance of the EM algorithm when a skewness parameter for each component density is considered.

We consider the following methods to obtain initial values:

The True Values Method (TVM): consists in to initialize the algorithm with the distributional parameter values used to generate the artificial random samples. Of course, the best results are expected when using this method. It has been included in our experiment playing a gold standard role.

The Random Values Method (RVM): to fit the FM-SN model with k components, we first divide the generated random sample into k sub-samples. To do this, the *k-means* method is employed – see Johnson & Wichern (2007) to obtain details about this clustering method. Let φ_i be the sub-sample i. Consider the following points artificially generated from uniform distributions over the specified intervals

$$\begin{aligned} \hat{\xi}_i^{(0)} &\sim U(\min\{\varphi_i\}, \max\{\varphi_i\}), \\ \hat{\omega}_i^{(0)} &\sim U(0, Var\{\varphi_i\}), \\ \hat{\gamma}_i^{(0)} &\sim U(-0.9953, 0.9953), \end{aligned} \tag{6}$$

where $\min\{\varphi_i\}$, $\max\{\varphi_i\}$ and $Var\{\varphi_i\}$ denote, respectively, the minimum, the maximum and the sample variance of φ_i, $i = 1,..,k$. These quantities are taken as rough estimates for the mean, variance and skewness coefficient associated to sub-population i (from which sub-sample i is supposedly collected), respectively. The suggested interval for $\hat{\gamma}_i^{(0)}$ is inspired by the fact that the range for the skewness coefficient in SN models is $(-0.9953, 0.9953)$.

The starting points for the specific component locations, scale and skewness parameters are given respectively by

$$\begin{aligned} \hat{\mu}_i^{(0)} &= \hat{\xi}_i^{(0)} - \sqrt{2/\pi}\,\delta_{(\hat{\lambda}_i^{(0)})}\hat{\sigma}_i^{(0)}, \\ \hat{\sigma}_i^{(0)} &= \sqrt{\frac{\hat{\omega}_i^{(0)}}{1 - \frac{2}{\pi}\delta^2_{(\hat{\lambda}_i^{(0)})}}}, \\ \hat{\lambda}_i^{(0)} &= \pm\sqrt{\frac{\pi(\hat{\gamma}_i^{(0)})^{2/3}}{2^{1/3}(4-\pi)^{2/3} - (\pi-2)(\hat{\gamma}_i^{(0)})^{2/3}}}, \end{aligned} \tag{7}$$

where $\delta_{(\hat{\lambda}_i^{(0)})} = \hat{\lambda}_i^{(0)}/\sqrt{1+(\hat{\lambda}_i^{(0)})^2}$, $i = 1,..,k$ and the sign of $\hat{\lambda}_i^{(0)}$ is the same of $\hat{\gamma}_i^{(0)}$. They are obtained by replacing $E(Y)$, $Var(Y)$ and $\gamma(Y)$ in (2) with their respective estimators in (6) and solving the resulting equations in μ_i, σ_i and λ_i. The initial values for the weights p_i are obtained as

$$(\hat{p}_1^{(0)}, \ldots, \hat{p}_k^{(0)}) \sim \text{Dirichlet}(1, \ldots, 1),$$

– a Dirichlet distribution with all parameters equal to 1 – that is, a uniform distribution over the unit simplex $\{(p_1, \ldots, p_k);\ p_i \geq 0,\ \sum_{i=1}^k p_i = 1\}$.

Method of Moments (MM): in this case the initial values are obtained using equations (7), but replacing $\hat{\xi}_i^{(0)}$, $\hat{\omega}_i^{(0)}$ and $\hat{\gamma}_i^{(0)}$ with the mean, variance and skewness coefficient of sub-sample i, respectively. Let n be the sample size and n_i be the size of sub-sample i. The initial values for the weights are given by

$$\hat{p}_i^{(0)} = \frac{n_i}{n}, \qquad i = 1,..,k.$$

To compare the methods, we generated samples from the FM-SN model with $k = 2$ and $k = 3$ components. For each fixed sample we obtained estimates of the parameters using the algorithm presented in section 4.1 initialized by each method proposed.

Another important issue is the degree of heterogeneity of the components. For $k = 2$ we considered the "moderately separated" ($2MS$), "well separated" ($2WS$) and "poorly separated" ($2PS$) cases. For $k = 3$ we considered the "two poorly separated and one well separated" ($3PWS$) and "the three well separated" ($3WWS$) cases. These degrees of heterogeneity were obtained informally, based on the location parameter values. The main reason to consider them as an important factor to our study is that the convergence of the EM algorithm is typically affected when the components overlap largely – see Park & Ozeki (2009) and the references herein.

Figures 1 and 2 show some histograms exemplifying these degrees of heterogeneity. In tables 1 and 2 are presented the parameters values used in the study.

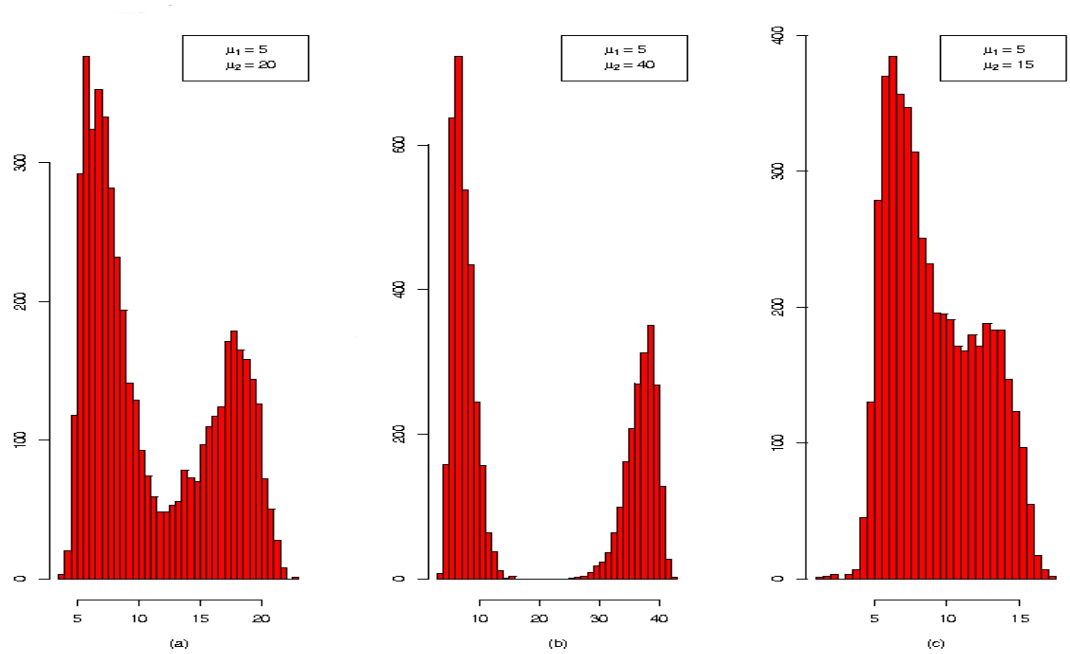

Figure 1: Histograms of FM-SN data: (a) $2MS$, (b) $2WS$ and (c) $2PS$.

Table 1: FM-SN parameters with $k = 2$.

Case	μ_1	μ_2	σ_1^2	σ_2^2	λ_1	λ_2	p_1	p_2
$2MS$	5	20	9	16	6	-4	0.6	0.4
$2WS$	5	40	9	16	6	-4	0.6	0.4
$2PS$	5	15	9	16	6	-4	0.6	0.4

All the computations were made using the R system (R Development Core Team, 2009). Sample sizes were fixed as $n = 500$, 1000, 5000 and 10000. For each combination of parameters and sample size, 500 samples from the FM-SN model were artificially generated. Then we computed the bias and mean squared

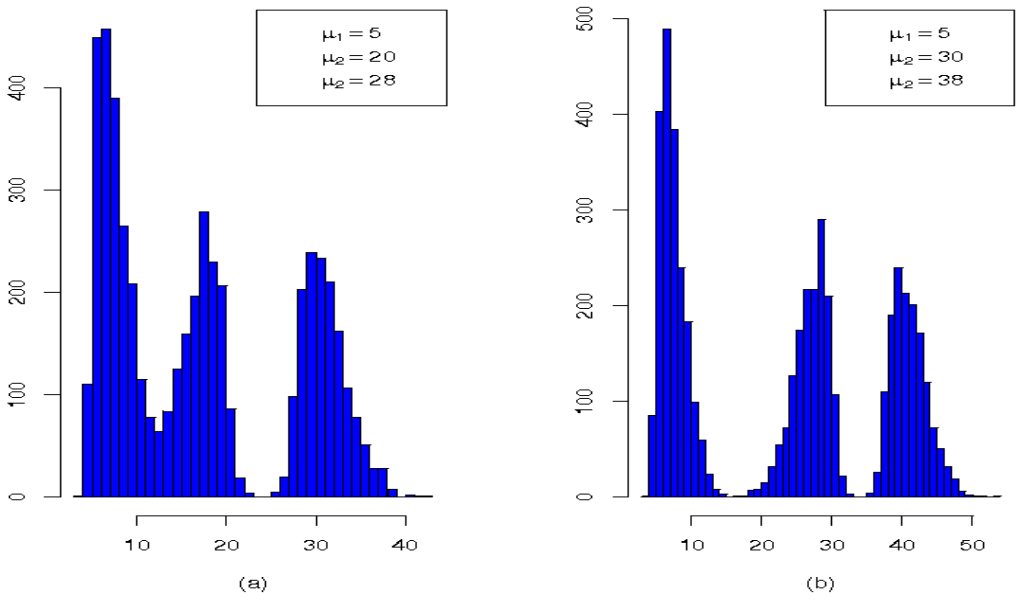

Figure 2: Histograms of FM-SN data: (a) $3PWS$ and (b) $3WWS$.

Table 2: FM-SN parameters with $k = 3$.

Case	μ_1	μ_2	μ_3	σ_1^2	σ_2^2	σ_3^2	λ_1	λ_2	λ_3	p_1	p_2	p_3
$3PWS$	5	20	28	9	16	16	6	-4	4	0.4	0.3	0.3
$3WWS$	5	30	38	9	16	16	6	-4	4	0.4	0.3	0.3

error (MSE) over all samples. For μ_i they are defined as

$$\text{bias} = \frac{1}{500}\sum_{j=1}^{500}\hat{\mu}_i^{(j)} - \mu_i \quad \text{and} \quad \text{MSE} = \frac{1}{500}\sum_{j=1}^{500}(\hat{\mu}_i^{(j)} - \mu_i)^2,$$

respectively, where $\hat{\mu}_i^{(j)}$ is the estimate of μ_i when the data is sample j. Definitions for the other parameters are obtained by analogy.

As a note about implementation, a expected consequence of the non-identifiability cited in section 4.2.2 is the permutation of the component labels when using the $k-$means method to perform an initial clustering of the data. It has serious implications when the main subject is the consistency of the estimates. To overcome this problem we adopted an order restriction on the initial values of the location parameters.

Tables 3 and 4 present, respectively, bias and MSE of the estimates in the $2MS$ case. With two moderately separated components and using the TVM and MM methods, the convergence of the estimates is evidenced, as we can conclude observing the decreasing values of bias and MSE when the sample size increases. They also show that the estimates of the weights p_i and of the location parameters μ_i have lower bias and MSE. On the other side, investigating the MSE values, we can note a different pattern of (slower) convergence to zero for the skewness parameters estimates. It is possibly due to well known inferential problems related to the skewness parameter (DiCiccio & Monti, 2004), suggesting the use of larger samples in order to attain the consistency property.

When we analyze the initialization methods performances, we can see that MM and TVM have a similar (good) one, presenting better results than the RVM method for all sample sizes and parameters. When using the RVM we can see that, in general, the absolute value of the biases of the estimates of σ_i^2 and λ_i are very large compared with that obtained when using the other methods. In general, according to our criteria, we can conclude that the MM method presented a satisfactory performance in all situations.

The bias and MSE of the estimates for the $2WS$ case are presented in tables 5 and 6, respectively. As in the $2MS$ case, their values decrease when the sample size increases (using the methods TVM and MM), although in a slower rate. Comparing the initialization methods, we can see again the poor performance of RVM, notably when estimating σ_i^2 and λ_i. The performances of MM and TVM are similar and satisfactory. Thus, we can say that, in general, the conclusions made for the $2MS$ case are still valid here.

We present the results for the $2PS$ case in tables 7 and 8. Bias and MSE are larger than in the $2MS$ and $2WS$ cases (for all sample sizes) when using MM and RVM. Also, the consistency of the estimates seems to be unsatisfactory, clearly not attained in the σ_i^2 and λ_i cases. According to the related literature, such drawbacks of the algorithm are expected when the population presents a remarkable homogeneity. An exception is made when the initial values are closer to the true parameter values see, for example, McLachlan & Peel (2000) and the above-mentioned work of Park & Ozeki (2009).

Results for the $3PWS$ case are showed in tables 9 and 10. It seems that consistency is achieved for $\hat{p}_i$, $\hat{\mu}_i$ and $\hat{\sigma}_i^2$, using TVM and MM. However, this is not the behavior in the $\hat{\lambda}_i$ case. This instability is common to all initialization methods, according to the MSE criterion. Using the RVM method we obtained, as before, larger values of bias and MSE.

Finally, tables 11 and 12 present the results for the $3PWW$ case. Concerning the estimates $\hat{p}_i$ and $\hat{\mu}_i$, very satisfactory results are obtained, with small values of bias and MSE when using TVM and MM. The values of MSE of $\hat{\sigma}_i^2$ exhibit a decreasing behavior when the sample size increases. On the other side, although we are in the well separated case, the values of bias and MSE of $\hat{\lambda}_i$ are larger, notably when using RVM as the initialization method.

Concluding this section, we can say that, as a general rule, MM and TVM methods presented very good similar results for the initialization of the EM algorithm. Considering TVM as the gold standard, this means that MM can be seen as a good alternative for real applications. If this condition is maintained, our study suggests that the consistency property holds for all EM estimates – maybe the convergence is slower for the scale parameter case – except for the skewness parameter, indicating that a sample size larger than 5000 is necessary to achieve consistency in this case. The study also suggests that the degree of heterogeneity of the population has a remarkable influence on the quality of the estimates.

5.2 Density Estimation Analysis

In this section we investigate the density estimation issue, that is, the point estimation of the parameter $\ell(\Theta)$, the log-likelihood evaluated at the true value of the parameter. We considered FM-SN models with two components and restricted ourselves to the cases $2MS$, $2WS$ and $2PS$, with sample sizes $n =$100, 500, 1000, 5000. The following measure was considered to compare the several methods of initialization

$$d_r(M) = \left| \frac{\ell(\Theta) - \ell_{(M)}(\hat{\Theta})}{\ell(\Theta)} \right| \times 100,$$

where $\ell_{(M)}(\hat{\Theta})$ is the log-likelihood evaluated at the EM estimate $\hat{\Theta}$, which was obtained using the initialization method M. According to this criterion, an initialization method M is better than M' if $d_r(M) < d_r(M')$. Obviously, we expect the lowest d_r values for the TVM method, the gold standard one.

Table 3: Bias of the estimates – $k = 2$, moderately separated ($2MS$)

n	Method	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{p}_1$	$\hat{p}_2$
500	TVM	0.01291666	-0.03518549	0.01661924	0.1847646	0.5253373	-0.4959544	-0.001556583	0.001556583
	RVM	0.9748264	-1.323073	-4.163036	-0.1236784	-1.867617	-3.058108	-0.005579744	0.005579744
	MM	0.01281975	-0.0419933	-0.1902714	0.3501801	0.5296119	-0.4772044	-0.001368731	0.001368731
1000	TVM	0.01460258	-0.02278766	-0.06984684	0.1100844	0.1304437	-0.1455234	-0.001570018	0.001570018
	RVM	1.072717	-1.295847	-4.469534	0.2355952	-2.595832	1.156491	-0.01132529	0.01132529
	MM	0.01437174	-0.02564698	-0.08612976	0.0923462	0.1344136	-0.1330872	-0.001409981	0.001409981
5000	TVM	0.001277386	0.0001436402	0.02193594	0.04569742	0.04218275	-0.04685099	-5.452673e-05	5.452673e-05
	RVM	0.974323	-1.002487	-4.040373	1.541311	-2.444670	0.7957775	-0.01203366	0.01203366
	MM	0.0009641692	-0.001388175	0.03096452	0.01537434	0.04726114	-0.03904657	6.324371e-05	-6.324371e-05
10000	TVM	-0.0005878508	0.0001176563	0.006640923	0.01797871	0.007833535	-0.02678607	0.0004952318	-0.0004952318
	RVM	1.003622	-1.067299	-4.154656	1.049366	-2.587216	0.9152174	-0.01212533	0.01212533
	MM	-0.000905076	-0.001093679	0.01476033	-0.007207103	0.01278570	-0.02045472	0.0005959484	-0.0005959484

Table 4: MSE of the estimates - $k = 2$, moderately separated ($2MS$)

n	Method	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{p}_1$	$\hat{p}_2$
500	TVM	0.01618735	0.1123241	2.609074	21.05012	16.89325	3.963948	0.0007687456	0.0007687456
	RVM	2.250586	4.266357	26.19319	78.74164	34.63418	6729.553	0.002770527	0.002770527
	MM	0.01627796	0.1324369	2.203446	18.58105	16.91288	3.977364	0.0007714305	0.0007714305
1000	TVM	0.007352563	0.04381211	1.320318	9.830591	1.570595	1.007072	0.0003967391	0.0003967391
	RVM	2.366453	4.088751	27.80727	70.83071	19.36655	9.247467	0.001719727	0.001719727
	MM	0.00737845	0.04473454	1.246387	9.680204	1.569277	1.015799	0.0003981356	0.0003981356
5000	TVM	0.001581752	0.007309233	0.2642777	2.0395	0.2449436	0.1845891	8.440737e-05	8.440737e-05
	RVM	2.165632	3.19515	26.12405	70.86125	17.22613	7.42687	0.001879775	0.001879775
	MM	0.001590433	0.00742466	0.2694587	2.076536	0.2465352	0.1866866	8.502204e-05	8.502204e-05
10000	TVM	0.0008475006	0.003877835	0.1238345	1.109195	0.1206761	0.08556848	4.341359e-05	4.341359e-05
	RVM	2.183152	3.346971	25.85164	69.09116	17.61139	7.636557	0.001412294	0.001412294
	MM	0.0008533806	0.003948246	0.1269059	1.134417	0.1216638	0.086953	4.378802e-05	4.378802e-05

Table 5: Bias of the estimates - $k = 2$, well separated ($2WS$)

n	Method	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{p}_1$	$\hat{p}_2$
500	TVM	-0.001342122	-0.0008946654	0.07087477	0.07346184	0.5476595	-0.3693064	0.001103623	-0.001103623
	RVM	1.145212	-1.609425	-3.282642	-4.080945	-2.568067	1.816600	0.001102224	-0.001102224
	MM	-0.001440393	0.001164155	0.07148351	0.1861907	0.8121083	-0.3944749	0.001069988	-0.001069988
1000	TVM	0.002673775	-0.00174725	0.004734621	0.005284774	0.1993675	-0.1665410	0.0004079317	-0.0004079317
	RVM	1.205876	-1.759603	-3.485664	-4.571723	-2.863520	2.133156	0.0004069826	-0.0004069826
	MM	0.002777323	-0.001922072	0.005182833	0.005432662	0.2025558	-0.1678265	0.0004079617	-0.0004079617
5000	TVM	-0.001089158	0.002580446	0.006617728	0.01874684	0.03130216	-0.03101032	-0.0007086542	0.0007086542
	RVM	1.298881	-1.804116	-3.685889	-4.807072	-3.238162	2.286944	-0.0007111243	0.0007111243
	MM	-0.001046705	0.002517380	0.00653431	0.01843972	0.0314678	-0.03095968	-0.0007086519	0.0007086519
10000	TVM	0.000376472	-0.001738574	0.003361113	-0.02059716	0.02255472	-0.01259483	-0.0002449717	0.0002449717
	RVM	1.416778	-1.773241	-3.851791	-4.859173	-3.541599	2.253168	-0.0002485395	0.0002485395
	MM	0.0003573312	-0.001752626	0.003659679	-0.02058202	0.02314823	-0.01267402	-0.0002449714	0.0002449714

Table 6: MSE of the estimates - $k = 2$, well separated ($2WS$)

n	Method	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{p}_1$	$\hat{p}_2$
500	TVM	0.01161586	0.0616181	0.8216382	5.17506	3.356231	1.954443	0.0004420863	0.0004420863
	RVM	2.781579	5.213882	16.54363	37.86230	18.38287	9.24199	0.0004420563	0.0004420563
	MM	0.01195556	0.06452464	0.8422866	10.64528	37.27305	2.199025	0.0004437283	0.0004437283
1000	TVM	0.006644214	0.0298738	0.3802331	2.291371	1.191168	0.7158823	0.0002491188	0.0002491188
	RVM	2.924958	5.630983	17.44514	40.69785	18.54422	9.177224	0.0002491342	0.0002491342
	MM	0.006828364	0.03030411	0.3902985	2.321045	1.230409	0.7301051	0.0002491197	0.0002491197
5000	TVM	0.001187191	0.005883913	0.07796818	0.4901996	0.2062550	0.1021211	4.950553e-05	4.950553e-05
	RVM	3.114875	5.686154	18.60732	41.63566	19.59477	9.201493	4.950964e-05	4.950964e-05
	MM	0.001215798	0.005946889	0.07978034	0.4948037	0.2115759	0.1033688	4.950552e-05	4.950552e-05
10000	TVM	0.0005889338	0.002968532	0.03659178	0.2307873	0.09602311	0.05295109	2.267009e-05	2.267009e-05
	RVM	3.390618	5.568737	19.85308	42.98142	21.32373	9.037181	2.266685e-05	2.266685e-05
	MM	0.0006025367	0.003005308	0.0375859	0.2332507	0.09867073	0.05370464	2.267008e-05	2.267008e-05

Table 7: Bias of the estimates - $k = 2$, poorly separated ($2PS$)

n	Method	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{p}_1$	$\hat{p}_2$
500	TVM	0.01130208	-0.1029151	0.8222888	-2.537293	0.7948747	-0.2598434	0.03027712	-0.03027712
	RVM	0.7377362	-1.413347	-4.782077	-3.800659	-1.60642	-1.592736	0.02066073	-0.02066073
	MM	0.5760954	-2.337444	-6.526634	-5.732598	-1.368408	3.878645	0.0914515	-0.0914515
1000	TVM	0.005135277	-0.03102672	0.512132	-1.235290	0.2332226	-0.1118209	0.01814267	-0.01814267
	RVM	0.6450937	-1.351040	-4.453074	-3.75638	-2.403474	2.001093	0.01954508	-0.01954508
	MM	0.5326432	-2.340892	-6.686511	-5.68292	-1.799719	4.054375	0.09628267	-0.09628267
5000	TVM	0.001401114	0.0008878685	0.0624029	-0.1614783	0.05408275	-0.03992003	0.002785076	-0.002785076
	RVM	0.465385	-1.019837	-3.645296	-2.647723	-1.8646	1.728286	0.01635918	-0.01635918
	MM	0.4132328	-2.313775	-6.531423	-5.741616	-2.127802	4.041619	0.0904191	-0.0904191
10000	TVM	0.001821830	0.0001505038	0.03025559	-0.1140996	0.006581757	-0.000269398	0.001758447	-0.001758447
	RVM	0.5061236	-1.031940	-3.692628	-2.634238	-2.105805	1.677271	0.00775284	-0.00775284
	MM	0.2871329	-2.205388	-6.481012	-5.305207	-1.909871	4.050629	0.1025349	-0.1025349

Table 8: MSE of the estimates - $k = 2$, poorly separated ($2PS$)

n	Method	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{p}_1$	$\hat{p}_2$
500	TVM	0.016901	0.153715	8.104147	31.83997	13.7952	10.93927	0.006690502	0.006690502
	RVM	1.46428	4.313297	34.95844	38.20203	205.0405	5609.656	0.01070596	0.01070596
	MM	1.258757	6.295013	44.2611	52.19197	11.58125	16.24227	0.01714522	0.01714522
1000	TVM	0.008539055	0.06420256	4.443361	14.55311	2.115562	1.764568	0.003219140	0.003219140
	RVM	1.290693	3.791282	32.55702	31.74903	14.69199	8.175651	0.008868588	0.008868588
	MM	1.035179	5.995885	45.38505	48.8999	10.09152	16.46602	0.01735244	0.01735244
5000	TVM	0.001800551	0.009271334	0.6618201	1.432464	0.374085	0.1521287	0.0004051785	0.0004051785
	RVM	0.8732212	2.477904	25.59724	19.83352	9.81654	6.563971	0.007593878	0.007593878
	MM	0.6356428	5.821094	43.6267	43.92463	8.171784	16.36373	0.01630810	0.01630810
10000	TVM	0.0007779437	0.005205021	0.3124418	0.7374107	0.1667618	0.08343909	0.0001882813	0.0001882813
	RVM	1.161408	2.757048	26.17153	20.13907	11.10304	6.230668	0.007534106	0.007534106
	MM	0.3658037	5.207123	42.81331	35.69968	5.819666	16.40898	0.01726078	0.01726078

Table 9: Bias of the estimates - $k = 3$, two poorly separated and one well separated ($3PWS$)

n	Method	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\sigma}_3^2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$
500	TVM	0.0124259	-0.058538	0.078976	0.075256	-0.095745	-0.462774	-2.038312	18.06389	-8.24194	-0.099099	0.09838	0.0007106
	RVM	0.839994	-1.62511	2.28647	-4.75611	-7.51127	1.389074	-0.32077	-3.716019	-3.224084	0.0096819	-0.025417	0.015735
	MM	0.012744	-0.099691	0.078937	-0.33422	-2.31877	1.94964	8.162306	-0.17509	-0.16034	-0.00051179	-0.014921	0.015433
1000	TVM	0.0009163	-0.013053	0.02771	0.236455	-0.0300267	-0.24516	-1.95914	10.48063	-8.181147	-0.1000959	0.101737	-0.001641
	RVM	0.841724	-1.586819	2.21955	-4.75936	-7.31627	0.695945	-1.03888	0.90239	-3.24466	0.005568	-0.017829	0.01226
	MM	0.0006666	-0.015744	0.027732	0.09991	-1.57201	1.40084	0.54447	-0.16887	-0.017631	0.0018592	-0.011644	0.0097849
5000	TVM	-0.0024898	-0.0071647	-0.0016895	0.063075	-0.051032	-0.021537	-1.963292	10.0949	-8.01175	-0.100198	0.10076	-0.00056612
	RVM	0.868112	-1.53414	2.20497	-4.7778	-7.30824	0.654812	-1.33676	1.72577	-3.25474	0.0032579	-0.013555	0.010297
	MM	-0.0029399	-0.0090219	-0.001699	0.075584	-0.81122	0.702509	0.10232	-0.0024549	0.036783	0.0008703	-0.0051312	0.0042608
10000	TVM	0.0007229	0.0004312	0.005227	-0.001277	0.082034	-0.0341	-2.00442	10.0445	-8.0362	-0.10048	0.10018	0.0003056
	RVM	0.7653	-1.4258	2.4471	-4.4972	-7.1377	0.48926	-1.3367	1.5905	-3.3251	0.00494	-0.01525	0.01030
	MM	0.000272	-0.001002	0.0052226	0.010146	-0.4979	0.51546	0.051814	-0.02867	-0.004384	0.0002749	-0.00335	0.003078

Table 10: MSE of the estimates - $k = 3$, two poorly separated and one well separated ($3PWS$)

n	Method	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\sigma}_3^2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$
500	TVM	0.0256408	0.126759	0.086256	4.111284	25.73358	6.28545	5.37295	24575.55	70.43225	0.0102611	0.0104158	0.00066346
	RVM	2.094829	5.42932	8.410817	27.17296	69.49805	181.8071	32.08280	6671.365	12.99524	0.0039025	0.0034232	0.00085144
	MM	0.0265016	0.266846	0.086721	2.900603	10.97478	16.8237	24090.79	2.73729	1.06819	0.00067316	0.00051696	0.00053893
1000	TVM	0.0123157	0.0568707	2.43316	13.12828	3.05556	4.53387	11.2794	68.3113	60.10207	0.010696	0.00032769	0.0003283
	RVM	1.951414	5.05061	7.466443	26.81183	66.32812	97.39908	13.45374	330.0578	13.07518	0.0017043	0.0014459	0.00044176
	MM	0.0123607	0.0574984	0.0364335	1.89177	6.05455	8.69406	3.052947	1.418105	0.601988	0.00034945	0.00027203	0.0002453
5000	TVM	0.0023085	0.0111053	0.0076361	0.43256	2.92208	0.621509	3.98858	102.270	64.4356	0.010079	0.010225	6.525753e-05
	RVM	1.97832	4.88315	7.39087	27.06606	65.756	79.2677	12.62484	9.90278	13.0587	0.001635	0.0010396	0.0004439
	MM	0.0023209	0.011292	0.0076426	0.44148	1.82228	1.7738	0.3738	0.250619	0.13556	7.335385e-05	5.681646e-05	4.850018e-05
10000	TVM	0.00110043	0.0050401	0.00410308	0.18212	1.3066	0.31966	4.0874	101.072	64.692	0.010118	0.010068	2.806801e-05
	RVM	1.74505	4.82237	9.016929	24.7309	64.0593	80.27004	11.1823	8.8820	13.332	0.00189	0.001888	0.0004551
	MM	0.001109	0.00514	0.004107	0.1873	0.7276	0.9279	0.1849	0.11501	0.069822	3.230155e-05	2.545253e-05	2.42436e-05

Table 11: Bias of the estimates - $k = 3$, the three well separated ($3WWS$)

n	Method	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\sigma}_3^2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$
500	TVM	0.002629	-0.030459	0.174474	0.02879	-0.0101093	-1.095813	-2.661707	23.33088	-8.294356	-0.0986793	0.0986485	3.087517e-05
	RVM	1.128429	-1.892981	2.323337	-4.006363	-7.90329	-1.37325	5.449758	0.727331	-3.472227	0.0034786	-0.0170508	0.0135722
	MM	0.002049	-0.073056	0.16449	-0.000293	-1.9645	0.941228	7.4967	-0.28628	-0.68906	-0.0008928	-0.01298	0.01387
1000	TVM	0.003582	-0.01047	0.09973	0.014815	-0.0449	-0.6072	-2.418028	10.31084	-8.15506	-0.09904	0.099609	-0.000569
	RVM	1.14596	-1.75138	2.246543	-3.825814	-7.67675	-1.014807	-2.106635	1.932	-3.3737	0.003889	-0.01303	0.0091405
	MM	0.003149	-0.01126	0.100995	0.020765	-1.33552	0.66874	0.3252	-0.153007	-0.42507	-0.000388	-0.0003888	-0.00928
5000	TVM	0.000879	-0.0031127	0.02638	-0.003719	-0.0240517	-0.16356	-2.113001	10.0626	-8.03146	-0.09987	0.100255	-0.000378
	RVM	1.34845	-2.1397	2.3268	-4.1174	-8.091005	0.6890144	-2.47597	2.40248	-3.49824	0.009667	-0.01142	0.001753
	MM	0.000467	-0.00347	0.02667	0.0001683	-0.55819	0.36519	0.06887	-0.030102	-0.11407	0.000255	-0.004465	0.0042097
10000	TVM	0.004409	0.0003288	0.01253	-0.02048	0.004985	-0.066435	-2.061703	10.01159	-8.01837	-0.099	0.10033	-0.0003497
	RVM	1.3549	-2.1503	2.406146	-4.1256	-7.93745	-0.763128	-2.57214	2.373008	-3.57646	0.007488	-0.009484	0.001996
	MM	0.0040304	9.503394e-05	0.01268	-0.017105	-0.3395	0.274816	0.016875	-0.01749	-0.06225	0.0003327	-0.003137	0.002804

Table 12: MSE of the estimates - $k = 3$, the three well separated ($3WWS$)

n	Method	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\sigma}_3^2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$
500	TVM	0.02284	0.08425	0.10313	1.30092	6.99418	6.3296	7.496	39122.85	70.743	0.010132	0.0102043	0.0004017
	RVM	3.0688	6.84256	8.3173	19.5307	71.4026	260.9848	25622.18	335.8204	13.6056	0.002224	0.001829	0.000625
	MM	0.023629	0.9625	0.18175	1.3752	7.4997	9.696	19921.21	2.09006	0.8895	0.000554	0.000496	0.000424
1000	TVM	0.009632	0.03823	0.048836	0.72127	3.352608	3.2399	6.13927	108.1138	67.279	0.0100098	0.010195	0.0002258
	RVM	3.0637	6.22514	7.78637	18.2895	69.6556	328.1986	13.1841	9.435052	13.32669	0.0019764	0.001447	0.00035106
	MM	0.00987	0.0386	0.049507	0.739038	3.86043	2.7949	1.9419	0.80634	0.47407	0.0002737	0.00021335	0.0002143
5000	TVM	0.001883	0.00792	0.00754	0.13726	0.70517	0.64068	4.5479	101.544	64.6403	0.010019	0.010098	3.962665e-05
	RVM	4.3828	8.95947	8.4774	19.954	74.574	762.535	14.1316	9.9473	14.04390	0.0040459	0.002173	0.0004728
	MM	0.001926	0.0080093	0.00765	0.1402	0.734012	0.6240084	0.2994	0.13834	0.097003	4.702413e-05	4.330964e-05	4.001327e-05
10000	TVM	0.00100849	0.003956	0.003519	0.06617	0.3612407	0.30119	4.3027	100.3758	64.37698	0.0100169	0.010093	2.122882e-05
	RVM	4.4666	10.10709	7.9793	20.12889	73.59039	419.6245	14.63398	9.8323	14.4596	0.002315	0.001867	0.0001507
	MM	0.0010254	0.00399	0.00355	0.06758	0.33538	0.33285	0.1479	0.0839	0.05650	2.672532e-05	2.090253e-05	2.074695e-05

Table 13 presents the means and standard deviations of d_r over 500 samples in the $2MS$ case. For TVM and MM, we can see that these values decrease when the sample size increases. The RVM presented the lowest mean value only when $n = 100$. For larger sample sizes its performance was inferior. In addition, for all sample sizes considered, it presented the largest standard deviation.

Table 13: Means and standard deviations ($\times 10^{-4}$) of d_r. $2MS$ case.

	Method					
	TVM		**RVM**		**MM**	
n	Mean	SD	Mean	SD	Mean	SD
100	1.50	15.90	1.20	21.00	1.50	15.96
500	0.27	2.87	0.87	15.13	0.27	2.87
1000	0.13	1.44	0.84	15.71	0.13	1.44
5000	0.03	0.29	0.88	17.16	0.03	0.29

Table 14: Means and standard deviations ($\times 10^{-4}$) of d_r. $2WS$ case.

	Method					
	TVM		**RVM**		**MM**	
n	Mean	SD	Mean	SD	Mean	SD
100	1.48	15.33	1.21	17.32	1.47	15.05
500	0.25	2.58	1.23	19.09	0.25	2.58
1000	0.12	1.26	1.40	20.97	0.12	1.26
5000	0.03	0.25	1.60	22.72	0.03	0.25

For the $2WS$ case we have again MM and TVM presenting a similar performance – see Table 14, with a clear monotonicity when the sample size increases. Excepting the RVM, all methods have smaller standard deviations in this well separated case, maybe due to the improved ability of the k-means method to cluster this heterogeneous data set.

Table 15: Means and standard deviations ($\times 10^{-4}$) of d_r. $2PS$ case.

	Method					
	TVM		**RVM**		**MM**	
n	Mean	SD	Mean	SD	Mean	SD
100	1.70	18.53	1.29	22.52	1.30	19.19
500	0.30	3.18	0.44	6.54	0.36	6.99
1000	0.15	1.68	0.43	8.53	0.40	7.83
5000	0.02	0.21	0.39	8.76	0.48	5.61

Results for the $2PS$ case are shown in Table 15. In this case the MM method has an opposed pattern. We do not observe a monotone behavior for d_r and the standard errors are larger than that presented in the $2MS$ and $2WS$ cases. It is reasonable to suppose that the homogeneous structure of the data is negatively influencing the ability of the k-means method to cluster it.

The main message is that the MM method seems to be suitable when we are interested in the estimation of the likelihood values, with some caution when the population is highly homogeneous.

5.3 Number of Iterations

It is well known that one of the major drawbacks of the EM algorithm is the slow convergence. The problem become more serious when there is a bad choice of the starting values (McLachlan & Krishnan, 2008). Consequently, an important issue is the investigation of the number of iterations necessary to the convergence of the algorithm. Here we consider the $2MS$, $2WS$ and $2PS$ cases, $k = 2$ and $n =$ 100, 500, 1000, 5000. For each combination of parameters and sample size, 500 samples were generated artificially, the number of iterations was observed and the means and standard deviations of this quantity were computed. The simulations results are reported in Tables 16, 17, and 18.

Table 16: Means and standard deviations of number of iterations. $2MS$ case

	Method					
	TVM		**RVM**		**MM**	
n	Mean	SD	Mean	SD	Mean	SD
100	2359.46	121.13	1612.09	100.79	2293.78	120.02
500	551.10	15.45	640.23	18.41	573.62	15.26
1000	468.65	8.17	666.58	11.06	501.72	7.67
5000	449.28	6.27	845.64	8.66	501.32	5.49

Table 17: Means and standard deviations of number of iterations. $2WS$ case.

	Method					
	TVM		**RVM**		**MM**	
n	Mean	SD	Mean	SD	Mean	SD
100	1519.10	105.98	942.32	79.56	1520.42	105.26
500	209.42	4.44	369.35	4.26	238.49	5.02
1000	189.85	2.60	403.22	3.88	220.24	2.68
5000	176.04	1.72	577.10	7.02	215.61	2.06

Results suggest that in the moderately and well separated cases (Tables 16 and 17, respectively), and using TVM and MM, the mean number of iterations decreases as the sample size increases, but the same is not true when RVM is adopted as the initialization method.

The $2PS$ case is illustrated in Table 18. As expected, we have a poor behavior possibly due to the population homogeneity, as we commented before. An interesting fact is that, comparing with MM, RVM has a smaller mean number of iterations.

6 A Simulation Study of Model Choice

Someone can argue that an arbitrary density can always be approximated by a finite mixture of normal distributions, see McLachlan & Peel (2000, Chapter 1), for example. However, it is not enough to increase the number of components in order to obtain a suitable fit. The main question is how to achieve this using

Table 18: Means and standard deviations of number of iterations. $2PS$ case.

	Method					
	TVM		**RVM**		**MM**	
n	Mean	SD	Mean	SD	Mean	SD
100	2470.30	105.94	1539.45	93.29	1713.71	90.74
500	1061.02	39.90	1090.39	38.42	1095.16	23.07
1000	900.66	30.85	1063.38	23.97	1135.90	15.85
5000	700.77	12.53	1274.86	19.16	1499.98	15.90

the smallest number of mixture components without overfit the data. One possible approach is to use some criteria function and compute

$$\hat{k} = \arg\min_k \{C(\hat{\Theta}_{(k)}),\ k \in \{k_{min}, ..., k_{max}\}\},$$

where $C(\hat{\Theta}_{(k)})$ is the criteria function evaluated at the EM estimate $\hat{\Theta}_{(k)}$, obtained by modeling the data using the FM-SN model with k components, and k_{min} and k_{max} are fixed positive integers.

Our main purpose in this section is to investigate the ability of some classical criteria to estimate the correct number of mixture components. We consider the Akaike Information Criterion (AIC) (Akaike, 1974), the Bayesian Information Criterion (BIC) (Schwarz, 1978), the Efficient Determination Criterion (EDC) (Bai *et al.*, 1989) and the Integrated Completed Likelihood Criterion (ICL) (Biernacki *et al.*, 2000). AIC, BIC and EDC have the form

$$-2\,\ell(\hat{\Theta}) + d_k\, c_n,$$

where $\ell(\cdot)$ is the actual log-likelihood, d_k is the number of free parameters that have to be estimated under the model with k components and the penalty term c_n is a convenient sequence of positive numbers. We have $c_n = 2$ for AIC and $c_n = \log(n)$ for BIC. For the EDC criterion, c_n is chosen so that it satisfies the conditions

$$c_n/n \to 0 \quad \text{and} \quad c_n/(\log\log(n)) \to 0$$

when $n \to \infty$. Here we compare the following alternatives

$$c_n = 0.2\sqrt{n}, \quad c_n = 0.2\log(n), \quad c_n = 0.2n/\log(n), \quad \text{and} \quad c_n = 0.5\sqrt{n}.$$

The ICL is defined as

$$-2\,\ell^*(\hat{\Theta}) + d_k \log(n),$$

where $\ell^*(\cdot)$ is the integrated log-likelihood of the sample and the indicator latent variables , given by

$$\ell^*(\hat{\Theta}) = \sum_{i=1}^{k} \sum_{j \in C_i} \log(\hat{p}_i \mathrm{SN}(y_j|\hat{\theta}_i)),$$

where C_i is a set of indexes defined as: j belongs to C_i if, and only if, the observation y_j is allocated to component i by the following clustering process: after the FM-SN model with k components was fitted using the EM algorithm we obtain the estimate of the posterior probability that an observation y_i belongs to the jth component of the mixture, $\hat{z}_{ij}$ – see equation (5). If $q = \arg\max_j\{\hat{z}_{ij}\}$ we allocate y_i to the component q.

In this study we simulated samples of the FM-SN model with $k = 3$, $p_1 = p_2 = p_3 = 1/3$, $\mu_1 = 5$, $\mu_2 = 20$, $\mu_3 = 28$, $\sigma_1^2 = 9$, $\sigma_2^2 = 16$, $\sigma_3^2 = 16$, $\lambda_1 = 6$, $\lambda_2 = -4$ and $\lambda_3 = 4$, and considered the sample sizes $n =$ 200, 300, 500,1000, 5000. Figure 3 shows a typical sample of size 1000 following this specified setup.

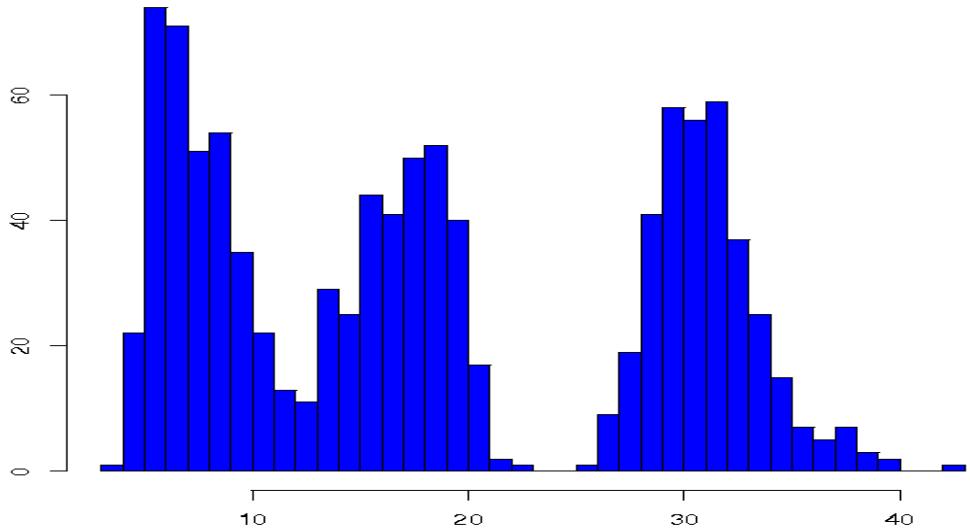

Figure 3: Histogram of a FM-SN sample with $k = 3$ and $n = 1000$

For each generated sample (with fixed number of 3 components) we fitted the FM-SN model with $k = 2$, $k = 3$ and $k = 4$, using the EM algorithm initialized by the method of moments. For each fitted model the criteria AIC, BIC, ICL and EDC were computed. We repeated this procedure 500 times and obtained the percentage of times some given criterion chooses the correct number of components. The results are reported in Table 19.

Table 19: Percentage of times the true model (with $k = 3$) is chosen using different criteria

n	**AIC**	**BIC**	**ICL**	**EDC**			
				$c_n = 0.2\log(n)$	$c_n = 0.2\sqrt{n}$	$c_n = 0.2n/\log(n)$	$c_n = 0.5\sqrt{n}$
200	94.2	99.2	99.2	77.8	98.4	99.4	99.4
300	94.0	98.8	98.8	78.2	98.4	98.8	98.8
500	95.8	99.8	99.8	86.4	99.8	99.8	99.8
1000	96.2	100.0	100.0	88.5	100.0	100.0	100.0
5000	95.6	100.0	100.0	92.8	100.0	100.0	100.0

We can see that BIC and ICL have a better performance than AIC for all sample sizes. Except for AIC, the rates presented an increasing behavior when the sample size increases. This possible drawback of AIC may be due to the fact that its definition does not take into account the sample size in its penalty term. Results for BIC and ICL were similar, while EDC showed some dependence on the term c_n. In general, we can say that BIC and ICL have equivalent abilities to choose the correct number of components and that, depending on the choice of c_n, ICL can be not as good as AIC or better than ICL and BIC.

7 Applications to Real Data Sets

Here we apply the modeling approach suggested in the previous sections to the analysis of two real data sets. That is, we model data using a FM-SN model, estimating parameters through the proposed EM-type algorithm, with initial values obtained by the method of moments and random values method and fixing as stopping rule $|\ell(\boldsymbol{\Theta}^{(k+1)}) - \ell(\boldsymbol{\Theta}^{(k)})| < 10^{-5}$.

7.1 GDP data set

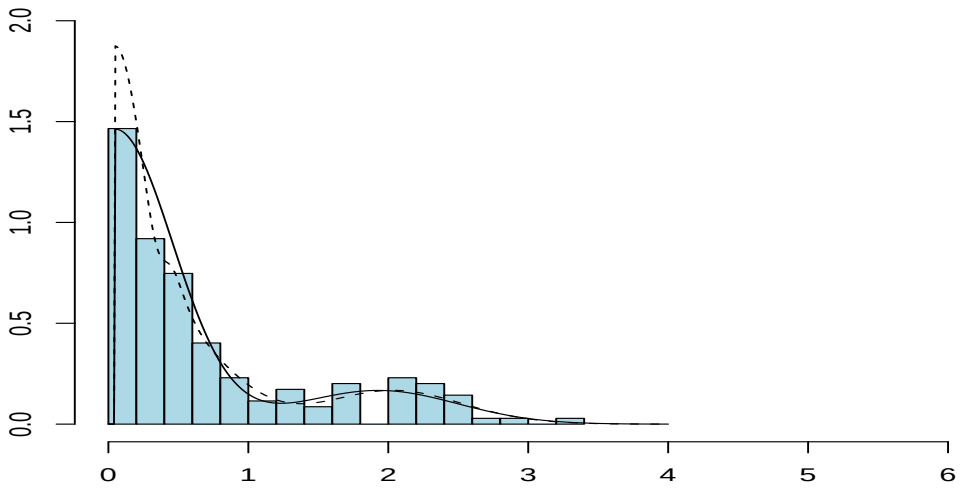

Figure 4: Histogram of the GDP data set and fitted FM-SN densities with 2 (solid line) and 3 (dotted line) components

In this application we consider observations from the gross domestic product (GDP) per capita (PPP US$) in 1998 from 174 countries. This data set appeared before in Dias & Wedel (2004) where, before fitting a mixture of two normal distributions to the data, was made a logarithmic transformation, in order to eliminate the clear strong skewness. In figure 4 we have a histogram of the data, suggesting a bimodal or trimodal pattern. Thus, a mixture of 2 or 3 SN distributions seems to be a natural candidate to model this data, avoiding a possibly unnecessary data transformation. For this application, each observation in the original data set was divided by 10^3.

Table 20 presents the initial values obtained by the MM and RVM methods. Table 21 shows the EM estimates. We can verify that, besides an apparent label switching, there are differences between starting points and EM estimates, which are more marked for the skewness parameter λ_i. In this case we have a small sample, and some care is needed when interpreting this results, because of the deficiencies of the algorithm in estimating this parameter, commented in the previous sections.

As in Lin *et al.* (2007a), we applied the Kolmogorov-Smirnov (K-S) test to evaluate the goodness of fit under different methods of initialization of the algorithm. The $p-$ values are presented in Table 22 and indicate a better fit using the model with 3 components, regardless of the initialization method.

Table 20: GDP data: initial values obtained by MM and RVM methods

		Method	
Parameter	k	RVM	MM
μ_i	2	(1.8365; 0.3911)	(0.0047; 1.6078)
	3	(2.4609; 0.9195; 0.0903)	(1.7354; 0.6208; 0.0896)
σ_i^2	2	(0.3617; 0.1542)	(0.2168; 0.4216)
	3	(0.0445; 0.0660; 0.0007)	(0.3689; 0.1730; 0.0522)
λ_i	2	(1.0619; -3.9085)	(11.5960; 1.5634)
	3	(-2.2742; -3.2476; 0.6745)	(3.3291; 2.4475; 2.1934)
p_i	2	(0.5419; 0.4580)	(0.7873; 0.2126)
	3	(0.3566; 0.2274; 0.4158)	(0.1724; 0.2068; 0.6206)
$\ell(\Theta)$	2	-260.8486	-106.5715
	3	-417.0599	-111.6361

Table 21: GDP data: EM estimates

		Method	
Parameter	k	RVM	MM
μ_i	2	(1.4124; 0.3043)	(0.0452; 1.8755)
	3	(1.9950; 0.4773; 0.0456)	(1.7915; 0.3968; 0.0454)
σ_i^2	2	(0.5620; 0.0380)	(0.2056; 0.2897)
	3	(0.4937; 0.0370; 0.0137)	(0.3102; 0.1840; 0.0521)
λ_i	2	(0.2926; -0.0665)	(2387.4927; 0.2941)
	3	(-0.4830; -0.0053; 1825.5701)	(0.7189; 7.4099; 2127.1691)
p_i	2	(0.3388; 0.6611)	(0.7799; 0.2200)
	3	(0.2857; 0.3813; 0.3329)	(0.2059; 0.2565; 0.5375)
$\ell(\Theta)$	2	-123.0055	-95.6176
	3	-94.4735	-92.0709

Fixing MM as the initialization method, we computed the model choice criteria AIC, BIC and EDC (with $c_n = 0.2\sqrt{n}$) for the models with 2 and 3 components. Table 23 shows the results, besides the log-likelihood evaluated at the EM estimates and the number of iterations needed to convergence. Although the model with 3 components has a smaller log-likelihood value, it has been penalized by the criteria because of the higher number of parameters. Surprisingly, the model with 2 components was rejected using the K-S test.

Figure 4 also presents the plug-in densities – that is, the FM-SN density with 2 and 3 components in which the parameters have been replaced by the EM estimates – superimposed on a single set of coordinate axes. Based on the graphical visualization, it appears that the model with 2 components has a quite reasonable and better fit.

Table 22: GDP data: Kolmogorov-Smirnov test

	2 components		3 components	
	RVM	MM	RVM	MM
K-S statistic	0.0687	0.0966	0.0365	0.0396
p-value	0.0250	0.0190	0.1940	0.1820

Table 23: GDP data: criteria values and number of iterations

k	$\ell(\hat{\Theta})$	AIC	BIC	EDC	Number of iterations
2	−94.95	205.24	227.35	214.32	7960
3	−92.08	206.15	240.90	220.43	7164

7.2 Old Faithful Geyser data set

Now we consider the famous Old Faithful Geyser data taken from Silverman (1986), consisting of 272 univariate measurements of eruptions lengths (in minutes) of the Old Faithful Geyser in Yellowstone National Park, Wyoming, USA. It has been analyzed by many authors, like Lin *et al.* (2007b), fitting FM-SN and FM-NOR models with two components, for example.

The histogram of this data, showed in figure 5, clearly suggests the fit of a bimodal distribution with two heterogenous asymmetric sub-populations. We made an analysis similar to that in the previous example, fitting SN models with 2 and 3 components. Results are displayed in tables 24, 25 and 26. In this case the

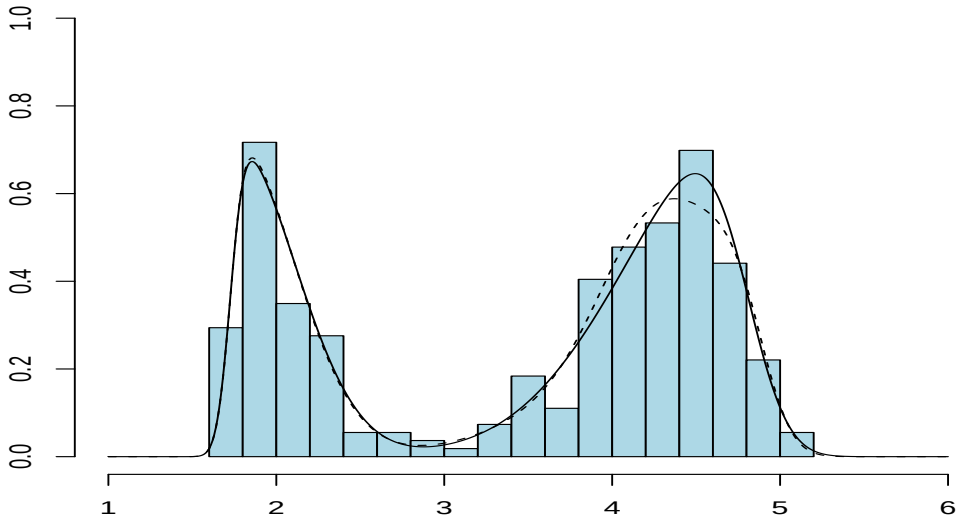

Figure 5: Histogram of the Old Faithful Geyser data set and fitted FM-SN densities with 2 (solid line) and 3 (dotted line) components

K-S test and the criteria agree on the choice of the model with 2 components.

Table 24: Old Faithful data: initial values obtained by MM and RVM methods

		Method	
Parameter	k	RVM	MM
μ_i	2	(4.7851; 2.8597)	(4.6874; 1.6713)
	3	(5.7732; 3.5815; 4.5693)	(4.4258; 4.7396; 4.6198)
σ_i^2	2	(0.0801; 0.0571)	(0.3125; 0.2236)
	3	(1.4072; 1.4842; 0.8971)	(2.3636; 2.5202; 2.7715)
λ_i	2	(1.4607; -1.5402)	(-1.7855; 122.9431)
	3	(-3.4674; -3.0006; -2.8209	(-1.4248; -2.3205; -1.8137)
p_i	2	(0.7718; 0.2281)	(0.6397; 0.3602)
	3	(0.2499; 0.1149; 0.6351)	(0.3308; 0.3566; 0.3125)
$\ell(\Theta)$	2	-1886.2170	-274.9906
	3	-3228.8101	-411.5698

Table 25: Old Faithful data: EM estimates

		Method	
Parameters	k	RVM	MM
μ_i	2	(4.8193; 2.0185)	(4.7994; 1.7264)
	3	(5.1021; 1.9993; 4.7367)	(2.0002; 4.9903; 4.6653)
σ_i^2	2	(0.5333; 0.0477)	(0.4688; 0.1451)
	3	(1.4799; 0.0444; 0.3476)	(0.0442; 1.0963; 0.2489)
λ_i	2	(-3.9733; -0.0731)	(-3.4798; 5.8263
	3	(-1.6523e+03; -3.4013e-03; -3.6525e+00)	(-0.0109; -10.2598; -2.3059)
p_i	2	(0.6584; 0.3415)	(0.6511; 0.3488)
	3	(0.1883; 0.3340; 0.4776)	(0.3344; 0.3005; 0.3650)
$\ell(\Theta)$	2	-226.4929	-257.5662
	3	-264.4344	-265.2770

8 Final Conclusion

In this work we presented a simulation study in order to investigate the performance of the EM algorithm for maximum likelihood estimation in finite mixtures of skew-normal distributions with component specific parameters. The results show that the algorithm produces quite reasonable estimates, in the sense of consistency and the total number of iterations, when using the method of moments to obtain the starting points. The study also suggested that the random initialization method used is not a reasonable procedure. When the EM estimates were used to compute some model choice criteria, namely the AIC, BIC, ICL or EDC criteria (in this case, depending on the penalization therm used), a good alternative to estimate the number of components of the mixture was obtained. On the other side, these patterns do not hold when the mixture components are poorly separated, notably for the skewness parameters estimates which, in addition, showed

Table 26: Old Faithful data: Kolmogorov-Smirnov test

	2 components		3 components	
	RVM	MM	RVM	MM
K-S statistic	0.0491	0.03569	0.0481	0.0485
$p-$value	0.1060	0.1730	0.0940	0.1010

Table 27: Old Faithful data: criteria values and number of iterations

k	$\ell(\hat{\Theta})$	AIC	BIC	EDC	Number of Iterations
2	-257.57	529.13	554.37	538.22	805
3	-257.26	536.52	576.19	550.81	1641

a performance strongly dependent on large samples. Possible extensions of this work include the multivariate case and a wider family of skewed distributions, like the class of skew-normal independent distributions (Lachos *et al.*, 2010).

References

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, **19**, 716–723.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, **12**, 171–178.

Azzalini, A. (2005). The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics*, **32**, 159–188.

Bai, Z. D., Krishnaiah, P. R. & Zhao, L. C. (1989). On rates of convergence of efficient detection criteria in signal processing with white noise. *IEEE Transactions on Information Theory*, **35**, 380–388.

Basso, R. M., Lachos, V. H., Cabral, C. R. B. & Ghosh, P. (2009). Robust mixture modeling based on scale mixtures of skew-normal distributions. *Computational Statistics and Data Analysis*. doi:10.1016/j.csda.2009.09.031.

Bayes, C. L. & Branco, M. D. (2007). Bayesian inference for the skewness parameter of the scalar skew-normal distribution. *Brazilian Journal of Probability and Statistics*, **21**, 141–163.

Biernacki, C., Celeux, G. & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**, 719–725.

Biernacki, C., Celeux, G. & Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis*, **41**, 561–575.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Dias, J. G. & Wedel, M. (2004). An empirical comparison of EM, SEM and MCMC performance for problematic gaussian mixture likelihoods. *Statistics and Computing*, **14**, 323–332.

DiCiccio, T. J. & Monti, A. C. (2004). Inferential aspects of the skew exponential power distribution. *Journal of the American Statistical Association*, **99**, 439–450.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer Verlag.

Greselin, F. & Ingrassia, S. (2009). Constrained monotone EM algorithms for mixtures of multivariate t distributions. *Statistics and Computing*. doi: 10.1007/s11222-008-9112-9.

Hathaway, R. J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture models. *The Annals of Statistics*, **13**, 795–800.

Henze, N. (1986). A probabilistic representation of the skew-normal distribution. *Scandinavian Journal of Statistics*, **13**, 271–275.

Ingrassia, S. (2004). A likelihood-based constrained algorithm for multivariate normal mixture models. *Statistical Methods and Applications*, **13**, 151–166.

Ingrassia, S. & Rocci, R. (2007). Constrained monotone EM algorithms for finite mixture of multivariate gaussians. *Computational Statistics & Data Analysis*, **51**, 5339–5351.

Johnson, R. A. & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. Prentice Hall, 6th edition.

Karlis, D. & Xekalaki, E. (2003). Choosing initial values for the EM algorithm for finite mixtures. *Computational Statistics & Data Analysis*, **41**, 577–590.

Lachos, V. H., Ghosh, P. & Arellano-Valle, R. B. (2010). Likelihood based inference for skew normal independent linear mixed models. *Statistica Sinica*, **20**, 303–322.

Lin, T. & Lin, T. (2009). Supervised learning of multivariate skew normal mixture models with missing information. *Computational Statistics*. 10.1007/s00180-009-0169-5.

Lin, T. I. (2009a). Maximum likelihood estimation for multivariate skew normal mixture models. *Journal of Multivariate Analysis*, **100**, 257–265.

Lin, T. I. (2009b). Robust mixture modeling using multivariate skew t distributions. *Statistics and Computing*. doi: 10.1007/s11222-009-9128-9.

Lin, T. I., Lee, J. C. & Ni, H. F. (2004). Bayesian analysis of mixture modelling using the multivariate t distribution. *Statistics and Computing*, **14**, 119–130.

Lin, T. I., Lee, J. C. & Hsieh, W. J. (2007a). Robust mixture modelling using the skew t distribution. *Statistics and Computing*, **17**, 81–92.

Lin, T. I., Lee, J. C. & Yen, S. Y. (2007b). Finite mixture modelling using the skew normal distribution. *Statistica Sinica*, **17**, 909–927.

McLachlan, G. J. & Krishnan, T. (2008). *The EM Algorithm and Extensions*. John Wiley and Sons, second edition.

McLachlan, G. J. & Peel, G. J. (2000). *Finite Mixture Models*. John Wiley and Sons.

Meng, X. L. & Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, **80**, 267–278.

Nityasuddhi, D. & Böhning, D. (2003). Asymptotic properties of the EM algorithm estimate for normal mixture models with component specific variances. *Computational Statistics & Data Analysis*, **41**, 591–601.

Park, H. & Ozeki, T. (2009). Singularity and slow convergence of the EM algorithm for gaussian mixtures. *Neural Process Letters*, **29**, 45–59.

Peel, D. & McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing*, **10**, 339–348.

R Development Core Team (2009). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Shoham, S. (2002). Robust clustering by deterministic agglomeration EM of mixtures of multivariate t-distributions. *Pattern Recognition*, **35**, 1127–1142.

Shoham, S., Fellows, M. R. & Normann, R. A. (2003). Robust, automatic spike sorting using mixtures of multivariate t-distributions. *Journal of Neuroscience Methods*, **127**, 111–122.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Titterington, D. M., Smith, A. F. M. & Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley and Sons.

Yakowitz, S. J. & Spragins, J. D. (1968). On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, **39**, 209–214.

Yao, W. (2010). A profile likelihood method for normal mixture with unequal variance. *Journal of Statistical Planning and Inference*. doi:10.1016/j.jspi.2010.02.004.