JIRSS (2017)
Vol. 16, No. 01, pp 33-51
DOI: 10.18869/acadpub.jirss.16.1.1003

On the Finite Mixture Modeling via Normal Mean-Variance Birnbaum-Saunders Distribution

Mehrdad Naderi [1,2], Alireza Arabpour [1], Ahad Jamalizadeh[1]

[1]Department of Statistics, Faculty of Mathematics and Computer, Shahid Bahonar University of Kerman, Kerman, Iran.
[2]Young Researchers Society, Shahid Bahonar University of Kerman, Kerman, Iran.

Received: 10/07/2016, Revision received: 23/11/2016, Published online: 22/03/2017

Abstract. This paper presents a new finite mixture model using the normal mean-variance mixture of Birnbaum-Saunders distribution. The proposed model is multi-modal with wider ranges of skewness and kurtosis. Moreover, it is useful for modeling highly asymmetric data in various theoretical and applied statistical problems. The maximum likelihood estimates of the parameters of the model are computed iteratively by feasible EM algorithm. To illustrate the finite sample properties and performance of the estimators, we conduct a simulation study and illustrate the usefulness of the new model by analyzing a real dataset.

Keywords. Birnbaum-Saunders distribution, ECM-algorithm, Finite mixture model, Mean-variance mixture distribution.

MSC: 62E10, 62E15, 62E07.

Corresponding Author: Alireza Arabpour (arabpour@uk.ac.ir)
Mehrdad Naderi (Mehrdad.Naderi@ymail.com)
Ahad Jamalizadeh (a.jamalizadeh@uk.ac.ir)

1 Introduction

Finite mixture (FM) models are one of the most applicable and flexible models aiming to provide a statistical tool for studying the heterogeneity in clustering and classification analysis and pattern recognition problems. This model is a weighted sum of g distribution functions which are known as mixture components and their number is held fixed. Due to the usefulness of the FM models, many kinds of different FM models were introduced in the past decade via considering various mixture components. For example, see the article by Lin *et al.* (2007) and Lin (2010) and monographs by McLachlan and Basford (1988) , Frühwirth-Schnatter (2006).

Although the FM of normal distributions (FM-N) has several properties and is useful in applied statistics, the model is not flexible enough for fat tails or discrepant data. To avoid this deficiency, the FM of multivariate t distributions (FM-T) was proposed by Peel and McLachlan (2000). Even though the new model has a heavy tail the FM-T model cannot be ideal in the presence of highly asymmetric observations. To produce a flexible model with wider ranges of skewness and kurtosis, Basso *et al.* (2010) studied a class of mixture models where the component densities are scale mixtures of the (univariate) skew-normal distributions, which include skew-normal, skew-t, skew-contaminated normal and skew-slash distributions as special cases. Following their work, Cabral *et al.* (2012) introduced the finite mixture of multivariate scale mixtures of the skew-normal distributions and studied its properties.

Another class of skewed distributions which have heavier tails than normal is the class of normal mean-variance mixture (NMV) distributions. Recently, various members of this class, which is also known as a location-scale mixture distributions, have been considered and studied in the literature (see *e.g.* Arslan (2010, 2015), Nematollahi *et al.* (2016)). As a special case of NMV distributions, Barndorff-Nielsen (1977) introduced the family of Generalized Hyperbolic (GH) distributions for modeling dune movements. The GH family includes symmetric distributions such as normal, Student t, Laplace, elliptically symmetric distributions and asymmetrical distributions such as skewed t, skewed Laplace and hyperbolic distributions as special cases. Pourmousa *et al.* (2015) proposed a normal mean-variance mixture distribution based on Birnbaum-Saunders distribution (NMVBS). They showed that all maximum likelihood (ML) estimators of the parameters in this family have an explicit form and provide a better fit than skew-t to the data. They also studied some properties of the family and introduced the NMVBS-ARCH model.

In this paper, we first show that the NMVBS distribution has wider ranges of skewness and kurtosis and then we construct its finite mixture model (FM-NMVBS).

Some properties of the model are studied and the ML estimators of the parameters are computed via the expectation-conditional maximization (ECM) algorithm (Meng and Rubin , 1993). To show the asymptotic property of the ML estimators, a simulation study is undertaken. Finally, we fit the FM-NMVBS model to a real dataset and compare this model with the finite mixture of scale mixtures of the skew-normal distributions.

The rest of the paper is organized as follows. In Section 2, we briefly review the GH and univariate NMVBS distributions. Section 3 describes the univariate finite mixture of NMVBS distributions, some of its properties and ECM procedure for computing the ML estimators of the parameters. A simulation study and a real data analysis are reported in Section 4. With some concluding remarks in Section 5, the paper will be closed.

2 Preliminaries

2.1 Generalized Hyperbolic distribution

A random variable X is said to have a GH distribution if its probability density function (pdf) is

$$f_{GH}(x;\mu,\lambda,\sigma^2,\kappa,\chi,\psi) = C\frac{K_{\kappa-0.5}\left(\sqrt{(\psi+\lambda^2/\sigma^2)(\chi+(x-\mu)^2/\sigma^2)}\right)}{\left(\sqrt{(\psi+\lambda^2/\sigma^2)(\chi+(x-\mu)^2/\sigma^2)}\right)^{0.5-\kappa}}\exp\{(x-\mu)\lambda/\sigma^2\}, \tag{2.1}$$

where $x\in\mathbb{R}$,

$$C=\frac{(\sqrt{\psi/\chi})^{\kappa}(\psi+\lambda^2/\sigma^2)^{0.5-\kappa}}{\sqrt{2\pi\sigma^2}K_\kappa(\sqrt{\psi\chi})},$$

$K_\kappa(.)$ denotes the modified Bessel function of the third kind, $\mu,\lambda,\kappa\in\mathbb{R}$ and the parameters χ and ψ are defined by

$$\begin{cases}\chi\geq 0,\psi>0 & \text{if } \kappa>0\\ \chi>0,\psi>0 & \text{if } \kappa=0\\ \psi\geq 0,\chi>0 & \text{if } \kappa<0.\end{cases}$$

It can easily be shown (McNeil *et al.* , 2005) that, if X follows the pdf (2.1), then it can be obtained from an NMV model, meaning that the random variable X has the representation of the form as

$$X\stackrel{d}{=}\mu+W\lambda+W^{1/2}Z, \tag{2.2}$$

where $\stackrel{d}{=}$ denotes equality in distribution, Z is a normal random variable with mean zero and variance σ^2, denoted by $Z \sim N(0, \sigma^2)$, and W is an independent random variable with generalized inverse Gaussian (GIG) distribution, introduced by Good (1953), which has the pdf as

$$f_{GIG}(w; \kappa, \chi, \psi) = (\frac{\psi}{\chi})^{\kappa/2} \frac{w^{\kappa-1}}{2K_\kappa(\sqrt{\psi\chi})} \exp\left\{\frac{-1}{2}\left(w^{-1}\chi + w\psi\right)\right\}, \quad w > 0.$$

The GIG distribution is widely used for modeling and analysing lifetime data. By representation (2.2), we can readily obtain the expectation and variance of X as

$$\begin{aligned} E[X] =& \mu + \lambda\Big(\frac{\chi}{\psi}\Big)^{\frac{1}{2}} R_{(\kappa,1)}(\sqrt{\chi\psi}), \\ Var(X) =& \Big(\frac{\chi}{\psi}\Big)^{\frac{1}{2}} R_{(\kappa,1)}(\sqrt{\chi\psi})\sigma^2 + \lambda^2\Big[\Big(\frac{\chi}{\psi}\Big) R_{(\kappa,2)}(\sqrt{\chi\psi}) - \Big(\Big(\frac{\chi}{\psi}\Big)^{\frac{1}{2}} R_{(\kappa,1)}(\sqrt{\chi\psi})\Big)\Big], \end{aligned} \tag{2.3}$$

where $R_{(\kappa,a)}(c) = K_{\kappa+a}(c) \big/ K_\kappa(c)$.

2.2 The NMVBS distribution

Let W be a random variable taking positive real values. It follows the Birnbaum-Saunders (BS) distribution, $W \sim BS(\alpha, \beta)$, if its cumulative distribution function (cdf) is

$$F(w; \alpha, \beta) = \Phi\left[\frac{1}{\alpha}\left(\sqrt{\frac{w}{\beta}} - \sqrt{\frac{\beta}{w}}\right)\right], \quad w > 0,\ \alpha > 0,\ \beta > 0, \tag{2.4}$$

where $\Phi(.)$ is the cdf of the standard normal distribution and α and β are the shape and scale parameters, respectively. The BS distribution was introduced by Birnbaum and Saunders (1969a,b) as a failure time distribution and has received considerable attention in the last three decades. This distribution has a positive skewness and is related to the GIG distribution. Desmond (1986) established that the BS distribution can be written as an equally weighted mixture of a GIG distribution and its complementary reciprocal. It means that we can obtain the cdf (2.4) in an alternative form as

$$F(w; \alpha, \beta) = \frac{1}{2} F_{GIG}\left(w; \frac{1}{2}, \frac{\beta}{\alpha^2}, \frac{1}{\beta\alpha^2}\right) + \frac{1}{2} F_{GIG}\left(w; \frac{-1}{2}, \frac{\beta}{\alpha^2}, \frac{1}{\beta\alpha^2}\right),$$

where $F_{GIG}(.)$ represents the cdf of the GIG distribution. Considering the random variable W in (2.2) with BS distribution, Pourmousa *et al.* (2015) obtained the NMVBS distribution as explained in the next definition.

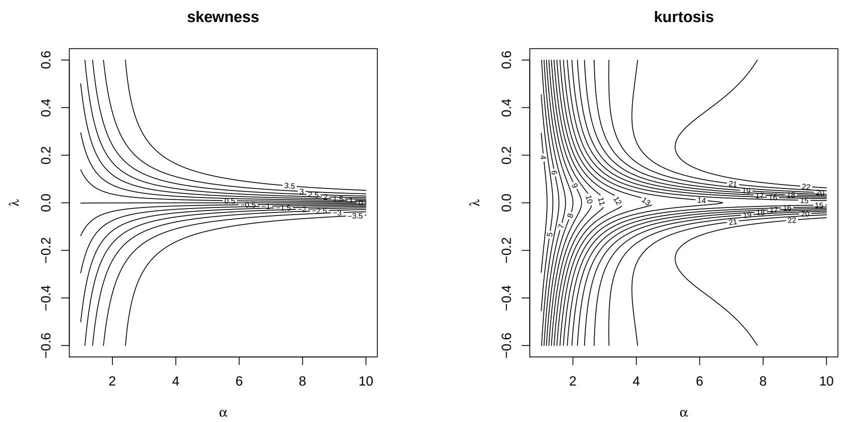

Figure 1: The contours plot of skewness and kurtosis of NMVBS distribution for various values of α and λ.

Definition 2.1. A random variable X is said to have a univariate NMVBS distribution if, in the representation (2.2), $W \sim BS(\alpha, 1)$.

Hereafter, when X has an NMVBS distribution, we shall write $X \sim \text{NMVBS}\,(\mu, \lambda, \sigma^2, \alpha)$. For $\beta = 1$, the NMVBS model is identifiable. It is simple to show that the pdf of NMVBS is a mixture of two pdfs of GH distributions. If $X \sim \text{NMVBS}\,(\mu, \lambda, \sigma^2, \alpha)$, then the pdf of X can be obtained as

$$f_{NMVBS}(x; \mu, \lambda, \sigma^2, \alpha) = \frac{1}{2} f_{GH}(x; \mu, \lambda, \sigma^2, \frac{1}{2}, \frac{1}{\alpha^2}, \frac{1}{\alpha^2}) + \frac{1}{2} f_{GH}(x; \mu, \lambda, \sigma^2, -\frac{1}{2}, \frac{1}{\alpha^2}, \frac{1}{\alpha^2}). \quad (2.5)$$

The skewness and kurtosis of X are obtained as

$$\gamma_x = \frac{\mu_3 - 3\mu_1\mu_2 + 2\mu_1^3}{(\mu_2 - \mu_1^2)^{1.5}} \quad \text{and} \quad \kappa_x = \frac{\mu_4 - 4\mu_1\mu_3 + 6\mu_1^2\mu_2 - 3\mu_1^4}{(\mu_2 - \mu_1^2)^2} - 3,$$

where

$$\begin{aligned}
\mu_1 &= E(X) = \lambda(1 + 0.5\alpha^2),\\
\mu_2 &= E(X^2) = 1 + \lambda^2(1 + 1.5\alpha^4) + \alpha^2(0.5 + 2\lambda^2),\\
\mu_3 &= E(X^3) = \lambda^3(1 + 7.5\alpha^6) + \lambda\Big[3 + \alpha^2\Big\{\alpha^2(4.5 + 9\lambda^2) + 6 + 4.5\lambda^2\Big)\Big\}\Big],\\
\mu_4 &= E(X^4) = \lambda^4(1 + 52.5\alpha^8) + 3(1 + 2\alpha^2 + 1.5\alpha^4)\\
&\quad + \lambda^2\Big\{6 + \alpha^2(27 + 8\lambda^2) + \alpha^4(54 + 30\lambda^2) + \alpha^6(45 + 57.5\lambda^2)\Big\}.
\end{aligned}$$

Figure 1 displays the contours plots of the skewness and kurtosis of the NMVBS distribution for various values of α and λ. It can be observed that the NMVBS distribution takes wider ranges of skewness and kurtosis compared to the skew-normal and skew-t distributions.

In the following theorem, we obtain a conditional distribution which is useful for the parameters estimation via the proposed EM-type algorithm.

Theorem 2.1. *Let X and W be the random variables with $NMVBS(\mu, \lambda, \sigma^2, \alpha)$ and $BS(\alpha, 1)$, respectively. Then, for any $x \in \mathbb{R}$, the pdf of W given $X = x$ is the mixture of two GIG distributions, i.e.,*

$$
\begin{aligned}
f_{W|X=x}(w|x; \mu, \lambda, \sigma^2, \alpha) =& p(x) f_{GIG}(w; 0, \chi(x, \mu, \sigma^2, \alpha), \psi(\lambda, \sigma^2, \alpha)) \\
&+ (1 - p(x)) f_{GIG}(w; -1, \chi(x, \mu, \sigma^2, \alpha), \psi(\lambda, \sigma^2, \alpha)),
\end{aligned} \tag{2.6}
$$

where

$$
p(x) = \frac{f_{GH}(x; \mu, \lambda, \sigma^2, 0.5, \alpha^{-2}, \alpha^{-2})}{f_{GH}(x; \mu, \lambda, \sigma^2, 0.5, \alpha^{-2}, \alpha^{-2}) + f_{GH}(x; \mu, \lambda, \sigma^2, -0.5, \alpha^{-2}, \alpha^{-2})},
$$

$\chi(x, \mu, \sigma^2, \alpha) = \{(x - \mu)/\sigma\}^2 + \alpha^{-2}$, $\psi(\lambda, \sigma^2, \alpha) = (\lambda/\sigma)^2 + \alpha^{-2}$ and $f_{GIG}(., \kappa, \chi, \psi)$ is the pdf of $GIG(\kappa, \chi, \psi)$. Furthermore, for $n = \pm 1, \pm 2, \ldots$,

$$
\begin{aligned}
E\Big[W^n \Big| X = x\Big] = \left(\frac{\chi(x, \mu, \sigma^2, \alpha)}{\psi(\lambda, \sigma^2, \alpha)}\right)^{n/2} \Big[& p(x) R_{(0,n)}\big(\sqrt{\psi(\lambda, \sigma^2, \alpha)\chi(0, \mu, \sigma^2, \alpha)}\big), \\
&+ (1 - p(x)) R_{(-1,n)}\big(\sqrt{\psi(\lambda, \sigma^2, \alpha)\chi(x, \mu, \sigma^2, \alpha)}\big)\Big].
\end{aligned} \tag{2.7}
$$

Proof. By Bayes' rule and properties of GIG distribution, the proof of theorem is obvious. □

3 Finite Mixture of NMVBS distributions

In this section, we show the construction of FM-NMVBS model and demonstrate how to employ the ECM-type algorithm to find ML estimation of its parameters.

A random variable X follows an FM-NMVBS distribution if its pdf is a weighted sum of g components of NMVBS pdfs. More precisely, the pdf of the FM-NMVBS can be written as

$$
f(x; \Theta) = \sum_{i=1}^{g} p_i \, f_{NMVBS}(x; \boldsymbol{\theta}_i), \tag{3.1}
$$

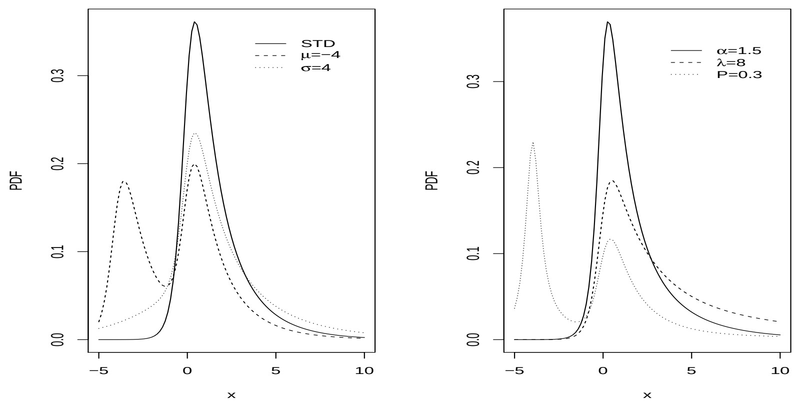

Figure 2: Probability density function of FM-NMVBS.

where p_i's are positive mixing proportions subject to $\Sigma_{i=1}^{g} p_i = 1$, $f_{NMVBS}(.; \boldsymbol{\theta}_i)$ is the NMVBS density obtained in Equation (2.5) with $\boldsymbol{\theta}_i = (\mu_i, \lambda_i, \sigma_i^2, \alpha_i)$ and $\boldsymbol{\Theta} = (p_1, ..., p_{g-1}, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_g)$. Figure 2 presents the curve of (3.1) for $g = 2$. In this figure, STD is represented as the standard case where $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = (0, 1, 1, 1)$ and $p = 0.5$.

Consider n independent random variables $X_1, ..., X_n$, which are taken from FM-NMVBS distribution. For data $\boldsymbol{x} = (x_1, ..., x_n)$, the observed Log-likelihood function of $\boldsymbol{\Theta}$, resulting from (3.1), is

$$\ell(\boldsymbol{\Theta}|X = x) = \sum_{j=1}^{n} log\left(\sum_{i=1}^{g} p_i\, f_{NMVBS}(x_j; \boldsymbol{\theta}_i)\right).$$

Theoretically, the ML estimator of the parameters can be obtained by maximizing $\ell(\boldsymbol{\Theta}|X = x)$ with respect to $\boldsymbol{\Theta}$. However, a direct maximization of this function is complicated and requires numerical algorithms.

An accurate and efficient numerical algorithm is the Expectation-Maximization (EM) which was introduced by Dempster *et al.* (1977). This is the standard tool for ML estimation in FM models. In the EM framework, the key idea is to solve a difficult incomplete-data problem by repeatedly solving tractable complete data problems. To apply this approach to FM-NMVBS model, it is convenient to construct a log-likelihood function by introducing a set of allocation zero-one variables $V_j = (V_{1j}, ..., V_{gj})$ for $j = 1, ..., n$, to describe the unknown population membership. Note that the ith element $V_{ij} = 1$ if y_j belongs to the component i, and $V_{ij} = 0$ otherwise. This implies that

V_j independently follows a multinomial distribution with one trial and probabilities $(p_1, ..., p_g)$, denoted as $V_j \sim M(1; p_1, ..., p_g)$. It also follows from (2.2) that the hierarchical formulation of (3.1) can be represented by

$$\begin{aligned} X_j \big| (W_j = w_j, V_{ij} = 1) &\sim N(\mu_i + w_j\lambda_i, w_j\sigma_i^2), \\ W_j \big| V_{ij} = 1 &\sim BS(\alpha_i, 1), \\ V_j &\sim M(1, p_1, p_2, ..., p_g). \end{aligned} \tag{3.2}$$

Following the hierarchical structure (3.2), based on the observed data $\boldsymbol{x}$, latent data $\boldsymbol{w} = (w_1, ..., w_n)$ and $\boldsymbol{V}_j = (V_1, \ldots, V_n)$, the complete data log-likelihood function of $\boldsymbol{\Theta}$, omitting additive constants, is as follows

$$\begin{aligned} \ell_c(\boldsymbol{\Theta}) = \sum_{j=1}^{n}\sum_{i=1}^{g} V_{ij}\Big[\log p_i - \log \alpha_i - \frac{1}{2}(\log \sigma_i^2) - \frac{(w_j - 1)^2}{2\alpha_i^2 w_j} \\ - \frac{(x_j - \mu_i)^2}{2\sigma_i^2 w_j} - \frac{\lambda_i^2 w_j}{2\sigma_i^2} + \frac{\lambda_i(x_j - \mu_i)}{\sigma_i^2}\Big]. \end{aligned} \tag{3.3}$$

To compute the ML estimates of the unknown parameters involved in (3.3), we apply ECM-type algorithm which is an extension of EM algorithm (Dempster *et al.* , 1977) with the maximization (M) step of EM replaced by a sequence of computationally simpler conditional maximization (CM) steps. For the given initial values $\hat{\boldsymbol{\Theta}}^{(0)}$, the ECM-type algorithm iterates between the following E-step and M-step:

- E-step: At the iteration k, we compute the so-called Q-function, defined as the conditional expected value of complete data log-likelihood (3.3) by fixing $\boldsymbol{\Theta} = \hat{\boldsymbol{\Theta}}^{(k)}$

$$Q(\boldsymbol{\Theta}|\hat{\boldsymbol{\Theta}}^{(k)}) = E[\ell_c(\boldsymbol{\Theta})|X = x, \hat{\boldsymbol{\Theta}}^{(k)}]. \tag{3.4}$$

The necessary conditional expectations involved in (3.4) include $\hat{v}_{ij}^{(k)} = E[V_{ij}|X_j, \hat{\boldsymbol{\Theta}}_i^{(k)}]$, $\hat{w}_{ij}^{(k)} = E[w_j|X_j, V_{ij} = 1, \hat{\boldsymbol{\Theta}}_i^{(k)}]$ and $\hat{t}_{ij}^{(k)} = E[w_j^{-1}|X_j, V_{ij} = 1, \hat{\boldsymbol{\Theta}}_i^{(k)}]$ which can be obtained by using (2.7) in Theorem 2.1 as

$$\begin{aligned} \hat{v}_{ij}^{(k)} &= \frac{\hat{p}_i f_{NMVBS}(x_j; \hat{\boldsymbol{\theta}}_i^{(k)})}{f(x_j; \hat{\boldsymbol{\Theta}}^{(k)})}, \\ \hat{w}_{ij}^{(k)} &= \left(\frac{\chi_{ij}}{\psi_i}\right)^{1/2} \left[p(x)R_{(0,1)}(\sqrt{\psi_i\chi_{ij}}) + (1 - p(x))R_{(-1,1)}(\sqrt{\psi_i\chi_{ij}})\right], \\ \hat{t}_{ij}^{(k)} &= \left(\frac{\psi_i}{\chi_{ij}}\right)^{1/2} \left[p(x)R_{(0,1)}(\sqrt{\psi_i\chi_{ij}}) + (1 - p(x))R_{(1,1)}(\sqrt{\psi_i\chi_{ij}})\right], \end{aligned} \tag{3.5}$$

where $p(x)$, $\chi_{ij} = \chi(x_j, \mu_i, \sigma_i^2, \alpha_i)$, $\psi_i = \psi(\lambda_i, \sigma_i^2, \alpha_i)$ and $R_{(\kappa,a)}(c)$ are defined in (2.6) and (2.3), respectively. So, the Q-function can be written as

$$Q(\boldsymbol{\Theta}|\hat{\boldsymbol{\Theta}}^{(k)}) = \sum_{j=1}^{n}\sum_{i=1}^{g} \hat{v}_{ij}^{(k)}\Big[\log p_i - \log\alpha_i - \frac{1}{2}\log(\sigma_i^2) - \frac{\hat{s}_{ij}^{(k)}}{2\alpha_i^2} - \frac{(x_j-\mu_i)^2}{2\sigma_i^2}\hat{t}_{ij}^{(k)} - \frac{\hat{w}_{ij}^{(k)}\lambda_i^2}{2\sigma_i^2} + \frac{\lambda_i(x_j-\mu_i)}{\sigma_i^2}\Big], \tag{3.6}$$

where $\hat{s}_{ij}^{(k)} = \hat{w}_{ij}^{(k)} + \hat{t}_{ij}^{(k)} - 2$.

- M-step: Let $n_i = \sum_{j=1}^{n} \hat{v}_{ij}^{(k)}$, $A_i = \sum_{j=1}^{n} \hat{v}_{ij}^{(k)}\hat{t}_{ij}^{(k)}$, $B_i = \sum_{j=1}^{n} x_j\hat{v}_{ij}^{(k)}\hat{t}_{ij}^{(k)}$ and $C_i = \sum_{j=1}^{n} \hat{w}_{ij}^{(k)}\hat{v}_{ij}^{(k)}$. To update parameter $\hat{\boldsymbol{\Theta}}_i^{(k)}$, maximize (3.6) over $\hat{\boldsymbol{\Theta}}$. This leads to the following CM estimators:

$$\begin{aligned}
\hat{p}_i^{(k+1)} &= \frac{n_i}{n}, \qquad \hat{\alpha}_i^{(k+1)} = \sqrt{\frac{\sum_{j=1}^{n} \hat{s}_{ij}^{(k)}\hat{v}_{ij}^{(k)}}{n_i}},\\
\hat{\lambda}_i^{(k+1)} &= \frac{A_i\sum_{j=1}^{n} x_j\hat{v}_{ij}^{(k)} - n_iB_i}{A_iC_i - n_i^2},\\
\hat{\mu}_i^{(k+1)} &= \frac{B_i - n_i\hat{\lambda}_i^{(k+1)}}{A_i},\\
\hat{\sigma}_i^{2(k+1)} &= \frac{1}{n_i}\left[\sum_{j=1}^{n} \hat{v}_{ij}^{(k)}\hat{t}_{ij}^{(k)}(x_j - \mu_i^{(k+1)})^2 - \hat{\lambda}_i^{2(k+1)}C_i\right].
\end{aligned}$$

This process is iterated until a suitable convergence rule is satisfied, *e.g.* if $||\hat{\boldsymbol{\Theta}}^{(k+1)} - \hat{\boldsymbol{\Theta}}^{(k)}||$ is sufficiently small, or until some distance involving two successive evaluations of the actual log-likelihood such as $||\ell(\hat{\boldsymbol{\Theta}}^{(k+1)}) - \ell(\hat{\boldsymbol{\Theta}}^{(k)})||$ is small enough.

3.1 Notes on implementation

In the EM algorithm, its slow convergence and the dependence of the algorithm on both the used stopping criterion and the initial values are the main drawbacks. Concerning the stopping criterion in order to avoid an indication of lack of progress of the algorithm (McNicholas *et al.*, 2010), we recommend adopting the Aitken acceleration method (Aitken , 1927) as the stopping criterion. At iteration k, we first compute the Aitken acceleration factor

$$a^{(k)} = \frac{\ell(\hat{\boldsymbol{\theta}}^{(k+1)}) - \ell(\hat{\boldsymbol{\theta}}^{(k)})}{\ell(\hat{\boldsymbol{\theta}}^{(k)}) - \ell(\hat{\boldsymbol{\theta}}^{(k-1)})}.$$

Following Böhning *et al.* (1994), the asymptotic estimate of the log-likelihood at iteration $k+1$ is

$$\ell_{\infty}(\hat{\boldsymbol{\theta}}^{(k+1)}) = \ell(\hat{\boldsymbol{\theta}}^{(k+1)}) + \frac{1}{1-a^{(k)}}\Big\{\ell(\hat{\boldsymbol{\theta}}^{(k+1)}) - \ell(\hat{\boldsymbol{\theta}}^{(k)})\Big\}.$$

As pointed by Lindsay (1995), the algorithm can be considered to have reached convergence when $|\ell_{\infty}(\hat{\boldsymbol{\theta}}^{(k+1)}) - \ell(\hat{\boldsymbol{\theta}}^{(k)})| < \varepsilon$. In our study, the tolerance ε is considered as 10^{-5}.

Since the mixture models may provide a multimodal log-likelihood, the method of estimation via EM algorithm may not give global maximum solution if the starting value ($\hat{\Theta}^{(0)}$) is far from the real parameter value. So, the choice of starting points plays an important role in the parameter estimation. We use the following simple procedure in our study:

- Separate the sample into the g groups using the k-means cluster algorithm via R code *kmeans*.
- Compute the proportion of data points belonging to the same cluster j, and use them as an initial value of p_j.
- For each group, compute the initial values $\mu_j^{(0)}, \lambda_j^{(0)}, \sigma_j^{2(0)}$ and $\alpha_j^{(0)}$ using FM-NMVBS estimation method described above for $g = 1$.

It can be worthwhile to note that the log-likelihood function of FM-NMVBS is related to the normal mixture with unequal variances via representation (3.2). Since in this model the likelihood is unbounded for $\mu = x$ and $\sigma \to 0$, the log-likelihood of FM-NMVBS is unbounded and corresponding global maximum likelihood estimator (MLE) is undefined. One way to solve this problem is to put a constraint on the parameter space such that the likelihood is bounded. Not only identification of appropriate parameter space is difficult, but also, in many cases, different choices of the parameter space may give different constrained global ML estimators. A practical approach to find the constraint parameter is to propose a profile log-likelihood to solve the unboundedness issue of the likelihood function. More details of the profile log-likelihood method can be found in Yao (2010).

3.2 Estimation of observed information matrix

A practical way to compute the asymptotic covariance of the ML estimates is the information-based method (Basford *et al.* , 1997). Based on the Meilijson's formula (Meilijson , 1989), the empirical information matrix is

$$I_e(\boldsymbol{\theta}|\boldsymbol{x}) = \sum_{j=1}^{n} \mathbf{s}(x_j|\boldsymbol{\theta})\mathbf{s}^T(x_j|\boldsymbol{\theta}) - \frac{1}{n}\mathbf{S}(x|\boldsymbol{\theta})\mathbf{S}^T(x|\boldsymbol{\theta}), \tag{3.7}$$

where $\mathbf{S}(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{j=1}^{n} \mathbf{s}(x_j|\boldsymbol{\theta})$ and the individual score $\boldsymbol{s}(x_j|\boldsymbol{\theta})$ can be determined from the result of Louis (1982) as

$$\boldsymbol{s}(x_j|\boldsymbol{\theta}) = \frac{\partial f(x_j|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = E\Big(\frac{\partial \ell_c(\boldsymbol{\theta}|x_j, w_j)}{\partial \boldsymbol{\theta}}\Big|x_j, \boldsymbol{\theta}\Big).$$

Substituting the ML estimates ($\hat{\boldsymbol{\theta}}$) for theta in (3.7), the empirical information matrix reduces to

$$I_e(\boldsymbol{\theta}|\boldsymbol{X}) = \sum_{j=1}^{n} \hat{\mathbf{s}}_j\hat{\mathbf{s}}_j^T, \tag{3.8}$$

where $\hat{\mathbf{s}}_j$ is an individual score vector with the elements

$$(\hat{\mathbf{s}}_{j,p_1}, ..., \hat{\mathbf{s}}_{j,p_{g-1}}, \hat{\mathbf{s}}^T_{j,\mu_1}, ..., \hat{\mathbf{s}}_{j,\mu_g}, \hat{s}_{j,\lambda_1}, ..., \hat{\mathbf{s}}_{j,\lambda_g}, \hat{\mathbf{s}}_{j,\sigma_1^2}, ..., \hat{\mathbf{s}}_{j,\sigma_g^2}, \hat{\mathbf{s}}_{j,\alpha_1}, ..., \hat{\mathbf{s}}_{j,\alpha_g}),$$

which can be computed, for $r = 1, ..., g$, as

$$\begin{aligned}
\hat{\mathbf{s}}_{j,p_r} &= \frac{\hat{\upsilon}_{rj}}{\hat{p}_r} - \frac{\hat{\upsilon}_{gj}}{\hat{p}_g}, \\
\hat{\mathbf{s}}_{j,\alpha_r} &= \hat{\upsilon}_{rj}(\frac{\hat{s}_{rj}}{\hat{\alpha}_r^3} - \frac{1}{\hat{\alpha}_r}), \\
\hat{\mathbf{s}}_{j,\mu_r} &= \hat{\upsilon}_{rj}\hat{\sigma}_r^{-2}\Big[(x_j - \hat{\mu}_r)\hat{t}_{rj} - \hat{\lambda}_r\Big], \\
\hat{\mathbf{s}}_{j,\lambda_r} &= \hat{\upsilon}_{rj}\hat{\sigma}_r^{-2}\Big[(x_j - \hat{\mu}_r) - \hat{\lambda}_r\hat{w}_{rj}\Big], \\
\hat{\mathbf{s}}_{j,\sigma_r^2} &= \frac{\hat{\upsilon}_{rj}}{2\hat{\sigma}_r^4}\left[\hat{t}_{rj}(x_j - \hat{\mu}_r)^2 + \hat{w}_{rj}\hat{\lambda}_r^2 - 2\hat{\lambda}_r(x_j - \hat{\mu}_r) - \hat{\sigma}_r^2\right],
\end{aligned}$$

where $\hat{v}_{rj}$, $\hat{w}_{rj}$ and $\hat{t}_{rj}$ are obtained by replacing $\hat{\mathbf{\Theta}}_r$ in (3.5). Therefore, the standard error of the ML estimator $\hat{\boldsymbol{\theta}}$ can be obtained by calculating the square roots of the diagonal elements of the inverse of (3.8).

4 Data analyses

4.1 Simulation study

In this subsection, we investigate bias and mean square error as two asymptotic properties of the estimates obtained using the suggested ECM algorithm. We consider two sets of the parameter values $\big(\mathbf{\Theta} = (p, \mu_1, \lambda_1, \sigma_1, \alpha_1, \mu_2, \lambda_2, \sigma_2, \alpha_2)\big)$ for this study as

$$\textit{Poorly separated}: \quad \mathbf{\Theta} = (0.4, -2, -2, 3, 1, 2, 1, 4, 0.5),$$
$$\textit{Well separated}: \quad \mathbf{\Theta} = (0.6, -3, 1, 1, 0.2, 3, 0.6, 1, 0.2).$$

The histogram of one sample taken from a FM-NMVBS population with above parameters and $n = 1000$ is shown in Figure 3. The left panel shows a mixture of NMVBS observations that largely overlap, meaning that the data are poorly separated. Although we have a two-component mixture in this figure, the histogram need not to be bimodal. On the other hand, for the well separated components, the histogram (the right panel) is bimodal.

For each combination of parameters and sample sizes $n = 100, 500, 1000, 5000$, we generate 500 samples from the FM-NMVBS model. Then, the absolute relative bias (R.Bias) and the mean squared error (MSE) are computed over all samples. For each parameter θ, they are defined as

$$R.Bias = \frac{1}{500}\sum_{i=1}^{500}\left|\frac{\hat{\theta}_i - \theta}{\theta}\right|, \qquad MSE = \frac{1}{500}\sum_{i=1}^{500}(\hat{\theta}_i - \theta)^2,$$

where $\hat{\theta}_i$ is the estimate of θ_i when the data constitutes the sample i.

Table 1 presents the results for the poorly and well separated cases. As a general rule, it is evident that the bias of the estimators decreases when the sample size increases; so does MSE. In addition, the values of MSE for $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ are more than 2 for the small sample (≤ 100). It is also observed that both R.bais and MSE of the scale, shape ($\boldsymbol{\alpha}$) and mixing proportion parameters in the well separated case are smaller than those of poorly separated. The worst case of estimation happens in estimating the scale parameter σ_2 which implies that, in the poorly separated case, the required sample size for obtaining a reasonable pattern of convergence is greater than 1000.

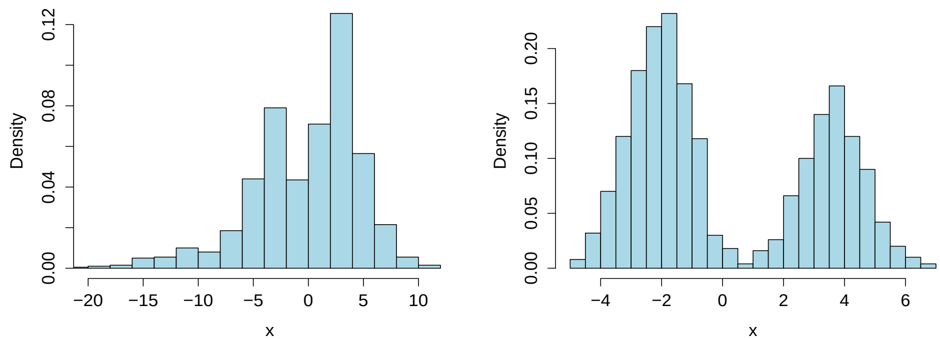

Figure 3: Artificial data with two components: poorly separated (left) and well separated (right) components.

4.2 Real data analysis

A powerful laser-based technique for scanning, outlining and sorting microscopic particles flowing in a stream of water is flow cytometry. Because of using this technique in both clinical and research biotechnological programs for rapid single-cell investigation of surface and intracellular markers, it is used in the literature and in a large number of biomedical applications such as molecular and cellular biology (*e.g.*, to measure DNA content), hematology (*e.g.*, to test leukemia samples) and immunology (*e.g.*, to conduct CD4 tests for HIV AIDS). Also, It was recently shown by Frühwirth-Schnatter and Pyne (2010) and Ho *et al.* (2012) that flow cytometric data are ideally suited for multimodal non-Gaussian (asymmetric) finite mixture modeling. Glynn (2006) provided a working dataset of flow cytometry in CC4-067-BM.fcs which consists of ten attributes measured on 5,634 cells. To illustrate our univariate mixture modeling approach, we analyze the data from the channel APC-Cy7. Considering the scale transformation (y-a)/(b-a) suggested by Hahne *et al.* (2009), we preprocessed the data by setting b = 1000 and a = 0 to compare our new FM-NMVBS model with FM-N and finite mixture of scale mixtures of skew-normal distribution models (including: skew-normal (FM-SN), skew-*t* (FM-ST), skew-contaminated-normal (FM-SCN) and skew-slash (FM-SSL)).

Table 2 shows the ML estimates along with the associated standard errors for the best fitted FM-NMVBS model with the corresponding values for the other five competing 2-component mixture models. Moreover, the values of Log-likelihood, AIC (Akaike , 1974) and BIC (Schwarz , 1978) reported in this table show that the FM-NMVBS provides the best fit. This result can also be seen from the histogram of the data and

Table 1: Bias and MSE for EM estimates of simulated data.

Measure	parameter	Poor				Well			
		sample size				sample size			
		100	250	500	1000	100	250	500	1000
R.Bias	p	0.0862	0.0729	0.0621	0.0357	0.0018	0.0013	0.0009	0.0006
	μ_1	0.0375	0.0353	0.0282	0.0088	0.2263	0.2087	0.1953	0.0852
	μ_2	0.2615	0.0222	0.0059	0.0020	0.1918	0.1678	0.1424	0.0962
	λ_1	0.2045	0.1487	0.0639	0.0099	0.6326	0.6183	0.5814	0.3259
	λ_2	0.4245	0.2351	0.1894	0.0865	0.9552	0.7918	0.6127	0.4298
	σ_1	0.2248	0.2119	0.1894	0.1561	0.1035	0.0378	0.0117	0.0085
	σ_2	0.5462	0.5315	0.5023	0.4238	0.1669	0.0618	0.0288	0.0139
	α_1	0.2389	0.2112	0.1815	0.1267	0.2809	0.1797	0.1134	0.0892
	α_2	0.0732	0.0471	0.0328	0.0125	0.0678	0.0325	0.0184	0.0057
MSE	p	0.0074	0.0033	0.0024	0.0013	0.0022	0.0010	0.0006	0.0002
	μ_1	3.1369	0.4611	0.1810	0.0272	2.5060	1.2196	0.7417	0.5474
	μ_2	4.4479	1.3482	0.5332	0.1880	3.0943	1.4335	0.8251	0.2395
	λ_1	3.1597	1.4629	0.9936	0.0467	2.3681	1.1773	0.7265	0.5328
	λ_2	4.5796	1.3422	0.5781	0.2928	2.9486	1.3661	0.8018	0.5124
	σ_1	1.5540	0.8816	0.5353	0.3588	0.0322	0.0081	0.0034	0.0013
	σ_2	5.0129	4.6542	4.5205	3.0078	0.0692	0.0164	0.0058	0.0022
	α_1	0.1264	0.0703	0.0466	0.0353	0.0892	0.0634	0.0179	0.0117
	α_2	0.0314	0.0174	0.0098	0.0019	0.0197	0.0126	0.0066	0.0038

the estimated pdf of models, plotted in Figure 4.

5 Conclusion

In this paper, we have dealt with a new family of mixture models based on the the NMVBS distribution, called the FM-NMVBS model, as a new distribution for flexible model-based clustering. This family of mixture models is attractive for modeling data as it can account for groups of data exhibiting patterns of asymmetry, multimodality and fat tails. We have presented a convenient hierarchical representation for the FM-NMVBS distribution and developed a simple ECM-type algorithm. The computer program coded in R language is available from the authors upon request. Numerical results suggest that the proposed FM-NMVBS is well suited to the experimental data and can be more robust against fat tailed and skewed observations than the FM models obtained by the scale mixture of skew-normal distribution competitors. The advantage of our current approach can be extended to the multivariate case and Bayesian approach of the FM-NMVBS model.

Table 2: Parameter estimates of APC-Cy7 data.

	FM-NMVBS		FM-ST		FM-SN		FM-SCN		FM-SSL		FM-N	
parameter	MLE	SE	MLE	SE	MLE	SE	MLE	SE	MLE	SE	MLE	SE
p	0.3617	0.0082	0.3681	0.0093	0.3871	0.0077	0.3610	0.0231	0.3629	0.0087	0.3703	0.0068
μ_1	-0.1885	0.0089	0.1046	0.0101	0.0799	0.0084	0.1219	0.0102	0.1193	0.0109	0.6933	0.0114
μ_2	2.7556	0.0095	2.8432	0.0101	2.8590	0.0110	2.8422	0.0124	2.8563	0.0101	2.5054	0.0059
σ_1	0.0697	0.0158	0.6793	0.0495	0.8549	0.0333	0.5270	0.0275	0.5465	0.0380	0.4407	0.0097
σ_2	0.2598	0.0264	0.4269	0.0127	0.4698	0.0125	0.3531	0.0253	0.3747	0.0111	0.3239	0.0027
λ_1	0.7883	0.1657	7.7599	1.2637	11.3959	1.7122	5.9769	0.8942	6.3904	0.9681	–	
λ_2	-0.2094	0.0458	-2.6398	0.1978	-2.4889	0.1599	-2.7527	0.3139	-2.8519	0.2074	–	
α_1	0.5786	0.8451	–		–		–		–		–	
α_2	0.9044	0.7564	–		–		–		–		–	
ν_1	–		5.4595	0.8995	–		0.3785	0.0478	1.7889	0.2392	–	
ν_2	–		5.4595	0.8995	–		0.1950	0.0537	–		–	
$\ell(\Theta)$	**-5403.468**		-5423.677		-5463.05		-5415.013		-5415.013		5795.045	
AIC	**10824.94**		10863.35		10940.1		10848.03		10846.03		11600.09	
BIC	**10884.67**		10916.45		10986.56		10907.75		10899.12		11633.27	

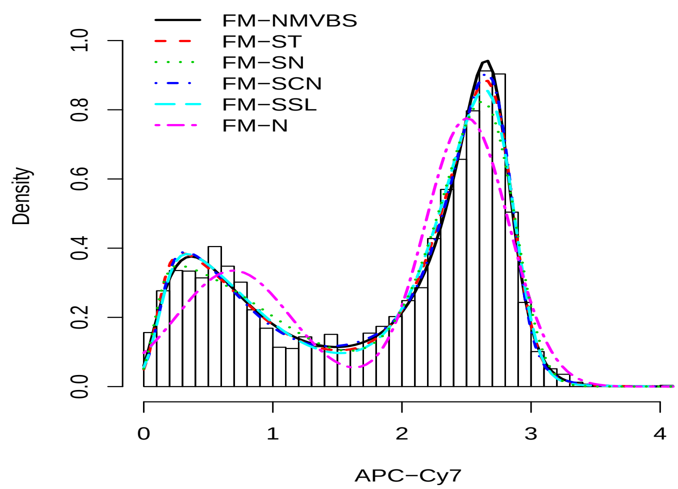

Figure 4: Histogram of the APC-Cy7 data with six fitted mixture models.

References

Aitken, A. C. (1927), On Bernoulli's Numerical Solution of Algebraic Equations. *Proceedings of the Royal Society of Edinburgh*, **46**, 289-305.

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

Arslan, O. (2010), An alternative multivariate skew Laplace distribution: properties and estimation. *Statistical Papers*, **51**, 865-887.

Arslan, O. (2015), Variance-mean mixture of the multivariate skew normal distribution. *Statistical Papers*, **56**, 353-378.

Barndorff-Nielsen, O. (1977), Exponentially decreasing distributions for the logarithm of particle size. *In Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* (Vol. 353, No. 1674, pp. 401-419). The Royal Society.

Basford, K. E., Greenway, D. R., McLachlan, G. J., and Peel, D. (1997), Standard errors of fitted component means of normal mixtures. *Computational Statistics*, **12**, 1-18.

Basso, R. M., Lachos, V. H., Cabral, C. R. B., and Ghosh, P. (2010), Robust mixture modeling based on scale mixtures of skew-normal distributions. *Computational Statistics and Data Analysis*, **54**, 2926-2941.

Birnbaum, Z. W., and Saunders, S. C. (1969a), A new family of life distributions. *Journal of Applied Probability*, 319-327.

Birnbaum, Z. W., and Saunders, S. C. (1969b), Estimation for a family of life distributions with applications to fatigue. *Journal of Applied Probability*, 328-347.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., and Lindsay, B. (1994), The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics*, **46**, 373-388.

Cabral, C. R. B., Lachos, V. H., and 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., and Rubin, D. B. (1977), Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, 1-38.

Desmond, A. F. (1986), On the relationship between two fatigue-life models. *Reliability, IEEE Transactions on*, **35**, 167-169.

Frühwirth-Schnatter, S. (2006), *Finite mixture and Markov switching models*. Springer Science and Business Media.

Frühwirth-Schnatter, S., and Pyne, S. (2010), Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. *Biostatistics*, **11**, 317-336.

Glynn, E.F. (2006), FCSExtract Utility. Stowers Institute for Medical Research. Online available at: http://research.stowers-institute.org/efg/ScientificSoftware/Utility/FCSExtract/.

Good, I. J. (1953), The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**, 237-264.

Ho, H. J., Pyne, S., and Lin, T. I. (2012), Maximum likelihood inference for mixtures of skew Student-t-normal distributions through practical EM-type algorithms. *Statistics and Computing*, **22**, 287-299.

Hahne, F., LeMeur, N., Brinkman, R. R., Ellis, B., Haaland, P., Sarkar, D., ... and Gentleman, R. (2009). flowCore: A Bioconductor package for high throughput flow cytometry. *Bmc Bioinformatics*, **10**, 106.

Lin, T. I. (2010), Robust mixture modeling using multivariate skew t distributions. *Statistics and Computing*, **20**, 343-356.

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

Lindsay, B.G., (1995), Mixture models: theory, geometry and applications. In NSF-CBMS regional conference series in probability and statistics, (pp. i-163). Institute of Mathematical Statistics and the American Statistical Association.

Louis, T. A. (1982), Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 226-233.

Meilijson, I. (1989), A fast improvement to the EM algorithm on its own terms. *Journal of the Royal Statistical Society. Series B (Methodological)*, 127-138.

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

McLachlan, G. J., and Basford, K. E. (1988), *Mixture models. Inference and applications to clustering*. Statistics: Textbooks and Monographs, New York: Dekker.

McNeil, A. J., Frey, R., and Embrechts, P. (2005), *Quantitative risk Management: Concepts, Techniques and Tools*. Princeton university press.

McNicholas, P.D., Murphy, T.B., McDaid, A.F., Frost, D. (2010), Serial and parallel implementations of modelbased clustering via parsimonious Gaussian mixture models. *Computational Statistics and Data Analysis*, **54**, 711-723.

Nematollahi, N., Farnoosh, R., and Rahnamaei, Z. (2016), Location-scale mixture of skew-elliptical distributions: Looking at the robust modeling. *Statistical Methodology*, **37**, 131-146.

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

Pourmousa, R., Jamalizadeh, A., and Rezapour, M. (2015), Multivariate normal mean-variance mixture distribution based on BirnbaumSaunders distribution. *Journal of Statistical Computation and Simulation*, **85**, 2736-2749.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria, 2015.

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

Yao, W. (2010), A profile likelihood method for normal mixture with unequal variance. *Journal of Statistical Planning and Inference*, **140**, 2089-2098.