Revista Colombiana de Estadística
Diciembre 2012, volumen 35, no. 3, pp. 457 a 478

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

Una comparación empírica de algunos métodos de inicialización EM y criterios de selección de modelos para mezclas de distribuciones normales asimetricas

José R. Pereira[a], Leyne A. Marques[b], José M. da Costa[c]

Departamento de Estatística, Instituto de Ciências Exatas, Universidade Federal do Amazonas, Manaus, Brasil

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, Mixture of distributions, Skewed distributions.

Resumen

El presente artículo muestra un estudio de simulación que evalúa el desempeño del algoritmo EM utilizado para determinar estimaciones por máxima verosimilitud de los parámetros de la mezcla finita de distribuciones normales asimétricas. Diferentes métodos de inicialización, así como el número de interacciones necesarias para establecer una regla de parada especificada y algunos criterios de selección del modelo para permitir estimar el número

[a]Associate professor. E-mail: jrpereira@ufam.edu.br
[b]Assistant professor. E-mail: leyneabuim@gmail.com
[c]Assistant professor. E-mail: zemirufam@gmail.com

apropiado de componentes de la mezcla han sido considerados. Los resultados indican que el algoritmo genera estimaciones razonables cuando los valores iniciales son obtenidos mediante el método de momentos, que junto con los criterios AIC, BIC, ICL o EDC constituyen una eficaz alternativa en la estimación del número de componentes de la mezcla. Resultados insatisfactorios se verificaron al estimar los parámetros de simetría, principalmente seleccionando un tamaño pequeño para la muestra, y en los casos conocidamente problemáticos en los cuales los componentes de la mezcla están suficientemente separados.

***Palabras clave*:** algoritmo EM, distribuciones asimétricas, mezcla de distribuciones.

1. Introduction

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, such as 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.

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, Laird & Rubin (1977), see also McLachlan & Krishnan (2008) and Ho, Pyne & Lin (2012). 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, Celeux & Govaert (2003)), asymptotic properties (Nityasuddhi & Böhning 2003) and comparisons of the standard EM with other algorithms (Dias & Wedel 2004).

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)), 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, Lee & Hsieh (2007), Lin, Lee & Yen (2007), Lin

(2009), Lin (2010) and Lin & Lin (2010). Here, we employ the algorithm presented in Basso, Lachos, Cabral & Ghosh (2010).

The goal of this work is to study the performance of the estimates produced by the EM algorithm, taking into account the method of moments and a random initialization method to obtain initial values, 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. We also investigated the density estimation issue by analyzing the estimates of the log-likelihood function at the true values of the parameters. The work is restricted to the univariate case.

The rest of the paper is organized as follows. In Sections 2 and 3, for the sake of completeness, we give a brief sketch of the skew-normal mixture model and of estimation via the EM algorithm, respectively. In Section 4, the simulation study about the initialization methods, the number of iterations and density estimation are presented. The study concerning model choice criteria is presented in Section 5. Finally, in Section 6 the conclusions of our study are draw and additional comments are given.

2. The Finite Mixture of SN Distributions Model

2.1. The Skew-Normal (SN) 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 $Y \sim \mathrm{SN}(\mu,\sigma^2,\lambda)$.

To simulate realizations of Y and to implement the EM-type algorithm a convenient stochastic representation is given by

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

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. (for proof see Henze (1986)). 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 (1) that

$$Y|T = t \sim N(\mu + \Delta t, \Gamma) \quad \text{and} \quad T \sim HN(0,1) \tag{2}$$

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

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

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

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

2.2. Finite Mixture of SN Distributions

The finite mixture of SN distributions model, hereafter 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

$$g(y_j|\Theta) = \sum_{i=1}^{k} p_i \text{SN}(y_j|\theta_i), \qquad j = 1, \ldots, n \tag{4}$$

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

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

Then it is straightforward to prove, integrating out Z_j, that y_j has density (4). If we combine this result with (2), we have the following stochastic representation for the FM-SN model

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

where

$$\Gamma_i = (1 - \delta_i^2)\sigma_i^2, \qquad \Delta_i = \sigma_i\delta_i, \qquad \delta_i = \lambda_i/\sqrt{1 + \lambda_i^2}, \qquad i = 1, \ldots, k \tag{5}$$

More details can be found in Basso et al. (2010) and references herein.

3. Estimation

3.1. An EM-type Algorithm

In this section 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. (2010) and we emphasize that, in order to do this, the representation (5) is crucial. The estimates are obtained using a faster extension of EM called the *Expectation-Conditional Maximization* (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, Lee & Yen (2007), because in the former case the updating equations for the component skewness parameter have a closed form. In what follows we consider the parametrization (5), and still use Θ to denote the vector with all parameters.

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 is to evaluate the expected value of the complete data function, known as the $Q-$*function* and 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. (2010). Here, ϕ denotes the standard normal density and we employ the following notations

$$\hat{z}_{ij} = E[Z_{ij}|y_j;\hat{\Theta}], \quad \hat{s}_{1ij} = E[Z_{ij}T_j|y_j;\hat{\Theta}] \quad \text{and} \quad \hat{s}_{2ij} = E[Z_{ij}T_j^2|y_j;\hat{\Theta}]$$

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

$$\begin{aligned}
\hat{z}_{ij}^{(m)} &= \frac{\hat{p}_i^{(m)} \mathrm{SN}(y_j|\hat{\theta}_i^{(m)})}{\sum_{i=1}^{k} \hat{p}_i^{(m)} \mathrm{SN}(y_j|\hat{\theta}_i^{(m)})} \qquad (6)\\
\hat{s}_{1ij}^{(m)} &= \hat{z}_{ij}^{(m)} \left[\hat{\mu}_{T_{ij}}^{(m)} + \frac{\phi\left(\hat{\mu}_{T_{ij}}^{(m)}/\hat{\sigma}_{T_i}^{(m)}\right)}{\Phi\left(\hat{\mu}_{T_{ij}}^{(m)}/\hat{\sigma}_{T_i}^{(m)}\right)} \hat{\sigma}_{T_i}^{(m)} \right]\\
\hat{s}_{2ij}^{(m)} &= \hat{z}_{ij}^{(m)} \left[(\hat{\mu}_{T_{ij}}^{(m)})^2 + (\hat{\sigma}_{T_i}^{(m)})^2 + \frac{\phi\left(\hat{\mu}_{T_{ij}}^{(m)}/\hat{\sigma}_{T_i}^{(m)}\right)}{\Phi\left(\hat{\mu}_{T_{ij}}^{(m)}/\hat{\sigma}_{T_i}^{(m)}\right)} \hat{\mu}_{T_{ij}}^{(m)} \hat{\sigma}_{T_i}^{(m)} \right]\\
\hat{\mu}_{T_{ij}}^{(m)} &= \frac{\hat{\Delta}_i^{(m)}}{\hat{\Gamma}_i^{(m)} + (\hat{\Delta}_i^{(m)})^2}(y_j - \hat{\mu}_i^{(m)}),\\
\hat{\sigma}_{T_i}^{(m)} &= \left(\frac{\hat{\Gamma}_i^{(m)}}{\hat{\Gamma}_i^{(m)} + (\hat{\Delta}_i^{(m)})^2} \right)^{1/2}
\end{aligned}$$

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 and several rules are proposed in the literature (see e.g., McLachlan & Krishnan (2008)). In this work our rule is to stop the process at stage m when $|\ell(\hat{\Theta}^{(m+1)})/\ell(\hat{\Theta}^{(m)}) - 1|$ is small enough.

3.2. Some Problems with Estimation in Finite Mixtures

It is well known that the likelihood of normal mixtures can be unbounded (see e.g., Frühwirth-Schnatter 2006, Chapter 6) and it is not difficult to verify

that the FM-SN models also have this feature. 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 e.g., Hathaway 1985, Ingrassia 2004, Ingrassia & Rocci 2007, Greselin & Ingrassia 2010). Thus, following Nityasuddhi & Böhning (2003), we investigate only the performance of the EM algorithm when considered component specific parameters (that is, unrestricted) of the mixture and we mention the estimates produced by the algorithm of section 3.1 as "EM estimates", that is, some sort of solution of the score equation, instead of "maximum likelihood estimates".

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 (see, Titterington, Smith & Makov 1985, Chapter 3 for details). 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 e.g., Peel & McLachlan 2000, Shoham 2002, Shoham, Fellows & Normann 2003, Lin, Lee & Ni 2004) do not present a formal proof of its identifiability. It is important to mention that 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.

4. A Simulation Study of Initial Values

4.1. Description of the Experiment

It is well known that the performance of the EM algorithm 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, we adopt a fixed rule to stop the process at stage m when

$$\left| \frac{\ell(\hat{\Theta}^{(m+1)})}{\ell(\hat{\Theta}^{(m)})} - 1 \right| < 10^{-6}$$

because we believe that this tolerance for the change in $\ell(\hat{\Theta})$ is quite reasonable in the applications where the primary interest is on the sequence of the log-likelihood values rather than the sequence of parameter estimates (McLachlan & Peel 2000, Section 2.11).

In the mixture context, the choice of starting values for the EM algorithm is crucial because, as noted by Dias & Wedel (2004), there are various saddle regions surrounding the possible local maximum 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 3.1, where an inter-

esting question is to investigate the performance of the EM algorithm with respect to the skewness parameter estimation for each component density in the FM-SN model. We consider the following methods to obtain initial values:

The Random Values Method (RVM): we first divide the generated random sample into k sub-samples employing the *k-means* method. The initialization of *k-means* algorithm is random, being recommended to adopt many different choices and we employ five random initializations (see Hastie, Tibshirani & Friedman 2009, Section 14.3). Let φ_i be the sub-sample i. Consider the following points artificially generated from uniform distributions over the specified intervals

$$\begin{aligned} \hat{\xi}_i^{(0)} &\sim U(\min\{\varphi_i\}, \max\{\varphi_i\}) \\ \hat{\omega}_i^{(0)} &\sim U(0, var\{\varphi_i\}) \,, \\ \hat{\gamma}_i^{(0)} &\sim sgn(sc\{\varphi_i\}) \times |U(-0.9953, 0.9953)| \end{aligned} \tag{7}$$

where $\min\{\varphi_i\}$, $\max\{\varphi_i\}$, $var\{\varphi_i\}$ and $sc\{\varphi_i\}$ denote, respectively, the minimum, the maximum, the sample variance and the sample skewness coefficient of φ_i, $i = 1, .., k$, also $|\cdot|$ denotes absolute value. These quantities are taken as rough estimates for the mean, variance and skewness coefficient associated to subpopulation i, respectively. The suggested form for $\hat{\gamma}_i^{(0)}$ is due to the fact that the range for the skewness coefficient in SN models is $(-0.9953, 0.9953)$ and to maintain the sign of the sample skewness coefficient.

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

$$\begin{aligned} \hat{\mu}_i^{(0)} &= \hat{\xi}_i^{(0)} - \sqrt{2/\pi}\delta_{(\hat{\lambda}_i^{(0)})}\hat{\sigma}_i^{(0)} \\ \hat{\sigma}_i^{(0)} &= \sqrt{\frac{\hat{\omega}_i^{(0)}}{1 - \frac{2}{\pi}\delta^2_{(\hat{\lambda}_i^{(0)})}}} \\ \hat{\lambda}_i^{(0)} &= \pm\sqrt{\frac{\pi(\hat{\gamma}_i^{(0)})^{2/3}}{2^{1/3}(4-\pi)^{2/3} - (\pi-2)(\hat{\gamma}_i^{(0)})^{2/3}}} \end{aligned} \tag{8}$$

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 (3) with their respective estimators in (7) 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, namely, a uniform distribution over the unit simplex $\left\{(p_1, \ldots, p_k);\ p_i \geq 0,\ \sum_{i=1}^k p_i = 1\right\}$.

Method of Moments (MM): the initial values are obtained using equations (8), 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, $i = 1, ..k$, with the k sub-samples obtained by the *k-means* method with five random initializations . 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

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

We generated samples from the FM-SN model with $k = 2$ and $k = 3$ components, with sizes fixed as $n = 500$; 1000; 5000 and 1,0000. In addition, we consider different degree of heterogeneity of the components, for $k = 2$ the "moderately separated" ($2MS$), "well separated" ($2WS$) and "poorly separated" ($2PS$) cases and for $k = 3$ 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 and the reason to consider them as an 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). In Table 1 the parameters values used in the study are presented and the figures 1 and 2 show some histograms exemplifying these degrees of heterogeneity.

TABLE 1: Parameters values for FM-SN models

Case	p_1	μ_1	σ_1^2	λ_1	p_2	μ_2	σ_2^2	λ_2	p_3	μ_3	σ_3^2	λ_3
$2MS$	0.6	5	9	6	0.4	20	16	-4				
$2WS$	0.6	5	9	6	0.4	40	16	-4				
$2PS$	0.6	5	9	6	0.4	15	16	-4				
$3PWS$	0.4	5	9	6	0.3	20	16	-4	0.3	28	16	4
$3WWS$	0.4	5	9	6	0.3	30	16	-4	0.3	38	16	4

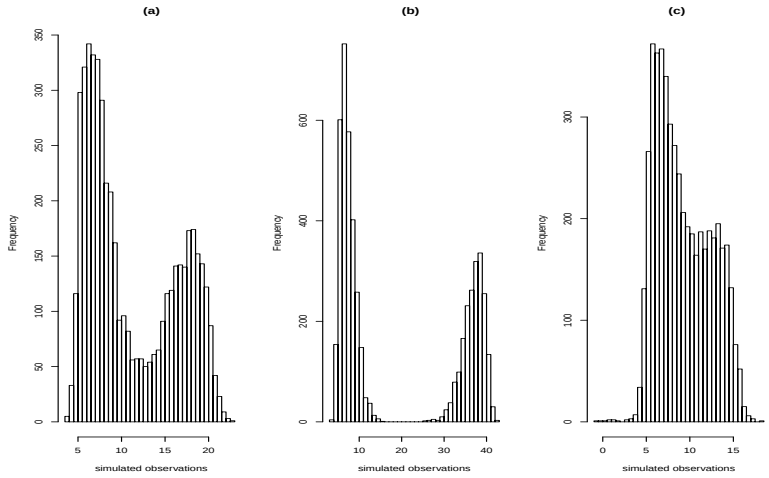

FIGURE 1: Histograms of FM-SN data: (a) $2MS$, (b) $2WS$ and (c) $2PS$.

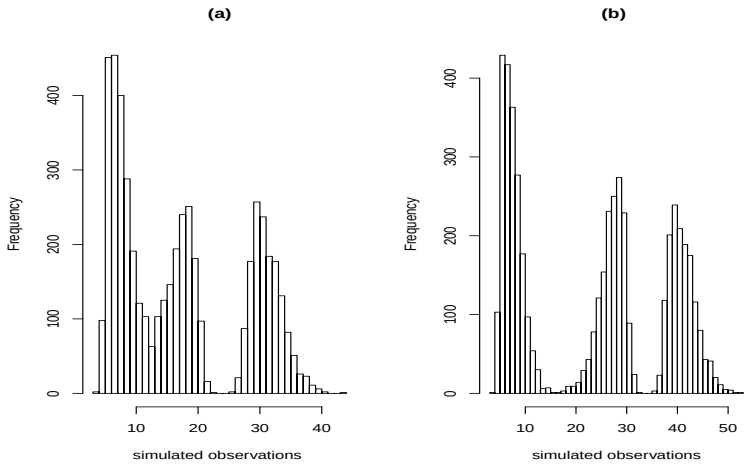

FIGURE 2: Histograms of FM-SN data: (a) $3PWS$ and (b) $3WWS$.

For each combination of parameters and sample size, samples from the FM-SN model were artificially generated and we obtained estimates of the parameters using the algorithm presented in section 3.1 initialized by each method proposed. This procedure was repeated 5,000 times and we computed the bias and mean squared error (MSE) over all samples, which for μ_i are defined as

$$\text{bias} = \frac{1}{5,000}\sum_{j=1}^{5,000}\hat{\mu}_i^{(j)} - \mu_i \quad \text{and} \quad \text{MSE} = \frac{1}{5,000}\sum_{j=1}^{5,000}(\hat{\mu}_i^{(j)} - \mu_i)^2,$$

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. All the computations were made using the R system (R Development Core Team 2009) and the implementation of the EM algorithm was computed by employing the R package `mixsmsn` (Cabral, Lachos & Prates 2012), available on CRAN.

As a note about implementation, an expected consequence of the non-identifiability cited in section 3.2 is the permutation of the component labels when using the *k-means* method to perform an initial clustering of the data. This label-witching problem seriously affects the determination of the MSE and consequently the evaluation of the consistency of the estimates (on this issue see e.g Stephens 2000). To overcome this problem we adopted an order restriction on the initial values of the location parameters and estimates for all parameters were sorted according to their true values before computing the bias and MSE. We emphasize that we employ this order restriction order to ensure the determination of the MSE, impartially, to compare the initialization methods.

4.2. Bias and Mean Squared Error (MSE)

Tables 2 and 3 present, respectively, bias and MSE of the estimates in the $2MS$ case. From these tables, we can see that, with both 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 the MM showed better performance than the RVM, for all sample sizes and parameters. When using the RVM, in general, the absolute value of the bias and MSE of the estimates of σ_i^2 and λ_i are very large compared with that obtained using the MM. In general, according to our criteria, we can conclude that the MM method presented a satisfactory performance in all situations.

Table 2: Bias of the estimates - two 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	RVM	0.14309	−0.39956	−0.45779	1.30675	1.32753	−1.10056	0.01927	−0.01927
	MM	−0.01842	−0.02842	0.39275	−0.12159	1.05604	−0.37835	0.00159	0.00159
1000	RVM	0.10815	−0.33814	−0.26339	1.09304	0.61695	−0.60402	0.01599	−0.01599
	MM	−0.02194	−0.01214	0.39285	−0.25737	0.58058	−0.10558	0.00212	−0.00212
5000	RVM	0.10776	−0.26897	−0.21237	0.68574	0.27081	−0.10679	0.01412	−0.01412
	MM	−0.02641	−0.01109	0.35592	−0.45453	0.44153	0.04753	0.00283	0.00283
10000	RVM	0.09904	−0.28784	−0.19722	0.67306	0.29762	0.04354	0.01451	−0.01451
	MM	−0.02783	−0.01026	0.35406	−0.44957	0.42318	0.05692	0.00275	0.00275

Table 3: MSE of the estimates - two 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	RVM	0.69489	2.62335	5.70944	74.43050	55.24861	108.27990	0.00797	0.00797
	MM	0.01509	0.09895	1.81950	13.41757	8.11938	5.48844	0.00071	0.00071
1000	RVM	0.49336	2.34634	4.40063	64.12685	13.10896	44.12675	0.00651	0.00651
	MM	0.00732	0.03158	0.99780	6.14854	1.94043	0.72317	0.00035	0.00035
5000	RVM	0.51481	1.91370	2.94570	44.01847	12.20841	7.70399	0.00544	0.00544
	MM	0.00203	0.00611	0.30162	1.24062	0.45855	0.11495	7.41e-05	7.41e-05
10000	RVM	0.48098	2.14718	3.13302	38.22046	10.01400	4.01044	0.00564	0.00564
	MM	0.00141	0.00598	0.22876	0.71719	0.30481	0.06338	3.97e-05	3.97e-05

The bias and MSE of the estimates for the $2WS$ case are presented in tables 4 and 5, respectively. As in the $2MS$ case, their values decrease when the sample size increases. Comparing the initialization methods, we can see again the poor performance of RVM, notably when estimating σ_i^2 and λ_i. The performance of

Table 4: Bias of the estimates - two 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	RVM	0.12911	−0.43583	−0.29445	7.32555	1.49636	−0.96377	0.01053	−0.01053
	MM	−0.01943	−0.00087	0.12457	0.09558	0.78816	−0.51038	−0.00013	0.00013
1000	RVM	0.11189	−0.39592	−0.24494	6.83576	1.06715	−0.42096	0.00938	−0.00938
	MM	−0.01896	0.00764	0.11564	0.09444	0.50192	−0.2533	0.00031	0.00031
5000	RVM	0.20668	−0.61846	−0.33728	8.68389	0.57418	−0.26060	0.01440	−0.01440
	MM	−0.01863	0.00770	0.10208	0.09145	0.29384	−0.09125	3.87e-05	−3.87e-05
10000	RVM	0.24109	−0.64407	−0.33492	8.68057	0.37535	−0.15859	0.01457	−0.01457
	MM	−0.01791	0.00615	0.10220	0.08006	−0.27967	−0.07295	−8.37e-05	8.37e-05

Table 5: MSE of the estimates - two 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	RVM	0.95939	6.72054	2.57082	2314.91800	204.77300	121.73630	0.00466	0.00466
	MM	0.01411	0.08918	0.86552	5.30267	4.67931	6.21511	0.00047	0.00047
1000	RVM	0.86198	6.26880	2.01054	2385.87300	157.64460	38.02320	0.00411	0.00411
	MM	0.00705	0.05053	0.45151	2.58499	1.64592	0.88093	0.00024	0.00024
5000	RVM	1.62846	10.58635	2.33173	2366.03700	80.19515	32.81540	0.00577	0.00577
	MM	0.00164	0.03406	0.62874	0.11866	0.30607	0.16967	4.79e-05	4.79e-05
10000	RVM	2.18692	10.97898	2.35874	2324.93000	31.51082	26.92428	0.00587	0.00587
	MM	0.00099	0.03115	0.07153	0.37187	0.18739	0.10765	2.36e-05	2.36e-05

Table 6: Bias of the estimates - two 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	RVM	0.54135	−3.05044	−6.45635	−4.92197	0.54115	3.93878	0.10598	−0.10598
	MM	0.67462	−2.97859	−6.68560	−5.96459	3.41267	4.80134	0.09124	−0.09124
1000	RVM	0.47419	−3.24148	−6.49341	−5.12707	−0.84129	3.81489	0.09821	−0.09821
	MM	0.67263	−2.76804	−6.44084	−5.94106	−1.59226	4.52825	0.09288	−0.09288
5000	RVM	0.11827	−3.38188	−6.31995	−4.61048	−0.32021	4.27876	0.10485	−0.10485
	MM	0.43959	−2.63605	−6.54154	−5.28364	−1.61009	4.58531	0.10837	−0.10837
10000	RVM	−0.01664	−3.32212	−6.26154	−4.56424	0.04022	4.33918	0.10793	−0.10793
	MM	0.34272	−2.58084	−6.36467	−5.15910	−0.35239	4.26839	0.11364	−0.11364

Table 7: MSE of the estimates - two 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	RVM	1.53553	12.77114	44.60716	61.74994	196.12690	48.65583	0.02622	0.02622
	MM	1.56413	10.28718	46.61793	49.10880	343.77230	30.24965	0.01746	0.01746
1000	RVM	1.67311	14.41824	44.47786	63.53965	56.47113	51.72255	0.02561	0.02561
	MM	1.39797	8.39757	42.72479	49.20213	25.63282	30.17503	0.01755	0.01755
5000	RVM	1.91642	16.52284	42.21555	60.68243	84.34230	37.77900	0.02987	0.02987
	MM	0.72188	7.45145	43.16745	35.72524	17.52604	25.12349	0.01734	0.01734
10000	RVM	2.18807	16.37261	41.17759	57.61406	117.18120	37.92872	0.02981	0.02981
	MM	0.48657	7.39552	41.31111	32.51687	19.04316	19.12946	0.01783	0.01783

TABLE 8: Bias of the estimates - two poorly separated and one well separated ($3PWS$).

n	Method	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\sigma}_3^2$
500	RVM	3.86197	−2.22030	0.34248	0.02341	0.01208	−0.03549
	MM	−0.02634	−0.07339	0.00338	0.49861	−1.95137	1.98042
1000	RVM	0.13962	−0.32733	0.06264	−0.43484	−1.79386	4.25781
	MM	−0.02324	−0.04743	−0.01057	0.35349	−1.50284	1.19124
5000	RVM	0.09945	−0.27154	0.06481	−0.25861	−1.36352	3.16178
	MM	−0.02336	−0.04234	−0.01724	0.43048	−0.91359	0.40114
10000	RVM	0.08625	−0.25897	0.04110	−0.23192	−1.23694	3.24305
	MM	−0.02290	−0.04478	−0.01054	0.42927	−0.74921	0.32357

TABLE 9: Bias of the estimates - two poorly separated and one well separated ($3PWS$).

n	Method	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$
500	RVM	3.86197	−2.22029	0.34248	0.02341	−0.03549	0.01208
	MM	2.78183	−1.00171	0.37213	0.00378	−0.01641	0.01263
1000	RVM	1.46755	−0.59057	0.17228	0.01745	−0.02687	0.00941
	MM	0.92051	−0.11912	0.22981	0.00218	−0.01139	0.00917
5000	RVM	0.66370	−0.05447	0.02601	0.01554	−0.01996	0.00441
	MM	0.57731	0.10412	0.13295	0.00285	−0.00614	0.00328
10000	RVM	0.53239	0.12565	0.01455	0.01376	−0.01651	0.00274
	MM	0.53269	0.10807	0.11858	0.00261	−0.00474	0.00212

TABLE 10: MSE of the estimates - two poorly separated and one well separated ($3PWS$).

n	Method	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\sigma}_3^2$
500	RVM	0.73592	4.19665	1.33364	7.66927	13.88521	443.83960
	MM	0.02326	0.11884	0.10291	2.91056	9.62248	16.26670
1000	RVM	0.63416	3.13945	1.31999	5.46298	9.99991	335.50440
	MM	0.01062	0.12819	0.04066	1.56864	5.58594	6.68605
5000	RVM	0.46937	2.92722	1.35910	4.07581	5.74955	360.56410
	MM	0.00305	0.08728	0.00844	0.68559	1.76444	5.77978
10000	RVM	0.38973	2.83554	1.24776	3.41887	5.16622	349.93790
	MM	0.00204	0.07621	0.00434	0.54306	1.11403	3.81096

TABLE 11: MSE of the estimates - two poorly separated and one well separated ($3PWS$).

n	Method	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$
500	RVM	401.77671	659.16871	15.66503	0.00691	0.00077	0.00633
	MM	89.47438	57.35417	2.67579	0.00068	0.00043	0.00056
1000	RVM	74.62215	95.01346	8.04131	0.00553	0.00053	0.00506
	MM	3.92305	1.23581	0.88842	0.00033	0.00022	0.00027
5000	RVM	19.78691	120.88661	1.57637	0.00433	0.00021	0.00427
	MM	0.73751	0.30914	0.18520	7.20e-05	3.75e-05	6.93e-05
10000	RVM	18.46118	7.03221	1.60081	0.00417	0.00031	0.00363
	MM	0.50922	0.24319	0.09469	4.23e-05	1.82e-05	4.10e-05

Table 12: Bias of the estimates - three well separated ($3WWS$).

n	Method	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\sigma}_3^2$
500	RVM	0.24221	−0.48999	0.07992	−0.70713	−1.98125	11.24986
	MM	−0.02717	−0.02233	0.00361	0.19686	−1.50136	1.60453
1000	RVM	0.24327	−0.49722	0.04203	−0.63272	−1.46754	9.55012
	MM	−0.02742	−0.01908	−0.01276	0.10826	−0.97860	1.18908
5000	RVM	0.33293	−0.67092	0.01975	−0.65223	−1.02971	10.20861
	MM	−0.02265	−0.01872	−0.01604	0.11568	−0.39018	0.58417
10000	RVM	0.33457	−0.80115	−0.07909	−0.63707	−0.90179	8.87343
	MM	−0.02328	−0.01509	−0.01704	0.10738	−0.28876	0.43182

Table 13: Bias of the estimates - three well separated ($3WWS$).

n	Method	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$
500	RVM	3.63818	−0.59785	0.16417	0.02216	0.01369	−0.03586
	MM	1.98187	−0.83366	0.31660	−0.00014	0.01385	−0.01371
1000	RVM	1.68173	−0.44696	0.07427	0.02026	0.00987	−0.03014
	MM	0.83325	−0.29503	0.25427	9.95e-05	0.00977	−0.00987
5000	RVM	0.85209	−0.16743	0.11006	0.02449	0.00285	−0.02735
	MM	0.38848	−0.08961	0.12745	−1.00e-05	0.00435	−0.00434
10000	RVM	0.45846	0.06753	0.13875	0.02299	0.00213	−0.02512
	MM	0.34917	−0.04635	0.10236	−1.45-e05	0.00308	−0.00306

Table 14: MSE of the estimates - three well separated ($3WWS$).

n	Method	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\sigma}_3^2$
500	RVM	1.92196	9.35109	2.52362	6.19796	10.97185	1848.86500
	MM	0.02247	0.17468	0.09472	1.43362	6.87787	8.26021
1000	RVM	2.02266	10.57483	2.68414	5.17628	9.12883	1517.61500
	MM	0.01091	0.14971	0.04303	0.79533	3.51761	3.92717
5000	RVM	2.93899	12.82373	3.23907	5.17035	8.56857	1510.38300
	MM	0.00249	0.14126	0.00836	0.21648	0.89455	0.81992
10000	RVM	3.08463	14.30240	2.85946	4.53532	9.51754	1383.25000
	MM	0.00156	0.13665	0.00412	0.17027	0.79811	0.42657

Table 15: MSE of the estimates - three well separated ($3WWS$).

n	Method	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$
500	RVM	361.45080	31.09128	9.18895	0.00727	0.00102	0.00712
	MM	34.37280	18.59549	2.08116	0.00047	0.00042	0.00041
1000	RVM	132.29800	28.30730	3.20316	0.00656	0.00089	0.00632
	MM	3.81642	0.92243	1.51732	0.00024	0.00022	0.00021
5000	RVM	41.27475	23.97153	1.03027	0.00845	0.00111	0.00684
	MM	0.50781	0.30959	0.18493	4.86e05	4.15e-05	4.15e-05
10000	RVM	22.70630	10.52715	1.66676	0.00754	0.00091	0.00647
	MM	0.30177	0.29030	0.08916	2.41e-05	2.11e-05	2.07e-05

MM is satisfactory and 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 6 and 7. Bias and MSE are larger than in the $2MS$ and $2WS$ cases (for all sample sizes) with both methods. 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 abovementioned work of Park & Ozeki (2009).

For the $3PWS$ case the results for the bias are shown in tables 8 and 9 and for the MSE in tables 10 and 11. It seems that consistency is achieved for $\hat{p}_i$, $\hat{\mu}_i$ and $\hat{\sigma}_i^2$, using MM. However, this is not the behavior for $\hat{\lambda}_i$. 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. These results are similar to that obtained for the FM-SN model with two components

Finally, for the $3PWW$ case, the bias of the estimates is presented in tables 12 and 13 and the EQM are shown in tables 14 and 15. Concerning the estimates $\hat{p}_i$ and $\hat{\mu}_i$, very satisfactory results are obtained, with small values of bias and MSE when using the 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, the 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 (it may be slower for the scale parameter!), except for the skewness parameter, indicating that a sample size larger than 5,000 is necessary to achieve consistency in the case of this parameter. The study also suggests that the degree of heterogeneity of the population has a remarkable influence on the quality of the estimates.

TABLE 16: Means and standard deviations ($\times 10^{-4}$) of d_r.

Method	n	Cases		
		$2MS$	2WS	2PS
RVM	500	1.43 (2.55)	2.51 (7.31)	2.23 (1.31)
	1000	1.05 (2.21)	2.24 (6.48)	0.84 (0.69)
	5000	0.75 (2.24)	2.01 (8.67)	1.02 (0.75)
	10000	0.67 (2.35)	2.14 (8.86)	0.67 (0.69)
MM	500	0.90 (0.66)	1.32 (0.95)	2.20 (1.21)
	1000	0.62 (0.48)	1.23 (0.80)	0.76 (0.57)
	5000	0.33 (0.24)	0.34 (0.26)	0.79 (0.46)
	10000	0.22 (0.16)	0.42 (0.26)	0.44 (0.33)

4.3. Density Estimation Analysis

In this section we consider 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 =500; 1,000; 5,000 and 10,000. For each combination of parameters and sample size, 5000 samples were generated and the following measure was considered to compare the 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 $d_r(M) < d_r(M')$. Table 16 presents the means and standard deviations of d_r.

For $2MS$ case, we can see that these values decrease when the sample size increases with both methods and that the MM presented the lowest mean value and standard deviation for all sample sizes. For the $2WS$ case, we do not observe a monotone behavior for d_r, the mean values and the standard errors are larger than that presented in the $2MS$ case, with poor performance of RVM. In this $2PS$ case, although we also do not observe a monotone behavior for d_r, we can see that the MM presented a better performance than the TVM.

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.

4.4. Number of Iterations

It is well known that one of the major drawbacks of the EM algorithm is the slow convergence. The problem becomes 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 for the convergence of the algorithm. As in subsection 4.3, here we consider only the $2MS$, $2WS$ and $2PS$ cases, with the same sample sizes and number of replications. For each generated sample, we observed the number of iterations and the means and standard deviations of this quantity were computed. The simulations results are reported in Table 17.

Results suggest that in the three cases, using 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. For the $2PS$ case, as expected, we have a poor behavior possibly due to the population homogeneity, as we commented before. An interesting fact is that, in the $2PS$ case, the RVM has a smaller mean number of iterations.

TABLE 17: Means and standard deviations of number of iterations.

Method	n	Cases		
		$2MS$	2WS	2PS
RVM	500	306.14 (387.70)	129.81 (117.01)	337.82 (289.01)
	1,000	283.57 (289.32)	126.87 (130.61)	319.70 (242.08)
	5,000	280.28 (260.36)	128.29 (213.99)	336.85 (206.38)
	10,000	286.31 (271.83)	131.33 (190.77)	353.61 (211.99)
MM	500	147.53 (72.79)	126.62 (26.80)	457.97 (167.28)
	1,000	129.37 (33.34)	119.70 (15.39)	429.42 (119.77)
	5,000	116.35 (11.57)	113.91 (5.90)	372.14 (71.89)
	10,000	115.10 (8.37)	113.29 (4.23)	352.33 (97.54)

5. A Simulation Study of Model Choice

There is a key issue with the use of finite mixtures to estimate the number of components in order to obtain a suitable fit. One possible approach is to use some criteria function and compute

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

where $C(\hat{\Theta}_{(k)})$ is the criterion 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 (for other approaches see McLachlan & Peel 2000).

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, Krishnaiah & Zhao 1989) and the Integrated Completed Likelihood Criterion (ICL) (Biernacki, Celeux & Govaert 2000). The AIC, BIC and EDC criteria 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

$$\lim(c_n/n) = 0 \quad \text{and} \quad \lim_{n\to\infty}(c_n/(\log\log n)) = \infty$$

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 (6)). 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 set up.

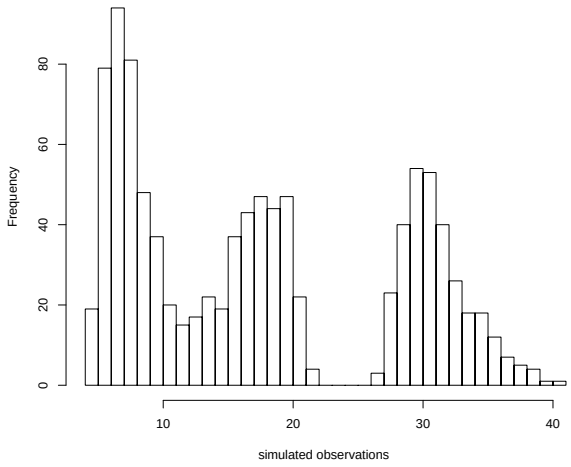

FIGURE 3: Histogram of a FM-SN sample with $k = 3$ and $n = 1,000$.

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 18

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

TABLE 18: Percentage of times that the criteria chosen the correct model.

n	AIC	BIC	ICL	EDC/ c_n			
				$0.2\log(n)$	$0.2\sqrt{n}$	$0.2n/\log(n)$	$0.5\sqrt{n}$
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

to choose the correct number of components and that, depending on the choice of c_n, ICL can not be as good as AIC or better than ICL and BIC.

6. Final Remarks

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 (AIC, BIC, ICL and EDC), the results suggest that the EDC, with the penalization term appropriate, provides a good alternative to estimate the number of components of the mixture. 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 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 (see Cabral et al. (2012)).

[Recibido: agosto de 2011 — Aceptado: octubre de 2012]

References

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE Transactions on Automatic Control* **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. (2010), 'Robust mixture modeling based on scale mixtures of skew-normal distributions', *Computational Statistics and Data Analysis* **54**, 2926–2941.

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 and Data Analysis* **41**, 561–575.

Cabral, C. R. B., Lachos, V. H. & Prates, M. O. (2012), 'Multivariate mixture modeling using skew-normal independent distributions', *Computational Statistics and Data Analysis* **56**, 126–142.

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. (2010), 'Constrained monotone EM algorithms for mixtures of multivariate t distributions', *Statistics and Computing* **20**(1), 9–22.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Second Edition*, Springer, USA.

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.

Ho, H. J., Pyne, S. & Lin, T. I. (2012), 'Maximum likelihood inference for mixtures of skew Student-t-normal distributions through practical EM-type algorithms', *Statistics and Computing* **22**(1), 287–299.

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 and Data Analysis* **51**, 5339–5351.

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

Lin, T. I. (2009), 'Maximum likelihood estimation for multivariate skew normal mixture models', *Journal of Multivariate Analysis* **100**, 257–265.

Lin, T. I. (2010), 'Robust mixture modeling using multivariate skew t distributions', *Statistics and Computing* **20**(3), 343–356.

Lin, T. I., Lee, J. C. & Hsieh, W. J. (2007), 'Robust mixture modelling using the skew t distribution', *Statistics and Computing* **17**, 81–92.

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. & Yen, S. Y. (2007), 'Finite mixture modelling using the skew normal distribution', *Statistica Sinica* **17**, 909–927.

Lin, T. & Lin, T. (2010), 'Supervised learning of multivariate skew normal mixture models with missing information', *Computational Statistics* **25**(2), 183–201.

McLachlan, G. J. & Krishnan, T. (2008), *The EM Algorithm and Extensions*, 2 edn, John Wiley and Sons.

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 and 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.

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.

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

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* **140**, 2089–2098.