Stat Methods Appl manuscript No.
(will be inserted by the editor)

Model-based clustering and classification with non-normal mixture distributions

Sharon X. Lee · Geoffrey J. McLachlan

the date of receipt and acceptance should be inserted later

Abstract Non-normal mixture distributions have received increasing attention in recent years. Finite mixtures of multivariate skew-symmetric distributions, in particular, the skew normal and skew t-mixture models, are emerging as promising extensions to the traditional normal and t-mixture models. Most of these parametric families of skew distributions are closely related, and can be classified into four forms under a recently proposed scheme, namely, the restricted, unrestricted, extended, and generalised forms. In this paper, we consider some of these existing proposals of multivariate non-normal mixture models and illustrate their practical use in several real applications. We first discuss the characterizations along with a brief account of some distributions belonging to the above classification scheme, then references for software implementation of EM-type algorithms for the estimation of the model parameters are given. We then compare the relative performance of restricted and unrestricted skew mixture models in clustering, discriminant analysis, and density estimation on six real datasets from flow cytometry, finance, and image analysis. We also compare the performance of mixtures of skew normal and t-component distributions with other non-normal component distributions, including mixtures with multivariate normal-inverse-Gaussian distributions, shifted asymmetric Laplace distributions and generalized hyperbolic distributions.

Keywords Mixture models · Skew distributions · Multivariate skew normal distribution · Multivariate skew t-distribution · EM algorithm

G. J. McLachlan
Department of Mathematics, University of Queensland,
St Lucia, 4072, Australia
E-mail: g.mclachlan@uq.edu.au

1 Introduction

Finite mixtures of symmetric distributions, in particular normal mixtures, are an important tool in statistical modelling and analysis. In recent years, mixtures of asymmetric distributions have emerged as a powerful alternative to the traditional normal and t-mixture models. For a comprehensive survey of mixture models and their applications, the reader is referred to the monographs by Everitt and Hand (1981), Titterington et al. (1985), McLachlan and Basford (1988), Lindsay (1995), Böhning (1999), McLachlan and Peel (2000), Frühwirth-Schnatter (2006), and the edited volume of Mengersen et al. (2011); see also the papers by Banfield and Raftery (1993) and Fraley and Raftery (1999).

The past few years have seen an increasing use of skew mixture distributions to provide improved modelling and clustering of data that consists of asymmetric clusters with outliers. They have been widely applied to datasets from a variety of fields, including biostatistics, bioinformatics, finance, image analysis, and medical science, among others. Some recent examples are given in Pyne et al. (2009a), Soltyk and Gupta (2011), Contreras-Reyes and Arellano-Valle (2012), and Riggi and Ingrassia (2013).

Recent advances in the mixture model-based clustering literature has focused on the development of mixture distributions with more flexible parametric component densities that can better accommodate data exhibiting non-normal features, including asymmetry, multimodality, heavy-tails, and the presence of outliers. Some notable examples include the skew normal mixture model (Lin, 2009, Cabral et al., 2012), the skew t-mixture model (Lin, 2010, Lee and McLachlan, 2011, Vrbik and McNicholas, 2012), the skew t-normal mixture model (Lin et al., 2013), and some other non-elliptical approaches (Karlis and Xekalaki, 2003, Franczak et al., 2012, Browne and McNicholas, 2013).

In this paper, the aforementioned non-normal mixture distributions are discussed. Particular attention is paid to the two most popular skew symmetric models, namely, the skew normal and skew t-mixture distributions, which have received increasing attention in the past few years. Following Lee and McLachlan (2013c), these component distributions can be systematically classified as the restricted, unrestricted, extended, and generalized versions based on their characterizations. This aids in the understanding of the link between various algorithms applied to the fitting of these distributions and mixtures of them.

Lee and McLachlan (2013c) showed that the same scheme can be extended directly to study other families of multivariate skew symmetric distributions.

One of the major uses of finite mixture models is in clustering and supervised classification (discriminant analysis). In cluster analysis applications, the usual goal is to provide a partition of the data into several groups or clusters, based on the assumption of cluster homogeneity; that is, observations from the same cluster are more similar to each other than those from a different cluster. Under this setting, each observation is assumed to come from one of

the components of the mixture model, and the probabilistic clustering of the data is based on their estimated posterior probabilities of component membership in the mixture model. Clustering is an unsupervised approach, where no information of group memberships are given, even if the existence of groups were known *a priori*. In contrast, in supervised classification, the group memberships for some observations are known for each group. We illustrate the usefulness of these skew mixture models in both situations in handling data with non-normal features.

The remainder of the paper is organized as follows. In Section 2, we give an overview of various multivariate skew symmetric distributions. In Section 3, we discuss several recently proposed mixtures of multivariate skew distributions, and the availability of software for these fitting in practice. In Section 4, we briefly outline several other flexible multivariate mixture models which are not elliptically-contoured. The non-normal mixture models are illustrated on a flow cytometric dataset in Section 5.1, where the four-component skew mixture models compare favourably with other state-of-the-art automated gating algorithms. Section 5 presents some further applications to real datasets, including the clustering of health and finance related data, estimation of the Value-at-Risk from a portfolio of Australian stock returns, discriminant analysis of wheat kernels, and natural colour image segmentation. Finally, we conclude with some remarks on the performance of restricted and unrestricted skew symmetric models in Section 6.

2 Multivariate skew symmetric distributions

We begin with a brief discussion of multivariate skew symmetric distributions, in particular, the skew normal and skew t-distributions. In Lee and McLachlan (2013c), the existing multivariate skew normal and skew t-distributions have been classified into four forms according to their characterizations, namely, the restricted, unrestricted, extended, and generalized forms. The same scheme can be applied to classify the broader class of multivariate skew symmetric distributions.

An asymmetric density can be generated by perturbing a symmetric density, yielding a so-called multivariate skew symmetric (MSS) density (Azzalini and Capitanio, 2003). Typically, a MSS density can be expressed as a product of a (multivariate) symmetric function $f_p(\cdot)$ and a perturbation (or skewing) function $h_q(\cdot)$, where $h_q(\cdot)$ is a function that maps a q-dimensional parameter into the scalar unit interval; that is,

$$f_p(\boldsymbol{y};\boldsymbol{\mu})\, h_q(\cdot), \tag{1}$$

where $f_p(\cdot)$ is symmetric around $\boldsymbol{\mu}$. Based on this formulation, a MSS density can be classified into different forms according to the value of q and the functional form of $h_q(\cdot)$. For example, when $q = 1$ and $h(\cdot) = 2F_1(\cdot)$, where $F_1(\cdot)$ denotes the (univariate) distribution function corresponding to $f_p(\cdot)$, we obtain a restricted characterization. Assuming $f_p(\cdot)$ is a p-dimensional symmetric

density, if $q = p$ and $h_p(\cdot) = 2^p F_p(\cdot)$, where $F_p(\cdot)$ denotes the p-dimensional distribution function corresponding to $f_p(\cdot)$, then the MSS density is said to have an unrestricted form. We shall discuss now some of the important skew symmetric distributions using the four forms introduced in Lee and McLachlan (2013c).

2.1 Restricted multivariate skew distributions

The study of skew distributions was pioneered by the work on the (univariate) skew-normal distribution of Azzalini (1985) and its extension to the multivariate case by Azzalini and Dalla Valle (1996). This and some of the first few skew distributions appearing in the literature belong to the restricted form. To establish notation, a random vector $\boldsymbol{Y}$ has a (restricted) multivariate skew normal (rMSN) distribution if its density is given by

$$f(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\delta}) = 2\phi_p(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma})\,\Phi_1\left(\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu});0,1-\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}\right), \qquad (2)$$

where $\boldsymbol{\mu}$ is a location vector, $\boldsymbol{\Sigma}$ is a scale matrix, and $\boldsymbol{\delta}$ is a skewness vector. Here, we let $\phi_p(.;\boldsymbol{\mu},\boldsymbol{\Sigma})$ be the density of the p-variate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and $\Phi_p(.;\boldsymbol{\mu},\boldsymbol{\Sigma})$ be the corresponding distribution function. The rMSN distribution admits a convenient stochastic representation, as given in Pyne et al. (2009a),

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\delta}|U_0| + \boldsymbol{U}_1, \qquad (3)$$

where U_0 and $\boldsymbol{U}_1$ are independent normal random variables. More specifically, U_0 is a standard (scalar) normal variable, or equivalently, $|U_0| \sim HN(0,1)$, where $HN(0,1)$ denotes the standard (univariate) half-normal distribution, and $\boldsymbol{U}_1$ has a centered p-variate normal distribution with covariance matrix $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} - \boldsymbol{\delta}\boldsymbol{\delta}^T$. Alternatively, (2) can be formulated via a conditioning approach, given by

$$\boldsymbol{Y} = \boldsymbol{\mu} + (\boldsymbol{Y}_1 \mid Y_0 > 0), \qquad (4)$$

where $\boldsymbol{Y}_1 \sim N_p(\boldsymbol{0},\boldsymbol{\Sigma})$, $Y_0 \sim N(0,1)$ and $\mathrm{cov}(\boldsymbol{Y}_1, Y_0) = \boldsymbol{\delta}$. Following Azzalini and Dalla Valle (1996), the notation $(\boldsymbol{Y}_1 \mid Y_0 > 0)$ implies the vector $\boldsymbol{Y}_1$ if $Y_0 > 0$ and $-\boldsymbol{Y}_1$ otherwise.

The (restricted) multivariate skew t-distribution is a natural extension of (3), in which the random variables U_0 and $\boldsymbol{U}_1$ now follow a corresponding t-distribution. Following Pyne et al. (2009a), the p-variate (restricted) multivariate skew t (rMST) distribution is given by

$$f(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\delta},\nu) = 2t_p(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)T_1\left(\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\sqrt{\frac{\nu+p}{\nu+d(\boldsymbol{y})}};0,\lambda,\nu+p\right), \qquad (5)$$

where $\lambda = 1 - \boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$, $d(\boldsymbol{y}) = (\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$ is the squared Mahalanobis distance between $\boldsymbol{y}$ and $\boldsymbol{\mu}$ with respect to $\boldsymbol{\Sigma}$. Here, we let $t_p(.;\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)$ denote the p-dimensional t-distribution with location vector $\boldsymbol{\mu}$, scale matrix

$\boldsymbol{\Sigma}$, and degrees of freedom ν. Also, we let $T_p(.;\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)$ be the corresponding distribution function. Similar to the rMSN distribution, the rMST distribution can be obtained via two stochastic mechanisms, the convolution of multivariate t- and truncated t-variables, and the conditioning of t-variables. There exists various definitions and/or variants of the multivariate skew t-distributions in the literature. It was pointed out in Arellano-Valle and Azzalini (2006) and Lee and McLachlan (2013c) that some of these formulations are equivalent after appropriate reparameterizations. In particular, it was noted in Lee and McLachlan (2013c) that the versions considered by Branco and Dey (2001), Azzalini and Capitanio (2003), Gupta (2003), and Lachos et al. (2010) are equivalent to the rMST distribution (5).

The MSN and MST distributions are only two important special cases of a broader class of skew distributions. The skewing mechanism can be applied directly to other parametric densities to produce a broader family of asymmetric distributions, such as the extensively studied class of skew-elliptical distributions; see, for example, Azzalini and Capitanio (1999), Branco and Dey (2001), and Azzalini and Capitanio (2003). This more general family originates from the elliptically-contoured (EC) distributions, whose densities are constant on ellipsoids (Fang et al., 1990); hence its name. The p-dimensional EC density, denoted by $EC(\boldsymbol{\mu},\boldsymbol{\Sigma};\tilde{f})$, takes the form

$$f_p(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma}) = c\,|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\tilde{f}\left((\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right), \tag{6}$$

where $\tilde{f}$ is any suitable parametric function from $\mathbb{R}^+$ to $\mathbb{R}^+$, known as the density generator of f_p and c is a normalizing constant. Then the density of the (restricted) multivariate skew-elliptical (rMSE) class takes the form

$$f(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\delta}) = 2\; f_p(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma})F_{d(\boldsymbol{y})}(w;0,\lambda), \tag{7}$$

where $f_p(.)$ is the density of an elliptically-contoured random variable as defined in (6), $w = \boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$, $F_{d(\boldsymbol{y})}(.;\boldsymbol{\mu},\boldsymbol{\Sigma})$ is the distribution function corresponding to $EC(\boldsymbol{\mu},\boldsymbol{\Sigma};\tilde{f}_{d(\boldsymbol{y})})$, and

$$\tilde{f}_{d(\boldsymbol{y})}(w) = \frac{\tilde{f}(w+d(\boldsymbol{y}))}{\tilde{f}(d(\boldsymbol{y}))}. \tag{8}$$

This class can be obtained by (3) and (4), with $\boldsymbol{Y}_1$ and Y_0 following EC distributions and $\boldsymbol{U}_1$ and U_0 have a joint EC distribution.

2.2 Unrestricted multivariate skew distributions

As mentioned previously, the unrestricted form corresponds to (1) with $q = p$. This corresponds to replacing the latent variable U_0 in (3) and Y_0 in (4) with a p-dimensional version, $\boldsymbol{U}_0$ and $\boldsymbol{Y}_0$, respectively. In this case, the constraint $\boldsymbol{Y}_0 > \boldsymbol{0}$ is taken as a set of component-wise inequalities, that is, $Y_{0i} > 0$ for

$i = 1, \ldots, p$. This family of (unrestricted) multivariate skew-elliptical (uMSE) class was studied in Sahu et al. (2003), the density of which is given by

$$f(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\delta}) = 2^p f_p(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) F_{d(\boldsymbol{y})}(\boldsymbol{\Delta}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}); \boldsymbol{0}, \boldsymbol{\Lambda}), \tag{9}$$

where $\boldsymbol{\Delta}$ is a diagonal matrix with elements given by $\boldsymbol{\delta}$, and $\boldsymbol{\Lambda} = \boldsymbol{I}_p - \boldsymbol{\Delta}\boldsymbol{\Omega}^{-1}\boldsymbol{\Delta}$. It should be stressed that the rMSE family and uMSE family match only in the univariate case, and one cannot obtain (7) from (9) when $p > 1$.

For the sake of completeness, we include the density of the unrestricted skew normal (uMSN) distribution and the unrestricted skew t (uMST) distribution here, given by

$$f(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\delta}) = 2^p \phi_p(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \Phi_p(\boldsymbol{\Delta}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}); \boldsymbol{0}, \boldsymbol{\Lambda}), \tag{10}$$

and

$$f(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\delta}, \nu) = 2^p t_p(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)\ T_p\left(\boldsymbol{\Delta}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})\sqrt{\frac{\nu + p}{\nu + d(\boldsymbol{y})}}; \boldsymbol{0}, \boldsymbol{\Lambda}, \nu + p\right), \tag{11}$$

respectively, where $\Lambda = \boldsymbol{I}_p - \boldsymbol{\Delta}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}$. The convolution- and conditioning-type stochastic representations for the unrestricted case extend from (3) and (4) directly, given by $\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\delta}|\boldsymbol{U}_0| + \boldsymbol{U}_1$ and $\boldsymbol{Y} = \boldsymbol{\mu} + (\boldsymbol{Y}_1 \mid \boldsymbol{Y}_0 > \boldsymbol{0})$, respectively, where, for the uMSN case, $\boldsymbol{U}_0$ and $\tilde{\boldsymbol{\Sigma}}^{\frac{1}{2}}\boldsymbol{U}_1$ are independent $N_p(\boldsymbol{0}, \boldsymbol{I}_p)$ random vectors, $\boldsymbol{Y}_0 \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$, $\boldsymbol{Y}_1 \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, and $\mathrm{cov}(\boldsymbol{Y}_1, \boldsymbol{Y}_0) = \boldsymbol{\Delta}$. Here we let $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} - \boldsymbol{\Delta}^2$.

2.3 Extended multivariate skew distributions

A more general extension of the rMSE and uMSE class is to consider a set of q constraints simultaneously on the latent variable of the form $\boldsymbol{Y}_0 + \boldsymbol{\tau} > \boldsymbol{0}$, which implies $Y_{0i} + \tau_i > 0$ for $i = 1, \ldots, q$. The random vector $\boldsymbol{\tau} \in \mathbb{R}$ can be interpreted as the mean vector of $\boldsymbol{Y}_0$. The density of this class can be easily computed via a well-known general relationship (see Theorem 5.1 of Arellano-Valle et al. (2002)), which can be written as

$$f(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Delta}) = f_p(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \frac{F_{d(\boldsymbol{y})}(\boldsymbol{\tau} + \boldsymbol{\Delta}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}); \boldsymbol{0}, \boldsymbol{\Lambda})}{F(\boldsymbol{\tau}; \boldsymbol{0}, \boldsymbol{\Gamma})}, \tag{12}$$

where $\boldsymbol{\Lambda}$ and $\boldsymbol{\Gamma}$ are positive-definite matrices and $F(.; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the distribution function corresponding to $EC(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \tilde{f})$. We shall refer to this as the extended MSE (eMSE) family. Observe that when $\boldsymbol{\tau} = \boldsymbol{0}$, the normalizing constant in the denominator of (12) reduces to 2^{-q}, which corresponds to the coefficient of 2 and 2^p in the rMSE and uMSE density, respectively.

This form incorporates the restricted and unrestricted class with appropriate restrictions on the parameters. The eMSE class, or the unified skew-elliptical (SUE) family, has been studied by various authors, although with

a slightly different parameterization, including Arellano-Valle and Azzalini (2006) and Arellano-Valle and Genton (2010b).

An important case of the extended form of skew normal distribution was studied by Arnold et al. (1993) and Azzalini and Capitanio (1999), with attention restricted to the case $q = 1$. The general case with an arbitrary q was subsequently studied by Liseo and Loperfido (2003), González-Farás et al. (2004), and Gupta et al. (2004), among others, in different contexts. A summary of these distributions is given in Arellano-Valle and Azzalini (2006).

The equivalent of SUN for the t-distribution, known as the unified skew t (SUT) distribution, is briefly sketched in Arellano-Valle and Genton (2010a) in a study of a restricted version of the SUT distribution.

Besides the flexibility in terms of both skewness and heavy tails, the members (with non-zero extension parameter) of the SUE family have desirable properties such as closure under conditioning and, also, the ability to model lighter tails than the normal distribution as well. It also enjoys parallel stochastic representations and various other properties.

2.4 Generalized multivariate skew distributions

When we further relax the distributional assumption of the latent variable $\boldsymbol{Y}_0$, allowing it to have a different distribution to $\boldsymbol{Y}_1$, we obtain a 'generalized form' of the multivariate skew distribution. Some notable examples of this class include the fundamental skew-elliptical (FUSE) family (Arellano-Valle and Genton, 2005) and the selection (SLCT) distributions (Arellano-Valle et al., 2006). In general, the density of this class takes the form

$$f(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = f_p(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\, Q(w(\boldsymbol{y})), \tag{13}$$

where $Q(.)$ is a distribution function (not necessarily related to f_p), and $w(\boldsymbol{y})$ is some odd function of $\boldsymbol{y}$.

An intuitive example of a generalized form of a skew distribution is the family of skew normal symmetric and skew t-symmetric distributions. In the univariate case, Nadarajah and Kotz (2003) considered a family of distributions, where $f(\boldsymbol{y})$ in (13) is the normal density, but the skewing function $Q(.)$ is taken to be the distribution function of other elliptical (or symmetric) distributions, for example, the t, Cauchy, logistic, Laplace and uniform distribution. The same approach was applied to a t-density in Nadarajah (2008) to construct a skew-t-symmetric family, which includes the skew t-normal, skew t-Cauchy, skew t-Laplace, skew t-logistic and skew t-uniform distribution. One interesting case of this family is the skew t-normal (STN) distribution, which takes the same set of parameters as the MST distribution, but model fitting is more computationally feasible than the MST distribution. The STN distribution, which was recently extended to the multivariate case in Lin et al. (2013), and is discussed further in Section 4.3.

3 Finite mixtures of skew normal and skew t-distributions

In the context of finite mixture models, the population density is assumed to be a convex linear combination of component densities. A p-dimensional random vector $\boldsymbol{Y}$ has a multivariate mixture distribution with g components if its density can be written as

$$f(\boldsymbol{y};\boldsymbol{\Psi}) = \sum_{h=1}^{g} \pi_h f(\boldsymbol{y};\boldsymbol{\theta}_h), \tag{14}$$

where $f(\boldsymbol{y};\boldsymbol{\theta}_h)$ are the parametric component densities, $\boldsymbol{\theta}_h$ is a vector containing the unknown parameters in the hth component density, the π_h are the nonnegative mixing proportions which sum to one, and $\boldsymbol{\Psi}$ is the vector of all unknown parameters in the mixture model.

3.1 Mixtures of normal and t-distributions

One of the first and most frequently used mixture models is the normal mixture model. It has become one of the most popular modelling tools due to its wide applicability and ease of fitting via the EM algorithm (McLachlan and Peel, 2000); see also Ganesalingam and McLachlan (1978) and McLachlan and Basford (1988). As a more robust alternative to the normal mixtures, finite mixture of t-distributions have also been widely applied to a variety of data (McLachlan and Peel, 1998). We shall denote these two models by FM-MN and FM-MT, respectively. Their densities are given by

$$f(\boldsymbol{y};\boldsymbol{\Psi}) = \sum_{h=1}^{g} \pi_h \phi_p(\boldsymbol{y};\boldsymbol{\mu}_h,\boldsymbol{\Sigma}_h), \tag{15}$$

and

$$f(\boldsymbol{y};\boldsymbol{\Psi}) = \sum_{h=1}^{g} \pi_h t_p(\boldsymbol{y};\boldsymbol{\mu}_h,\boldsymbol{\Sigma}_h,\nu_h). \tag{16}$$

Estimation of the model parameters of the FM-MN and FM-MT models are typically carried out by applying the expectation-maximization (EM) algorithm (Dempster et al., 1977); see also McLachlan and Krishnan (2008). Implementations are available in most mathematical and/or statistical software, including R (R Development Team, 2011) and MATLAB.

3.2 Mixtures of skew normal distributions

One of the first attempts at using skew component densities in a mixture model is the finite mixture of restricted skew normal distributions adopted in Pyne et al. (2009a). Three different variants of the restricted MSN mixture model were studied by Pyne et al. (2009a), Cabral et al. (2012), and

Frühwirth-Schnatter and Pyne (2010), the latter from a Bayesian perspective. A discussion of these models and the EM algorithm for fitting them can be found in Lee and McLachlan (2013c). For our purpose, the density of a g-component restricted multivariate skew normal mixture model is given by

$$f(\boldsymbol{y};\boldsymbol{\Psi}) = \sum_{h=1}^{g} \pi_h \, f_{rMSN}(\boldsymbol{y};\boldsymbol{\mu}_h,\boldsymbol{\Sigma}_h,\boldsymbol{\delta}_h), \tag{17}$$

where f_{rMSN} denotes the p-dimensional restricted MSN density (2); that is, we adopt the characterization as used in Pyne et al. (2009a). We refer to (17) as the finite mixture of restricted multivariate skew normal (FM-rMSN) distributions.

Similarly, for the unrestricted case, we shall use the notation FM-uMSN for the finite mixture of unrestricted multivariate skew normal distribution. This model was studied by Lin (2009), who gave an exact implementation of the EM algorithm. The density of the FM-uMSN distribution is given by

$$f(\boldsymbol{y};\boldsymbol{\Psi}) = \sum_{h=1}^{g} \pi_h \, f_{uMSN}(\boldsymbol{y};\boldsymbol{\mu}_h,\boldsymbol{\Sigma}_h,\boldsymbol{\delta}_h), \tag{18}$$

where f_{uMSN} refers to the density (10). To date, mixtures of more general versions of skew normal distributions have not been studied, although the non-mixture or single-component version of some extended skew normal distributions have been considered in several application; for example, in Rodrigues (2006).

3.3 Mixtures of skew t-distributions

Being a natural extension of the skew normal and t-mixture models, the skew t-mixture model has recently received much attention. It is more robust against outliers than the skew normal mixture, and retains reasonable tractability. Several versions of restricted skew normal and skew t-mixtures have been studied by a number of authors, including Pyne et al. (2009a), Frühwirth-Schnatter and Pyne (2010), Basso et al. (2010), and Vrbik and McNicholas (2012). It was shown in Lee and McLachlan (2013c) that they are essentially adopting the same distribution, although using different parameterizations. In the examples to follow, we refer to the characterization adopted in Pyne et al. (2009a) as the restricted skew t (FM-rMST) model, the density of which is given by

$$f(\boldsymbol{y};\boldsymbol{\Psi}) = \sum_{h=1}^{g} \pi_h \, f_{rMST}(\boldsymbol{y};\boldsymbol{\mu}_h,\boldsymbol{\Sigma}_h,\boldsymbol{\delta}_h,\nu_h), \tag{19}$$

where $f_{rMST}(.)$ refers to (5).

The unrestricted case has received much less attention, partly due to the density involving a p-dimensional skewing function, making it difficult to handle analytically. The unrestricted skew t-mixture, hereafter FM-uMST, has density

$$f(\boldsymbol{y};\boldsymbol{\Psi}) = \sum_{h=1}^{g} \pi_h \; f_{uMST}(\boldsymbol{y};\boldsymbol{\mu}_h,\boldsymbol{\Sigma}_h,\boldsymbol{\delta}_h,\nu_h), \tag{20}$$

where f_{uMST} is the unrestricted MST density given by (11). This mixture model was first studied by Lin (2010), and subsequently by Lee and McLachlan (2011, 2013a).

3.4 Software for fitting mixtures of skew distributions

Algorithms for fitting the restricted and unrestricted multivariate skew normal and skew t-distributions are given in a number of contributions mentioned above. Explicit expressions for the implementation of the EM algorithm for the FM-rMSN model can be found in Pyne et al. (2009a), Frühwirth-Schnatter and Pyne (2010), and Cabral et al. (2012), and software implementation in R for the versions considered in Pyne et al. (2009a) and Cabral et al. (2012) are available publicly from `EMMIX-skew` (Wang et al., 2009) and `mixsmsn` (Prates et al., 2011), respectively. The EM algorithm for fitting their corresponding FM-rMST distributions are also implemented in their packages. An alternative implementation of the FM-rMST model is presented in Vrbik and McNicholas (2012), where the conditional expectations involved in the E-step are expressed in terms of hypergeometric functions. For a discussion of the connections between these implementations, the reader is referred to Lee and McLachlan (2013c). For the unrestricted case, software implementation of the FM-uMSN and FM-uMST models is given in the R package `EMMIX-uskew` (Lee and McLachlan, 2013b).

4 Other non-normal mixture models

While the skew symmetric distributions plays a central role in the development of non-normal models, there are other alternative asymmetric models that have received some attention, in particular, the normal-inverse-Gaussian distribution and the recently proposed shifted asymmetric Laplace distribution.

4.1 Multivariate normal-inverse-Gaussian distribution

The multivariate normal inverse Gaussian (MNIG) distribution belongs to the generalized hyperbolic (GH) class (Barndorff-Nielsen, 1977). It can be obtained as a mean-variance mixture of normal distributions with an inverse Gaussian mixing distribution. As a parametric family with five parameters, the MNIG

distribution can take a range of shapes with arbitrary degrees of skewness and heaviness of tails. Let $\boldsymbol{Y}$ be a p-dimensional random vector. Then $\boldsymbol{Y}$ has a MNIG distribution if, conditional on a scalar inverse Gaussian variable W, $\boldsymbol{Y}$ has a multivariate normal distribution. More formally, we can write,

$$\boldsymbol{Y} \mid w \sim N_p(\boldsymbol{\mu} + w\boldsymbol{\lambda}\boldsymbol{\Sigma}, w\boldsymbol{\Sigma}),$$

$$W \sim IG(\xi, \sigma),$$

where $IG(\xi, \sigma)$ denotes the inverse Gaussian distribution with parameters ξ and σ, and $\boldsymbol{\lambda}$ is $p \times 1$ skewness parameter. It follows that the density of the MNIG distribution is given by

$$f_{MNIG}(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \xi, \sigma) = 2^{-(\frac{p-1}{2})}\sigma \left(\frac{\sqrt{\xi^2 + \boldsymbol{\lambda}^T\boldsymbol{\Sigma}\boldsymbol{\lambda}}}{\pi\sqrt{\sigma^2 + d(\boldsymbol{y})}} \right)^{\frac{p+1}{2}} e^{\xi\sigma + \boldsymbol{\lambda}^T(\boldsymbol{y}-\boldsymbol{\mu})} K_{\frac{p+1}{2}} \left(\sqrt{(\xi^2 + \boldsymbol{\lambda}^T\boldsymbol{\Sigma}\boldsymbol{\lambda})(\sigma^2 + d(\boldsymbol{y}))} \right), \quad (21)$$

where $d(\boldsymbol{y})$ is defined as in (5), and $K_r(.)$ denotes the modified Bessel function of the third kind of order r. The parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ have their usual meanings, ξ is a scalar parameter that affects the tails of the distribution, and σ affects the scale of the distribution. Note that when ξ and σ tend to infinity, the MNIG distribution approaches the multivariate normal distribution.

The finite mixture of MNIG distributions, studied by Karlis and Santourian (2009), has density given by

$$f(\boldsymbol{y}; \boldsymbol{\Psi}) = \sum_{h=1}^{g} \pi_h f_{MNIG}(\boldsymbol{y}; \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h, \boldsymbol{\lambda}_h, \xi_h, \sigma_h), \quad (22)$$

where the π_h are the mixing proportions as defined previously, and $\boldsymbol{\Psi}$ contains all the unknown parameters of the model. The ML estimates of the parameters of the model (22) can be obtained via the EM algorithm, with closed-form E- and M-steps involving modified Bessel functions.

4.2 Multivariate shifted asymmetric Laplace distribution

Mixtures of shifted asymmetric Laplace (SAL) distributions were recently introduced as another alternative to mixtures of skew elliptical distributions. The SAL distribution is a generalization of the Laplace distribution, and a limiting case of the generalized hyperbolic (GH) distribution (Kotz et al., 2001). Its density is given by

$$f_{MSAL}(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}) = \frac{2e^{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \left(\frac{d(\boldsymbol{y})}{2 + \boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}} \right)^{\frac{2-p}{2}} K_{\frac{2-p}{2}} \left(\sqrt{(2 + \alpha^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha})d(\boldsymbol{y})} \right), \quad (23)$$

where $\boldsymbol{\mu}$ is p-dimensional location vector, $\boldsymbol{\Sigma}$ is a $p \times p$ covariance matrix, and $\boldsymbol{\alpha}$ is a p-dimensional vector controlling the skewness of the MSAL distribution (Franczak et al., 2012). A finite mixture of MSAL (FM-MSAL) distributions has density

$$f(\boldsymbol{y};\boldsymbol{\Psi}) = \sum_{h=1}^{g} \pi_h f_{MSAL}(\boldsymbol{y};\boldsymbol{\mu}_h,\boldsymbol{\Sigma}_h,\boldsymbol{\alpha}_h), \tag{24}$$

where $\boldsymbol{\Psi}$ is the vector containing all the unknown parameters of the model. One advantage of employing the MSAL density as component distributions of the finite mixture model is that ML estimation of the parameters of (23) can be obtained in a straightforward manner via the EM algorithm, involving relatively simple expressions for the E- and M-steps.

4.3 Mixtures of skew t-normal distributions

The fitting of an (unrestricted) multivariate t-mixture model can become quite slow when p is large, due to the computation time involved in calculating the multivariate t-distribution function. In view of this, Lin et al. (2013) proposed a computational more feasible alternative, the (restricted) multivariate skew t-normal (rMSTN) distribution, where the skewing function in (11) is replaced by a (univariate) normal distribution function. Its density is given by

$$f_{rMSTN}(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda},\nu) = 2t_p(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)\Phi_1(\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{y}-\boldsymbol{\mu})), \tag{25}$$

and the corresponding mixture model is given by

$$f(\boldsymbol{y};\boldsymbol{\Psi}) = \sum_{h=1}^{g} \pi_h f_{MSTN}(\boldsymbol{y};\boldsymbol{\mu}_h,\boldsymbol{\Sigma}_h,\boldsymbol{\lambda}_h,\nu_h), \tag{26}$$

where $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and ν have the same meaning as in the MST distribution, and $\boldsymbol{\lambda}$ is p-dimensional parameter controlling the skewness of the distribution. With this formulation, considerable time is saved in the implementation of the E-step of the EM algorithm for fitting mixtures of rMSTN distributions. Similar to the EM algorithm for fitting FM-rMST distributions, closed-form expressions can be obtained for the E- and M-step, involving evaluation of distribution functions in one-dimension only.

The rMSTN distribution shares the same set of parameters as the rMST and uMST distributions (note that $\boldsymbol{\lambda}$ can be expressed in terms of $\boldsymbol{\delta}$), and was shown in Lin et al. (2013) to have competitive performance to the MST mixture model.

4.4 Mixtures of generalized hyperbolic distributions

The generalized hyperbolic (GH) distribution has become a popular model for describing financial data. It has also been used widely in modeling mass-size particle data; see, for example, Jones and McLachlan (1989) and the references therein. This family of GH distributions was originally introduced by Barndorff-Nielsen (1977) as a normal variance-mean mixture distribution with a mixing distribution given by a generalized inverse Gaussian (GIG) distribution. The GH family encompasses a number of non-normal distributions as special or limiting cases, including the aforementioned MING and MSAL distributions, the hyperbolic distribution, and another variant of the skew t-distribution known as the generalized hyperbolic skew t (GHST) distribution. Note that the GHST distribution is not equivalent to the rMST distribution, although it can be considered as a restricted form of skew distribution in the sense that the latent skewing variable is univariate.

The traditional characterization of the six-parameters multivariate generalized hyperbolic distribution (McNeil et al., 2005) suffers from an identifiability issue. To work around this, Browne and McNicholas (2013) considered an alternative parameterization by setting the scale parameter to a fixed value, resulting in a four-parameter density given by

$$f(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\alpha},\lambda,\omega) = \frac{e^{-\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\left(\frac{\omega+d(\boldsymbol{y})}{\omega+\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}\right)^{\left(\frac{\lambda}{2}-\frac{p}{4}\right)} \frac{K_{\lambda-\frac{p}{2}}\left(\sqrt{(\omega+\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha})\,(\omega+d(\boldsymbol{y}))}\right)}{K_\lambda(\omega)}, \tag{27}$$

where $d(\boldsymbol{y}) = (\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$. We shall refer to this by the multivariate generalized hyperbolic (MGH) distribution. The corresponding mixture density is given by

$$f(\boldsymbol{y};\boldsymbol{\Psi}) = \sum_{h=1}^{g}\pi_h f_{MGH}(\boldsymbol{y};\boldsymbol{\mu}_h,\boldsymbol{\Sigma}_h,\boldsymbol{\alpha}_h,\omega_h,\lambda_h). \tag{28}$$

The EM algorithm for the FM-MGH model presented in Browne and McNicholas (2013) does not exist in closed form.

4.5 Hierarchical mixtures of multivariate t-distributions

In a hierarchical mixture distribution, or mixture of mixture distributions, each component density is modelled by a mixture model, thus allowing it to capture asymmetric shapes; see, for example, the hierarchical mixtures of experts (HME) proposed by Jordan and Jacobs (1992) and the two-level model

by Calò et al. (2013). The density of a hierarchical (two-level) mixture of t (HMT) distributions (Nguyen and Wu, 2013) takes the form

$$\begin{aligned} f(\boldsymbol{y};\boldsymbol{\Psi}) &= \sum_{h=1}^{g} \pi_h f_{FM-MT}(\boldsymbol{y};\boldsymbol{\Psi}_h), \\ &= \sum_{h-1}^{g} \pi_h \sum_{i=1}^{g_i} \pi_{hi}\, t_p(\boldsymbol{y};\boldsymbol{\mu}_{hi},\boldsymbol{\Sigma}_{hi},\nu_{hi}), \end{aligned} \tag{29}$$

where π_h is the mixing proportion of the higher-level mixture model, and π_{hi} is the conditional probability of an observation $\boldsymbol{y}$ belonging to the ith component of the lower-level mixture model given that it belongs to the hth higher-level component. In (29), each higher-level component density is modelled by multiple multivariate t-distributions. Note that the number of lower-level components g_i can vary across the higher-level components, but for the example in Section 5.6, we fix $g_i = 2$ as used in Nguyen and Wu (2013). It should be also noted that fitting a FM-HMT is intrinsically different to fitting a traditional t-mixture model with $g^* > g$ components and then performing some merging procedure. The hierarchical structure of the FM-HMT model automatically registers each observation to a higher-level component.

For ease of reference, we include a summary of the above-mentioned non-normal distributions in Table 1.

5 Applications

5.1 Clustering flow cytometric data

We consider the clustering of a quadvariate dataset derived from a hematopoietic stem cell transplant (HSCT) experiment, collected by the British Columbia Cancer Agency. These data contain close to 10,000 samples, each stained with four fluorescent markers. Pairwise plots of the markers for this dataset are displayed in the left lower panels of Figure 1, where cells are displayed in four different colours according to manual expert clustering.

We compared the performance of four multivariate mixture models, namely, FM-uMST, FM-rMST, FM-MNIG, and FM-MSAL in assigning cells to the expert's clusters. Manual gating identified four clusters in this sample case. We therefore applied the algorithm for fitting each model with the number of components g predefined as 4. For the skew symmetric mixture models, the algorithms were initialized with a number of different initial labels given by k-means clustering, and the set of parameters associated with the highest relative (initial) log-likelihood is selected to start the EM iterations. For the FM-MSAL and FM-MNIG models, the initialization strategy described by their authors were used. The algorithms were terminated according to the stopping criteria described by their authors; see Table 1 for references.

To assess the performance of these three algorithms, we calculated the rate of misclassification against the benchmark expert clustering, which is taking as being the 'true' class membership. This is measured by choosing among the possible permutations of the cluster labels the one that gives the lowest value. A lower misclassification rate indicates a closer match between the 'true' labels and the cluster labels given by the candidate algorithm. We also report here another popular measure of clustering agreement between two clusters, namely, the adjusted rank index (ARI) (Hubert and Arabie, 1985). An ARI equal to 1 corresponds to an exact match between the two set of labels, while an ARI value of 0 indicates non-agreement between each pair of points. Note that in the calculation of misclassification rate and the ARI, dead cells were removed before computing these measures.

Mixture distribution	Component density	EM algorithms
Restricted Skew-normal (FM-rMSN)	$f(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\delta}) = 2\phi_p(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma})\,\Phi_1(\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu});0,\lambda)$ $\lambda = 1-\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$	Pyne et al. (2009a)
Unrestricted skew-normal (FM-uMSN)	$f(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\delta}) = 2^p\phi_p(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma})\,\Phi_p(\boldsymbol{\Delta}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu});\mathbf{0},\Lambda)$ $\boldsymbol{\Delta} = \text{diag}(\boldsymbol{\delta}),\ \Lambda = \boldsymbol{I}_p - \boldsymbol{\Delta}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}$	Lin (2009)
Restricted skew t (FM-rMST)	$f(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\delta},\nu) = 2t_p(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)\,T_1\left(\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\sqrt{\frac{\nu+p}{\nu+d(\boldsymbol{y})}};0,\lambda,\nu+p\right)$ $d(\boldsymbol{y}) = (\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}),\ \lambda = 1-\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$	Pyne et al. (2009a) Vrbik and McNicholas (2012)
Unrestricted skew t (FM-uMST)	$f(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\delta},\nu) = 2^p t_p(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)\,T_p\left(\boldsymbol{\Delta}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\sqrt{\frac{\nu+p}{\nu+d(\boldsymbol{y})}};\mathbf{0},\Lambda,\nu+p\right)$ $\boldsymbol{\Delta} = \text{diag}(\boldsymbol{\delta}),\ d(\boldsymbol{y}) = (\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}),\ \Lambda = \boldsymbol{I}_p - \boldsymbol{\Delta}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}$	Lin (2010) Lee and McLachlan (2013a)
Restricted skew t-normal (FM-rMSTN)	$f(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda},\nu) = 2t_p(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)\,\Phi_1(\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{y}-\boldsymbol{\mu});0,1)$	Lin et al. (2013)
Normal-inverse-Gaussian (FM-MNIG)	$f(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda},\xi,\sigma) = 2^{-(\frac{p-1}{2})}\sigma\left(\frac{\alpha}{\pi\sqrt{\beta}}\right)^{\frac{p+1}{2}} e^{\xi\sigma+\boldsymbol{\lambda}^T(\boldsymbol{y}-\boldsymbol{\mu})}K_{\frac{p+1}{2}}(\alpha\beta)$ $\alpha^2 = \xi^2+\boldsymbol{\lambda}^T\boldsymbol{\Sigma}\boldsymbol{\lambda},\ \beta^2 = \sigma^2+(\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$	Karlis and Santourian (2009)
Shifted asymmetric Laplace (FM-MSAL)	$f(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\alpha}) = \frac{2e^{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}}{(2\pi)^{\frac{p}{2}}\lvert\boldsymbol{\Sigma}\rvert^{\frac{1}{2}}}\left(\frac{d(\boldsymbol{y})}{\beta}\right)^{\frac{2-p}{2}}K_{\frac{2-p}{2}}\left(\sqrt{\beta\,d(\boldsymbol{y})}\right)$ $\beta = 2+\alpha^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha},\ d(\boldsymbol{y}) = (\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$	Franczak et al. (2012)
Generalized hyperbolic (FM-MGH)	$f(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\alpha},\lambda,\omega) = \frac{e^{-\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}}{(2\pi)^{\frac{p}{2}}\lvert\boldsymbol{\Sigma}\rvert^{\frac{1}{2}}}\left(\frac{\omega+d(\boldsymbol{y})}{\gamma}\right)^{\left(\frac{\lambda}{2}-\frac{p}{4}\right)}\frac{K_{\lambda-\frac{p}{2}}\left(\sqrt{\gamma(\omega+d(\boldsymbol{y}))}\right)}{K_\lambda(\omega)}$ $\gamma = \omega+\alpha^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha},\ d(\boldsymbol{y}) = (\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$	Browne and McNicholas (2013)

Table 1 Summary of the non-normal mixture distributions in Section 3 and 4.

Model	FM-uMST	FM-rMST	FM-MNIG	FM-MSAL
misclassification rate	0.0034	0.0042	0.3204	0.0052
ARI	0.9781	0.9782	0.5469	0.9751

Table 2 Clustering performance of various multivariate mixture models on the HSCT dataset. Cells identified as dead cells were not included in the calculation of the error rate.

Algorithm	flowKoh	flowClust	flowMeans	FLOCK	ADICyt
misclassification rate	0.2218	0.0445	0.0086	0.0138	0.0111
ARI	0.7177	0.9581	0.9695	0.9549	0.9592

Table 3 Clustering performance of various other automated gating algorithms on the HSCT dataset. Cells identified as dead cells were not included in the calculation of the error rate.

The summary of clustering results of the algorithms are listed in Table 2. The upper right panels of Figure 1 show the classification results of FM-uMST, with different colours indicating different clusters. The 'true' clustering is shown in the lower left panels of Figure 1. We observe from Table 2 that the skew t-mixture models achieved the lowest misclassification rate and the highest ARI, indicating a close match to the true clustering. The FM-rMST and FM-MSAL models also gave reasonable clustering results, achieving only slightly higher misclassification rate and lower ARI than the FM-uMST model.

For comparison, we applied various automated gating algorithms to the HSCT dataset, including flowkoh (Nikolic, 2010), flowClust (Lo et al., 2009, 2008), flowMeans (Pyne et al., 2009b), FLOCK (Qian et al., 2010), and ADICyt (Aghaeepour et al., 2013). The procedure flowKoh, developed by the British Columbia Institute of technology, employs the self-organizing map (SOM) to cluster and visualize high-dimensional data. The procedure flowClust performs the fitting of multivariate t-mixture model after Box-Cox transformation. The procedure flowMeans adopts the traditional k-means algorithm to perform clustering, and applies a change-point detection algorithm to determine the number of populations. The procedure FLOCK, short for Flow Clustering without K, uses a grid-based partitioning and merging scheme for the identification of cell clusters, and determines the number of clusters by examing the density gap between the partitioned data regions. The last procedure considered, ADICyt, is a commercial software designed for fast and effective analysis of flow cytometric data. The algorithm mimics manual gating by hierarchically splitting the data in a sequence of optimally selected 2D-projections, then applies an entropy-based merging scheme to obtain a clustering of the data. The misclassification rate and ARI values are listed in Table 3. It can be observed that mixtures of skew-elliptical distributions achieve the highest ARI and the lowest misclassification error. A plot of the fitted contours of the multivariate skew mixture models also show that these models can capture the shape of the data very well.

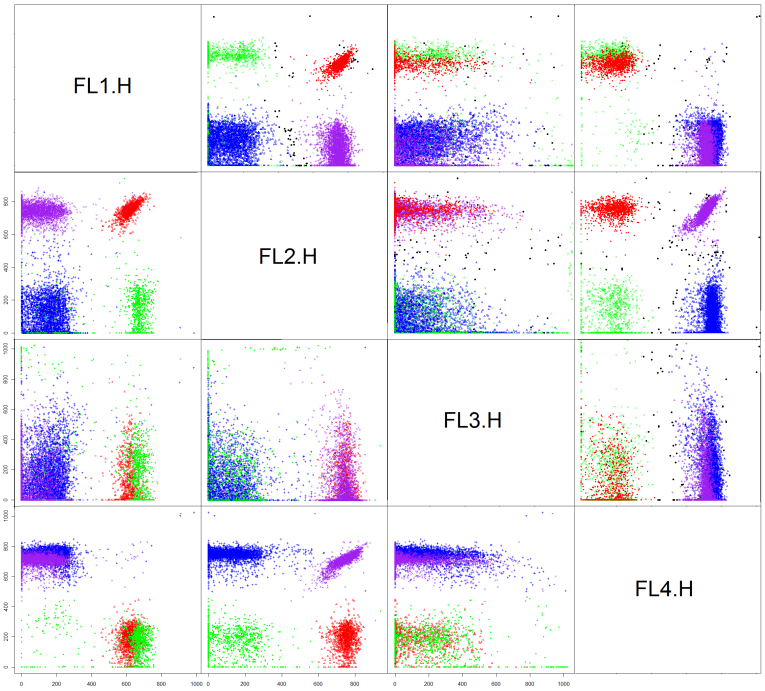


Fig. 1 HSCT dataset: Scatter plot of Fl1.H to FL4.H in four colours, corresponding to four clusters. Lower left panels: clustering given by FM-uMST; Upper right panels: clustering given by manual experts.

5.2 Australian Institute of Sports data

Our second example on real data concerns the Australian Institute of Sport (AIS) data, a dataset analyzed by Cook and Weisberg (1994) that has since been used extensively in the literature. These data contain 11 biomedical measurements on 202 Australian athletes (100 female and 102 male), of which we shall consider three variables here, namely, the body mass index (BMI), lean body mass (LBM), and the percentage of body fat (Bfat).

We fitted two-component mixtures of MN, MT, rMSN, rMSTN, rMST, uMST, MSAL, MNIG, MGH distributions to the data. The initial values were generated according to the strategy described by their authors. Table 4 lists the number of misclassified units against the true clustering (male and female) for each model. Figure 2 shows the contours of the fitted density of each model in two of the variables, LBM and Bfat. It can be observed from Table 4 that both the FM-uMST and the FM-MNIG models have misallocated four samples, the least number of misallocations in this case, with the other models yielding around double this number. However, the fitted contours depicted in Figure 2 reveal that the fitted contours of the FM-uMST model would appear to be more reasonable than the density estimated by the FM-MNIG

Model	number of misclassified units
FM-MN	8
FM-MT	9
FM-rMSN	9
FM-rMSTN	7
FM-rMST	7
FM-uMST	4
FM-MSAL	13
FM-MNIG	4
FM-MGH	8

Table 4 Clustering performance of various multivariate skew mixture models on the AIS dataset.

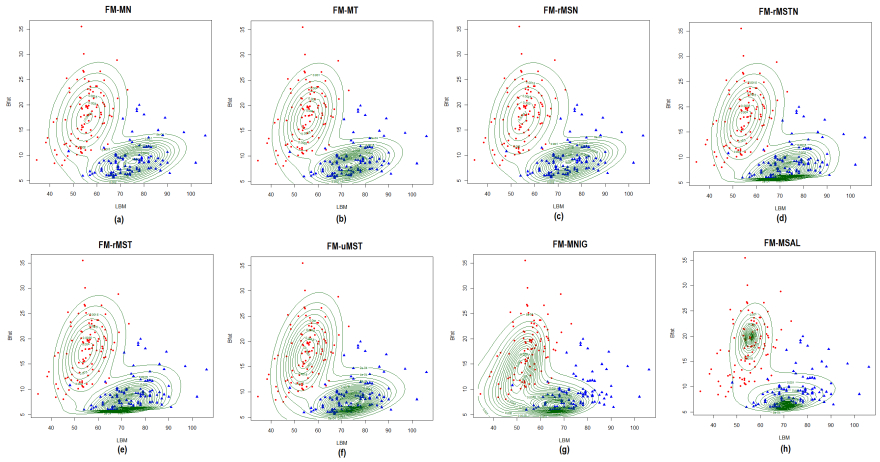


Fig. 2 AIS dataset: Contour plots of the fitted two-component mixture models on the trivariate data. Scatter plot of LBM and Bfat is given in two colours, red dots for male and blue triangles for female; (a) the fitted mixture contour of FM-MN; (b) contour plot of the fitted FM-MT model; (c) the density contours of the fitted FM-rMSN model; (d) the contours of the densities of the fitted FM-rMSTN model; (e) the fitted contours of the FM-rMST model; (f) the density contours of the fitted FM-uMST model; (g) the contours of the densities of the fitted FM-MNIG model (h) contour plot of the fitted FM-MSAL model.

model, although both achieve the same misclassification rate. The restricted asymmetric distributions (FM-rMST, FM-rMSTN, FM-MGH) yields similar results, with 7 or 8 misclassified observations.

5.3 Bankruptcy data

We consider a subset of the bankruptcy data studied by Altman in 1968 (Altman, 1968). The original sample consists of annual financial data of 66 American firms recorded in the form of ratios. Half of the selected firms had filed for bankruptcy. The data were collected approximately two years prior to their bankruptcy, and the other 33 samples were randomly chosen from those that were financially sound. Altman selected five financial variables from a list of 22

Model	number of misclassified firms	BIC
FM-rMSN	16	1340.51
FM-uMSN	13	1352.47
FM-rMST	14	1335.00
FM-uMST	2	1336.08

Table 5 Clustering performance of various restricted and unrestricted multivariate mixture models on the bankruptcy dataset.

ratios that are potentially significant in predicting bankruptcy. For this illustration, we consider two of these ratios, namely, the ratio of retained earnings (RE) to total assets, and the ratio of earnings before interests and taxes (EBIT) to total assets. The goal here is to predict whether a firm went bankrupt based on the two variables. This bivariate sample (Figure 3a) is apparently bimodal and appears to be asymmetric; hence we fit a two-component skew mixture model to the data. To compare the performance of restricted and unrestricted skew mixture models, we fitted finite mixtures of restricted multivariate skew normal (FM-rMSN) distributions, finite mixtures of (restricted) multivariate skew normal (FM-uMSN) distributions, finite mixture of restricted multivariate skew t (FM-rMST) model, and finite mixtures of unrestricted multivariate t (FM-uMST) distributions.

The results for the multivariate mixture models are presented in Table 5. Also reported are the BIC values of the fitted models. The contours of the fitted mixture densities FM-rMSN, FM-uMSN, FM-rMST, and FM-uMST are depicted in Figures 3b to 3f, respectively. To better demonstrate the shape of the fitted models, the estimated densities of each component are displayed rather than the mixture contours. It can be observed that heavy-tailed models have a lower number of misclassified firms compared to their respective skew normal mixture models. It can also be observed from Table 5 that improved clustering can be achieved with the unrestricted model. This is also evident in Figure 3 where they adapt much more closely the shape of the clusters, especially the cluster corresponding to the solvent firms (blue cluster) for which the restricted skew mixture distributions find difficult to model. It is interesting to observe that the unrestricted skew t-mixture model performs much better than the other three models for this case, with the number of incorrectly classified firms is lowered to two. We report that the other models with non-normal components did not achieve a comparable result to the FM-uMST model in this example (having 24 or more misclassified firms), the results of which are omitted here.

5.4 Estimation of Value-at-Risk

The Value-at-risk (VaR) is frequently used in financial risk management as a measure of the risks of investment loss. It is given by the predicted maximum loss over a specified holding period given a specified confidence level. More formally, consider a portfolio of p assets returns $Y_1, \ldots, Y_p$ and let $Y_R =$

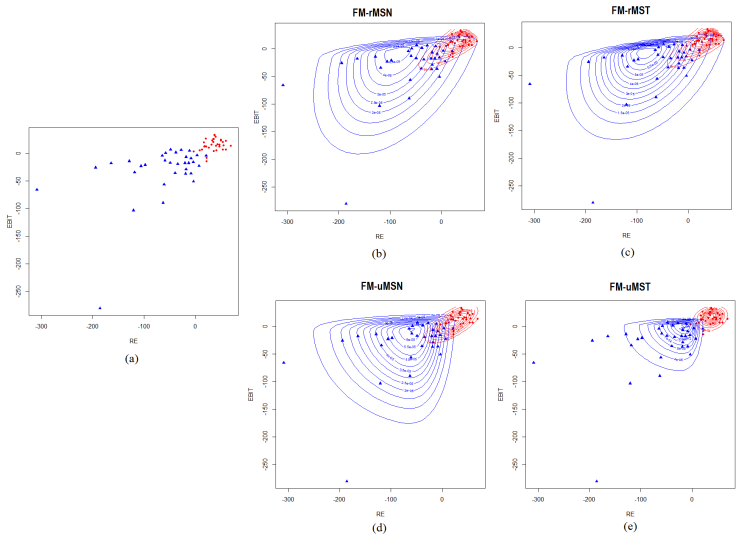


Fig. 3 Bankruptcy dataset: Contour plots of the fitted two-component mixture models on RE and EBIT. (a) Scatter plot of RE and EBIT in two colours, red dots for bankrupted firms and blue triangles for solvent firms; (b) the fitted component contours of the FM-MN model; (c) the contours of the component densities of the fitted FM-MT model; (d) the component density contours of the fitted FM-MSN model; (e) contour plot of the component densities of the fitted FM-uMST model.

$\sum_{j=1}^{p} Y_j$ be the (aggregate) return. Then the VaR is defined to be the negative of the largest value of y_α satisfying

$$\text{pr}\{Y_R < y_\alpha\} \leq \alpha, \tag{30}$$

where α is the significance level. Thus the VaR can be interpreted as the negative of the quantile of order α of the portfolio's (hypothetical) return distribution F_R; that is, VaR can be expressed as

$$-F_R^{-1}(\alpha). \tag{31}$$

Note that the negative sign preceding $F_R^{-1}(\alpha)$ ensures the VaR is a positive value, that is, a positive amount of 'losses'. For example, if α is 1% and the time period is one day, then a VaR of one million dollars can be interpreted as meaning that the probability of incurring a loss in excess of one million dollars for this portfoilio over this day is bounded by 0.01.

Current analytical calculations of VaR typically assume the distribution of the portfolio return to be a normal or log-normal distribution, which is rarely true in reality. It is well known that the historical return data exhibits heavy tails and skewness, which results in underestimation and overestimation of the VaR at high and low confidence levels, respectively, when a normal distribution is assumed. It is thus natural to speculate that fitting mixtures of skew distributions can potentially improve the accuracy of VaR estimation,

as shown by the promising results in an example given by Soltyk and Gupta (2011).

In this example, we consider a portfolio of three shares listed on the Australian Stock exchange (ASX). The data contain monthly returns for the shares Woolworths Limited (WOW), Woodside Petroleum Limited (WPL), and Westfield Group (WDC) for the period 1st January 2000 to 23rd November 2012. The return of each share is based on the adjusted closing price, and results are recorded as a percentage. The summary statistics of the stock returns suggest that the data do not satisfy normality assumption, and skewness and excess kurtosis are present in all three returns.

We fitted mixtures of normal, skew normal, skew t, skew t-normal, and shifted asymmetric Laplace distributions to this trivariate dataset. For this illustration, the mixture models are fitted with $g = 1$ to 4 components, and the model with the lowest BIC is selected. To estimate the VaR based on the fitted models, the simulation approach is used (see Soltyk and Gupta (2011) for an example using the FM-uMSN model). In brief, a large sample is generated from the fitted model, and an estimate of the VaR is given by the appropriate quantile of the total assets return of the simulated sample. The first row of Table 7 gives the estimated 1% VaR value given by the five models. Given that the empirical or historical VaR calculated from the data is \$26.92, it can be observed that the FM-uMST model gives the closest estimate. Both the normal and skew normal mixture models underestimated the VaR, while the other models overestimated it. Table 7 supports the results in Soltyk and Gupta (2011), which presented an example to show that the incorporation of a skewness parameter can provide a more accurate fit to stock returns. We observe here that further improvements can be made by fitting mixtures with asymmetric heavy-tailed component distributions.

There are various techniques proposed for evaluating the accuracy of a VaR measure. We shall consider two of the most common statistical tests available, namely the unconditional coverage test (or backtesting) and the (Markov) independence test. The backtest developed by Kupiec (1995) is one of the earliest and most widely used tests for VaR, focusing on the unconditional coverage property of the VaR estimate model. This test is concerned with whether or not the observed violation rate is statistically equal to the expected rate. Specifically, the test examines the proportion of violations, given by $r = v/n$, where v denotes the number of violations in the data and n is the number of observations, and determine whether it is statistically different from the expected rate of $\alpha \times 100\%$ of the sample. A violation occurs if the actual loss exceeds the VAR as implied by the model for the given level of significance. Under the null hypothesis that the model is adequate, v follows a binomial distribution, leading to a likelihood ratio (LR) test statistic of the form

$$\mathrm{LR}_{bt} = 2\log\left[\left(\frac{1-r}{1-\alpha}\right)^{n-v}\left(\frac{r}{\alpha}\right)^{v}\right], \tag{32}$$

which has a chi-squared distribution with one degree of freedom. Hence the test would reject a VaR model if it generates too many or too few violations. The backtest, however, cannot detect whether a VaR model satisfies the independence property.

In view of this, Christoffersen (1998) proposed a more elaborate test, which examines whether or not a VaR violation process is serially dependent. The underlying assumption is that, for a good and accurate VaR model, the probability of observing a VaR violation should not depend on whether or not a violation has occurred the previous observation, that is, the sequence of violations should be independently distributed. The LR test statistic for the Independence test can be expressed as

$$\mathrm{LR}_{idp} = -2\log\left[\frac{(1-q)^{N_1+N_2}q^{N_3+N_4}}{(1-q_1)^{N_1}q_1^{N_3}(1-q_2^{N_2})q_2^{N_4}}\right], \tag{33}$$

where N_1 denotes the number of observations with no violation followed by no violation, N_2 denotes the number of observations with violation followed by no violation, N_3 denotes the number of observations with no violation followed by a violation, and N_4 denotes the number of observations in which a violation has occurred followed by another violation. The proportions q, q_1, and q_2 are defined, respectively, as the proportion of observations in which a violation has occurred, the proportion of observing a violation given no violation has occurred in the previous observation, and the proportion of observing a violation given a violation had occurred in the previous observation. More formally, they are given by $q = \frac{N_3+N_4}{N_1+N_2+N_3+N_4}$, $q_1 = \frac{N_3}{N_1+N_3}$, and $q_2 = \frac{N_4}{N_2+N_4}$. The null hypothesis of $q_1 = q_2$ is tested against the alternative of first-order Markov independence, and the LR statistic (33) again has a χ^2_1 distribution.

The backtest and the independence test can be combined to give a joint test of the adequacy of the VaR forecast in terms of the unconditional coverage and independence properties. The relevant test statistic is simply the sum of the two test statistics (32) and (33) which, under the null hypothesis, has a chi-squared distribution with two degrees of freedom.

In addition, we consider the exceeding ratio (ER) (Choi and Min, 2011), defined as the ratio of the estimated number of violations over the expected number of violations. An ER value greater than one indicates the model is under-forecasting the VaR, and an ER value less than one indicates an over-forecast VaR estimate. The results from Table 7 suggest that all the models considered in this example did not fail the backtesting and independence test, but the FM-uMST, FM-MSTN and FM-MSAL models predict the risk more accurately. The exceeding ratio ranks the FM-uMST model best, followed by the FM-rMSTN and FM-MSAL models which performed equally well.

5.5 Seeds data

We now consider a discriminant analysis application. The seeds data contains seven geometric measurements taken on X-ray images of 210 wheat kernels

	WOW	**WPL**	**WDC**
minimum	-16.58	-22.70	-14.63
maximum	12.22	22.29	19.91
mean	0.97	0.92	0.68
std. dev.	5.37	7.65	5.50
skewness	-0.49	-0.29	0.25
kurtosis	3.04	3.90	3.94

Table 6 Summary statistics of the monthly returns of three Australian stocks for year 2000 to 2012.

	FM-MN	FM-rMSN	FM-uMST	FM-MSTN	FM-MSAL	FM-MGH
VaR	26.08	26.00	27.10	27.78	28.91	29.96
exceeding ratio	1.93	1.93	1.29	0.65	0.65	0.65
backtesting	0.30	0.30	0.72	0.63	0.63	0.63
independence	1.00	1.00	1.00	0.96	0.96	0.96

Table 7 Performance of various skew mixture models on estimating the 1% VaR of three Australian stocks. The backtesting and independence values refers to the p-value of the respective tests. The empirical VaR is \$26.92.

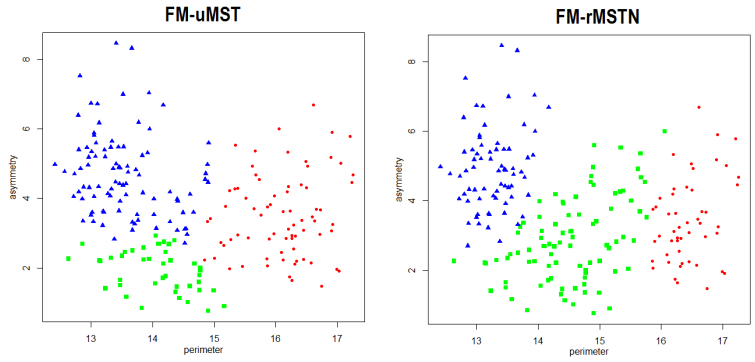
Model	MCR	ARI
FM-MN	0.1904	0.5536
FM-rMSN	0.2222	0.5103
FM-rMST	0.1904	0.5488
FM-rMSTN	0.1746	0.6005
FM-uMST	0.1587	0.6328
FM-MNIG	0.1904	0.5488
FM-MSAL	0.2381	0.5371
FM-HMT	0.2698	0.4493
FM-MGH	0.1904	0.5488

Table 8 Comparison of various mixture models on the classification of the seeds data.

(Charytanowicz et al., 2010). Each of the seeds belongs to one of three different varieties of wheat, namely, Kama, Rosa, and Canadian. For this analysis, we perform model-based classification of the data with $g = 3$ based on the two measurements perimeter and asymmetry. The algorithms were trained on 147 samples (70%) randomly selected from the dataset. The classification performance (MCR and ARI) of the FM-MN, FM-rMSN, FM-rMST, FM-rMSTN, FM-uMST, FM-MNIG, FM-MSAL, FM-MHT, and FM-MGH models are reported in Table 8. The cross-tabulations of the true and predicted classifications for the best two model (FM-uMST and FM-rMSTN) and their scatter plots are given in Figure 4. These results indicates that unrestricted mixture model outperforms the restricted model and other models in this example.

5.6 Image Segmentation

In this example, we consider the segmentation of real-world natural images from the Berkeley's image segmentation dataset (Meignen and Meignen, 2006). The original image is shown in Figure 5(a), and the hand segmented image



	1	2	3
Kama	9	10	1
Rosa	0	0	21
Canadian	22	0	0

	1	2	3
Kama	18	2	0
Rosa	9	0	12
Canadian	0	22	0

Fig. 4 Classification results of the best two models on the seeds data. Upper panels: scatter plot of the clustering results, where the colors represents the predicted class. Lower panels: cross-tabulation of true and predicted class memberships.

by human is given in Figure 5(e) which is taken as the ground truth. For this example, the objective is to segment each pixel of the image into two labels – 'background' and 'foreground'. A visual comparison of the segmentation results of the FM-MN, FM-rMSN, FM-rMST, FM-HMT, FM-MNIG and FM-MGH models (shown in Fig 5(b)-(d), (f)-(h) respectively) shows that the FM-rMST and FM-MGH models are quite good. Observe that this two models produced relatively sharp edges between the 'background' and 'foreground', and the 'noise' around each corner of the image had a lower impact on the segmentation result. To provide a quantitative comparison of the segmentation result given by the various mixture model, we report the miscalssification rate (MCR) (Zhang et al., 2001), Rand Index (RI) (Rand, 1971), Variation of Information (VI) (Meilă, 2005), and Global Consistency Error (GCE) (Martin et al., 2001) for each algorithm in Table 9. Note that for MCR, VI and GCE, a lower value indicates a closer match between the ground truth and the segmented image by an algorithm. For RI, a higher value is preferred. As can be observed in table 9, the result of the FM-rMST model has the highest RI and lowest MCR, VI and GCE. The segmentation accuracy of the FM-MGH model is quite close to that given by the FM-rMST model. A closer inspection of the 'branches' area on the far right of the segmented images reveal that the FM-rMST model is slightly better at distinguishing noises in this region.

Model	MCR	RI	VI	GCE
FM-MN	0.1274	0.7787	0.1535	0.8247
FM-rMSN	0.0748	0.8615	0.1089	0.6061
FM-rMST	0.0116	0.9699	0.0295	0.2153
FM-MNIG	0.2177	0.6606	0.2044	1.0892
FM-HMT	0.0619	0.8851	0.0933	0.5232
FM-MGH	0.0148	0.9648	0.0340	0.2407

Table 9 Comparison of image segmentation result on a natural colour image

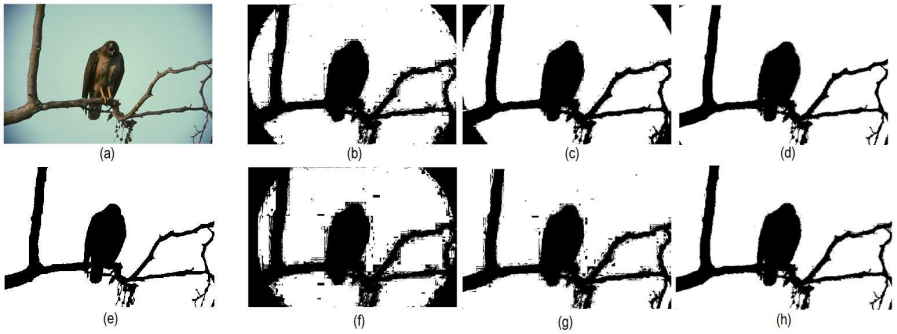


Fig. 5 Segmentation of natural colour image from the Berkeley's image segmentation data set (42049). Segmented image using various mixture model. (a) Original image; (b) segmentation result using the multivariate normal mixture model; (c) segmentation result using the multivariate skew normal mixture model; (d) segmentation result using the multivariate skew t-mixture model; (e) human segmentation; (f) segmentation result using the multivariate normal-inverse-Gaussian mixture model; (g) segmentation result using the multivariate hierarchical (two-level) mixture of t-distributions; (h) segmentation result using the multivariate generalized hyperbolic distribution.

6 Concluding Remarks

In this paper, we have discussed some of the more popular non-normal mixture models, and compared their performance on four real datasets. It has been observed from the examples considered here that, in general, the unrestricted models can improve the clustering results in comparison to their respective restricted models. They can also adapt to some unusual asymmetric shapes better than their restricted counterparts, for example, the shape of the cluster of solvent firms in the Bankruptcy data. We note that the restricted models do not seem to be able to produce this type of shape, which may be due to the way the skewness parameter is characterized in these models.

However, there is a higher computational cost involved in fitting the unrestricted mixture models. With the more general form of skew distributions, the conditional expectations involved in the E-step of the EM algorithm may not be able to be expressed in closed form. Numerical procedures for multidimensional integration can be computationally expensive, rendering some of the unrestricted models prohibitive in applications involving high-dimensional datasets. For example, consider the bivariate bankruptcy dataset example in Section 5.3, the average CPU time per iteration for the unrestricted skew-

normal mixture model is approximately double that of the restricted model, and the FM-uMST model is 1.5 times slower than the FM-rMST model in this example. In the trivariate AIS dataset example, the ratio between the computation time of the unrestricted and restricted models increase to 10 and 27 times, respectively, for the MSN and MST case. Future research is needed to look at developing faster algorithms for the fitting of mixtures of general skew distributions, or strategies for making them more practically affordable in higher dimensional applications.

Some researchers have considered mixtures of less well-known asymmetric distributions, such as the MSAL and MNIG distributions discussed in Section 4. These distribution can have a closed-form EM algorithm for model fitting, as in the case of FM-MSAL and FM-MNIG, where the EM-steps involve relatively simpler expressions. They thus have the potential to provide feasible alternatives to skew symmetric mixture models. In particular, they can be used to provide possible initial partitions of the data for the subsequent application of more computationally intensive models.

References

Aghaeepour N, Finak G, Consortium TF, Consortium TD, Hoos H, Mosmann TR, Brinkman R, Gottardo R, Scheuermann RH (2013) Critical assessment of automated flow cytometry data analysis techniques. Nature Methods 10:228–238

Altman EI (1968) Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. Journal of Finance 23(4):589–609

Arellano-Valle RB, Azzalini A (2006) On the unification of families of skew-normal distributions. Scandinavian Journal of Statistics 33:561–574

Arellano-Valle RB, Genton MG (2005) On fundamental skew distribtuions. Journal of Multivariate Analysis 96:93–116

Arellano-Valle RB, Genton MG (2010a) Multivariate extended skew-t distributions and related families. Metron - special issue on 'Skew-symmetric and flexible distributions' 68:201–234

Arellano-Valle RB, Genton MG (2010b) Multivariate unified skew-elliptical distributions. Chilean Journal of Statistics 1:17–33

Arellano-Valle RB, del Pino G, Martin ES (2002) Definition and probabilistic properties of skew-distributions. Statistics and Probability Letters 58(2):111 – 121

Arellano-Valle RB, Branco MD, Genton MG (2006) A unified view on skewed distributions arising from selections. The Canadian Journal of Statistics 34:581–601

Arnold BC, Beaver RJ, Meeker WQ (1993) The nontruncated marginal of a truncated bivariate normal distribution. Psychometrika 58:471–488

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

Azzalini A, Capitanio A (1999) Statistical applications of the multivariate skew-normal distribution. Journal of the Royal Statistical Society Series B 61(3):579–602

Azzalini A, Capitanio A (2003) Distribution generated by perturbation of symmetry with emphasis on a multivariate skew t distribution. Journal of the Royal Statistical Society Series B 65(2):367–389

Azzalini A, Dalla Valle A (1996) The multivariate skew-normal distribution. Biometrika 83(4):715–726

Banfield JD, Raftery A (1993) Model-based gaussian and non-gaussian clustering. Biometrics 49:803–821

Barndorff-Nielsen OE (1977) Exponentially decreasing distributions from the logarithm of of particle size. Proc RoySocLond A353:401–419

Basso RM, Lachos VH, Cabral CRB, Ghosh P (2010) Robust mixture modeling based on scale mixtures of skew-normal distributions. Computational Statistics and Data Analysis 54:2926–2941

Böhning D (1999) Computer-Assisted Analysis of Mixtures and Applications: Meta-Analysis, Disease Mapping and Others. Chapman and Hall/CRC Press, London

Branco MD, Dey DK (2001) A general class of multivariate skew-elliptical distributions. Journal of Multivariate Analysis 79:99–113

Browne RP, McNicholas PD (2013) A mixture of generalized hyperbolic distributions. arXiv:13051036 [statME]

Cabral CS, Lachos VH, Prates MO (2012) Multivariate mixture modeling using skew-normal independent distributions. Computational Statistics and Data Analysis 56:126–142

Calò AG, Montanari A, Viroli C (2013) A hierarchical modeling approach for clustering probability density functions. Computational Statistics and Data Analysis DOI 10.1016/j.csda.2013.04.013

Charytanowicz M, Niewczas J, Kulczycki P, Kowalski P, Lukasik S, Zak S (2010) A complete gradient clustering algorithm for features analysis of x-ray images. In: Pietka E, Kawa J (eds) Information Technologies in Biomedicine, Springer-Verlag, Berlin-Heidelberg, pp 15–24

Choi P, Min I (2011) A comparison of conditional and unconditional approaches in value-at-risk estimation. the Journal of the Japanese Economic Association 62:99–115

Christoffersen PF (1998) Evaluating interval forecasts. International Economic Review 39:841–862

Contreras-Reyes JE, Arellano-Valle RB (2012) Growth curve based on scale mixtures of skew-normal distributions to model the age-length relationship of cardinalfish (epigonus crassicaudus). arXiv:12125180 [statAP]

Cook RD, Weisberg S (1994) An Introduction to Regression Graphics. John Wiley & Sons, New York

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of Royal Statistical Society , Series B 39:1–38

Everitt BS, Hand DJ (1981) Finite Mixture Distributions. Chapman and Hall, London

Fang KT, Kotz S, Ng K (1990) Symmetric multivariate and related distributions. Chapman & Hall, London

Fraley C, Raftery AE (1999) How many clusters? which clustering methods? answers via model-based cluster analysis. Computer Journal 41:578–588

Franczak BC, Browne RP, McNicholas PD (2012) Mixtures of shifted asymmetric laplace distributions. arXiv:12071727 [statME]

Frühwirth-Schnatter S (2006) Finite mixture and Markov switching models. Springer, New York

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

Ganesalingam S, McLachlan GJ (1978) The efficiency of a linear discriminant function based on unclassified initial samples. Biometrika 65:658–662

González-Farás G, Domínguez-Molinz JA, , Gupta AK (2004) Additive properties of skew normal random vectors. Journal of Statistical Planning and Inference 126:521–534

Gupta AK (2003) Multivariate skew-t distribution. Statistics 37:359–363

Gupta AK, González-Faríaz G, Domínguez-Molina JA (2004) A multivariate skew normal distribution. Journal of Multivariate Analysis 89:181–190

Hubert L, Arabie P (1985) Comparing partitions. Journal of Classification 2:193–218

Jones PN, McLachlan GJ (1989) Modelling mass-size particle data by finite mixtures. Communications in Statistics - Theory and Methods 18:2629–2646

Jordan MI, Jacobs RA (1992) Hierarchies of adaptive experts. In: Moody J, Hanson S, Lippmann R (eds) Advances in Neural Information Processing Systems 4, California: Morgan Kaufmann, pp 985–993

Karlis D, Santourian A (2009) Model-based clustering with non-elliptically contoured distributions. Statistics and Computing 19:73–83

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

Kotz S, Kozubowski TJ, Podgórski K (2001) The Laplace Distribution and Generalizations: A Revisit With Applications to Communications, Economics, Engineering, and Finance. Birkhauser, Boston

Kupiec P (1995) Techniques for verifying the accuracy of risk management models. The Journal of Derivatives 3:73–84

Lachos VH, Ghosh P, Arellano-Valle RB (2010) Likelihood based inference for skew normal independent linear mixed models. Statistica Sinica 20:303–322

Lee S, McLachlan GJ (2011) On the fitting of mixtures of multivariate skew t-distributions via the EM algorithm. arXiv:11094706 [statME]

Lee S, McLachlan GJ (2013a) Finite mixtures of multivariate skew t-distributions: some recent and new results. Statistics and Computing DOI 10.1007/s11222-012-9362-4

Lee SX, McLachlan GJ (2013b) EMMIX-uskew: An `R` package for fitting mixtures of multivariate skew t-distributions via the EM algorithm. Journal of

Statistical Software Preprint arXiv:1211.5290
Lee SX, McLachlan GJ (2013c) On mixtures of skew-normal and skew t-distributions. Advances in Data Analysis and Classification DOI 10.1007/s11634-013-0132-8, preprint arXiv:1211.3602
Lin TI (2009) Maximum likelihood estimation for multivariate skew-normal mixture models. Journal of Multivariate Analysis 100:257–265
Lin TI (2010) Robust mixture modeling using multivariate skew t distribution. Statistics and Computing 20:343–356
Lin TI, Ho HJ, Kee CR (2013) Flexible mixture modelling using the multivariate skew-t-normal distribution. Statistics and Computing DOI DOI10.1007/s11222-013-9386-4
Lindsay BG (1995) Mixture Models: Theory, Geometry, and Applications. NSF-CBMS Regional Conference Series in probability and Statistics, Vol. 5 (Institute of Mathematical Statistics and the American Statistical Association), Alexandria, VA
Liseo B, Loperfido N (2003) A Bayesian interpretation of the multivariate skew-normal distribution. Statistics & Probability Letters 61:395–401
Lo K, Brinkman RR, Gottardo R (2008) Automated gating of flow cytometry data via robust model-based clustering. Cytometry part A 73:312–332
Lo K, Hahne F, Brinkman RR, Gottardo R (2009) flowclust: a bioconductor package for automated gating of flow cytometry data. BMC Bioinformatics 10:145
Martin D, Fowlkes C, Tal D, Malik J (2001) A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. Proc Int Conf Comput Vis 2:416–423
McLachlan GJ, Basford KE (1988) Mixture Models: Inference and Applications. Marcel Dekker, New York
McLachlan GJ, Krishnan T (2008) The EM Algorithm and Extensions, 2nd edn. Wiley-Interscience, Hokoben, N. J.
McLachlan GJ, Peel D (1998) Robust cluster analysis via mixtures of multivariate t-distributions. In: Amin A, Dori D, Pudil P, Freeman H (eds) Lecture Notes in Computer Science, Berlin: Springer-Verlag, pp 658–666
McLachlan GJ, Peel D (2000) Finite Mixture Models. Wiley Series in Probability and Statistics, New York
McNeil AJ, Frey R, Embrechts P (2005) Quantitative risk management : concepts, techniques and tools. Princeton University Press, USA
Meignen S, Meignen H (2006) On the modeling of small sample distributions with generalized gaussian density in a maximum likelihood framework. IEEE Transactions on Image Processing 15:1647–1652
Meilă M (2005) Comparing clusterings - an axiomatic view. In: In ICML 05: Proceedings of the 22nd international conference on Machine learning, ACM Press, pp 577–584
Mengersen KL, Robert CP, Titterington DM (2011) Mixtures: Estimation and Applications. John Wiley & Sons, NewYork
Nadarajah S (2008) Skewed distributions generated by the student's t kernel. Monte Carlo Methods and Applications 13:289–404

Nadarajah S, Kotz S (2003) Skewed distributions generated by the normal kernel. Statistics and Probability Letters 65:269–277

Nguyen TM, Wu QMJ (2013) A nonsymmetric mixture model for unsupervised image segmentation. IEEE Transactions on Cybernetics 43:751–765

Nikolic R (2010) `flowKoh`: Self-organizing map for flow cytometry data analysis. URL `http://commons.bcit.ca/radina_nikolic/docs/flowKoh_R_Code.zip`

Prates M, Lachos V, Cabral C (2011) `mixsmsn`: Fitting finite mixture of scale mixture of skew-normal distributions. URL `http://CRAN.R-project.org/package=mixsmsn`, R package version 0.3-2

Pyne S, Hu X, Wang K, Rossin E, Lin TI, Maier LM, Baecher-Allan C, McLachlan GJ, Tamayo P, Hafler DA, De Jager PL, Mesirow JP (2009a) Automated high-dimensional flow cytometric data analysis. Proceedings of the National Academy of Sciences USA 106:8519–8524

Pyne S, Hu X, Wang K, Rossin E, Lin TI, Maier LM, Baecher-Allan C, McLachlan GJ, Tamayo P, Hafler DA, De Jager PL, Mesirow JP (2009b) `FLAME`: Flow analysis with Automated Multivariate Estimation. URL `http://www.broadinstitute.org/cancer/software/genepattern/modules/FLAME/published\_data`

Qian Y, Wei C, Lee F, Campbell J, Halliley J, Lee J, Cai J, Kong Y, Sadat E, Thomson E (2010) Elucidation of seventeen human peripheral blood b-cell subsets and quantification of the tetanus response using a density-based method for the automated identification of cell populations in multidimensional flow cytometry data. Cytometry part B 78:S69–S82

R Development Team (2011) R: A Language and Environment for Statistical Computing. URL `http://www.R-project.org/`, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0

Rand WM (1971) Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association 66:846–850

Riggi S, Ingrassia S (2013) Modeling high energy cosmic rays mass composition data via mixtures of multivariate skew-t distributions. arXiv:13011178 [astro-phHE]

Rodrigues J (2006) A bayesian inference for the extended skew-normal measurement error model. Brazilian Journal of Probability and Statistics 20:179–190

Sahu SK, Dey DK, Branco MD (2003) A new class of multivariate skew distributions with applications to Bayesian regression models. The Canadian Journal of Statistics 31:129–150

Soltyk S, Gupta R (2011) Application of the multivariate skew normal mixture model with the EM algorithm to Value-at-Risk. MODSIM 2011 - 19th International Congress on Modelling and Simulation, Perth, Australia, December 12-16, 2011

Titterington DM, Smith AFM, Markov UE (1985) Statistical analysis of finite mixture distributions. Wiley, New York

Vrbik I, McNicholas PD (2012) Analytic calculations for the EM algorithm for multivariate skew t-mixture models. Statistics and Probability Letters

82:1169–1174

Wang K, McLachlan GJ, Ng SK, Peel D (2009) `EMMIX-skew`: EM Algorithm for Mixture of Multivariate Skew Normal/t Distributions. URL `http://www.maths.uq.edu.au/~gjm/mix\_soft/EMMIX-skew`, R package version 1.0-12

Zhang Y, Brady M, Smith S (2001) Segmentation of brain mr images through a hidden markov random field model and the expectation maximization algorithm. IEEE Transactions on Medical Imaging 20:45–57