The corrections below have been made in the amended version that follows.

(1) In Page 3, line above equation (3), "$\boldsymbol{\Omega}$" should be "$\boldsymbol{\omega}$."
(2) In the right-hand side of equation (3), replace "$\boldsymbol{\delta}$" and "$\boldsymbol{\delta}^T$" by "$\boldsymbol{\Delta}$" and "$\boldsymbol{\Delta}^T$", respectively.
(3) In the right-hand side of equation (6), replace "$\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\delta}$" and "$\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{\frac{1}{2}}$" each by "$\mathbf{O}$".
(4) In the right-hand side of equation (7), replace "$\boldsymbol{\Delta}$" and "$\boldsymbol{\Delta}^T$" each by "$\mathbf{O}$".
(5) In the right-hand side of equation (8), replace "$\boldsymbol{\delta}^T$" by "$\boldsymbol{\Delta}^T$".
(6) In the line below equation (8), replace "$\boldsymbol{\Delta}\boldsymbol{\Delta}^T$" by "$\boldsymbol{\Delta}^2$" in "$\boldsymbol{\Sigma} + \boldsymbol{\Delta}\boldsymbol{\Delta}^T$" and replace "$\boldsymbol{\Delta}^T$" by "$\boldsymbol{\Delta}''$ in "$\boldsymbol{\Delta}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Delta}''$.
(7) In the right-hand side of equation (10), replace "$\boldsymbol{\delta}$" by "$\mathbf{O}$" (twice).
(8) In the right-hand side of equation (12), replace "$\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\delta}$" and "$\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{\frac{1}{2}}$" each by "$\mathbf{O}$".
(9) In the right-hand side of equation (14), replace "$\boldsymbol{\Delta}$" by "$\mathbf{O}$" (twice).
(10) In Page 6 in the last paragraph of Section 3, replace "Azzalini and Dalla (1996)" with "Azzalini and Capitanio (2003)."
(11) In the right-hand side of equation (27), in the denominator of the first line, replace "$^{(k)}$" by "$\nu_h^{(k)}$" and "$T_{2,\nu_h^{(k)}}$" by "$T_{1,\nu_h^{(k)}}$".
(12) In the right-hand side of equation (29), in the second term, replace $\lambda_h^{(k)^2}$ by $\lambda_h^{(k)}\sqrt{\frac{\nu_h^{(k)}+p}{\nu_h^{(k)}+d_h^{(k)}(\boldsymbol{y}_j)}}$.

Alternatively, $e_{4,hj}^{(k)}$ can be reduced to $\lambda_h^{(k)^2} + q_{hj}^{(k)} e_{3,hj}^{(k)}$.

Finite mixtures of multivariate skew t-distributions: some recent and new results

Sharon Lee · Geoffrey J. McLachlan

Abstract Finite mixtures of multivariate skew t (MST) distributions have proven to be useful in modelling heterogeneous data with asymmetric and heavy tail behaviour. Recently, they have been exploited as an effective tool for modelling flow cytometric data. A number of algorithms for the computation of the maximum likelihood (ML) estimates for the model parameters of mixtures of MST distributions have been put forward in recent years. These implementations use various characterizations of the MST distribution, which are similar but not identical. While exact implementation of the expectation-maximization (EM) algorithm can be achieved for 'restricted' characterizations of the component skew t-distributions, Monte Carlo (MC) methods have been used to fit the 'unrestricted' models. In this paper, we review several recent fitting algorithms for finite mixtures of multivariate skew t-distributions, at the same time clarifying some of the connections between the various existing proposals. In particular, recent results have shown that the EM algorithm can be implemented exactly for faster computation of ML estimates for mixtures with unrestricted MST components. The gain in computational time is effected by noting that the semi-infinite integrals on the E-step of the EM algorithm can be put in the form of moments of the truncated multivariate non-central t-distribution, similar to the restricted case, which subsequently can be expressed in terms of the non-truncated form of the central t-distribution function for which fast algorithms are available. We present comparisons to illustrate the relative performance of the restricted and unrestricted models, and demonstrate the usefulness of the recently proposed methodology for the unrestricted MST mixture, by some applications to three real datasets.

Keywords Mixture models · EM algorithm · Skew normal distributions · Skew t component distributions

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

1 Introduction

Finite mixture distributions have become increasingly popular in the modelling and analysis of data due to their flexibility. This use of finite mixture distributions to model heterogeneous data has undergone intensive development in the past decades, as witnessed by the numerous applications in various scientific fields such as bioinformatics, cluster analysis, genetics, information processing, medicine, and pattern recognition. Comprehensive surveys on mixture models and their applications can be found, for example, in the monographs by Everitt and Hand (1981), Titterington et al. (1985), McLachlan and Basford (1988), Lindsay (1995), Böhning (1999), and Frühwirth-Schnatter (2006); see also the papers by Banfield and Raftery (1993) and Fraley and Raftery (1999).

Mixtures of multivariate t-distributions, as proposed by McLachlan and Peel (1998, 2000), provide extra flexibility over normal mixtures; see also Peel and McLachlan (2000). The thickness of tails can be regulated by an additional parameter – the degrees of freedom, thus enabling it to accommodate outliers better than normal distributions. However, in many practical problems, the data often involve observations whose distributions are highly asymmetric as well as having longer tails than the normal; for example, datasets from flow cytometry Pyne et al. (2009).

Azzalini (1985) introduced the so-called skew-normal (SN) distribution for modelling assymmetry in datasets. This sparked a renewed interest in the study of "skewed"

distributions. Following the development of the (univariate) SN and skew t-mixture models by Lin et al. (2007b), and Lin et al. (2007a), respectively, Basso et al. (2010) studied a class of mixture models where the component densities are scale mixtures of the (univariate) skew-normal distribution introduced by Branco and Dey (2001), known as the skew-normal independent (SNI) family, which include the classical skew-normal and skew t-distributions as special cases. Recently, Cabral, Lachos, and Prates (2012) have extended the work of Basso et al. (2010) to the multivariate case. Finite mixtures of multivariate skew-normal and skew t-distributions were also studied in Frühwirth-Schnatter and Pyne (2010), from a Bayesian viewpoint, where the characterization of Azzalini and Dalla Valle (1996) and Azzalini and Capitanio (2003) are adopted for the component distributions, respectively.

In a study of automated flow cytometry analysis, Pyne et al. (2009) proposed another finite mixture of multivariate skew normal and skew t-distributions based on a 'restricted' variant of the skew-elliptical distributions introduced by Sahu, Dey, and Branco (2003), hereafter referred to as the restricted multivariate skew normal (rMSN) and restricted skew t (rMST) distribution, respectively. Wang et al. (2009) obtained a closed-form EM algorithm (Dempster, Laird, and Rubin, 1977) for this model. Very recently, Virbik and McNicholas (2012) presented an alternative implementation for the rMST mixture model which involves hypergeometric functions.

It is important to note that the SNI skew-normal (SNI-SN) and skew-t (SNI-ST) distributions discussed in Cabral et al. (2012), and also the skewed distributions used in Frühwirth-Schnatter and Pyne (2010), are equivalent in reparameterization to the rMSN and rMST formulation, and hence can be considered as restricted characterizations of the MSN and MST distributions. A closed form implementation of the EM algorithm for mixtures of the SNI-SN and SNI-ST distributions is presented in Cabral et al. (2012).

An alternative to the skew t-mixture model was introduced by Karlis and Santourian (2009), known as the normal inverse Gaussian (NIG) mixture distribution. Like the skew t-model, the NIG mixture distribution can take flexible shapes, including heavy tail and skewness, and allows for a closed form EM-type algorithm for ML estimation.

Another promising alternative to the skew t-mixture model is the skew t-normal (STN) mixture model recently studied in Cabral et al. (2008) and Ho et al. (2012b), the former using a Bayesian approach. The (univariate) STN distribution, introduced by Gómez et al. (2007), was shown to have a larger range of skewness and kurtosis than the traditional skew distributions. The STN and ST distributions share the same set of parameters, but the former have a lower computational burden in estimation. Extension of existing results on the STN mixture to the multivariate case can potentially provide a favourable alternative to the MST mixture model.

In Lin (2010), a mixture model with unrestricted component skew t-distributions was considered, adopting the characterization by Sahu et al. (2003). However, with this more general formulation, maximum likelihood (ML) estimation via the EM algorithm can no longer be implemented in closed form due to the intractability of some of the conditional expectations involved on the E-step. To work around this, Lin (2010) proposed a Monte Carlo (MC) version of the E-step. One drawback of this approach is that the model fitting procedure relies on MC estimates which can be computationally expensive. Another issue is that there is no guarantee of an increase in the log likelihood at each iteration.

More recently, Lee and McLachlan (2011) and Ho et al. (2012a) independently developed exact expressions for two of the intractable conditional expectations involved in the EM algorithm for fitting mixtures of uMST distributions. They showed that the EM algorithm can be implemented exactly to calculate the ML estimates of the parameters for the (unrestricted) multivariate skew t-mixture model. This is achieved by using analytically reduced expressions for the conditional expectations, suitable for numerical evaluation using readily available software. A key factor in being able to compute the integrals quickly by numerical means is the recognition that they can be expressed as moments of a truncated multivariate non-central t-distribution, which in turn can be expressed in terms of the distribution function of a (non-truncated) multivariate central t-random vector, for which fast programs already exist. We note that, however, there are a few incorrectly specified results in the two papers. The corrections will be discussed in Section 5. In addition, Lee and McLachlan (2011) proposed a one-step late (OSL) approach to the EM implementation for mixtures of unrestricted MST which provides a simple closed-form expression for another intractable conditional expectation involved in the E-step.

In this paper, we provide an overview of recent developments concerned with fitting mixtures of multivariate skew t-distributions, with special reference to implementations based on the EM algorithm or extensions of it. The performance of restricted and unrestricted models in clustering real datasets are studied. In some cases, the clustering capacity of restricted

MST mixtures can be improved by adopting the unrestricted model. We also show that the closed-form EM implementation for mixtures of unrestricted MST distributions is more efficient compared to the version with a MC E-step. It produces more accurate results for which, if MC were to achieve comparable accuracy, a large number of draws would be necessary. Also, this implementation maintains stable and monotonic convergence to a local maximizer. Moreover, if the exact ECME implementation is adopted, efficient and stable monotonic convergence is guaranteed.

The remainder of the paper is organized as follows. In Section 2, we begin with a discussion of several variants of the multivariate skew normal distributions mentioned earlier, clarifying their relationships. Section 3 provides a description of various characterizations of multivariate skew t-distributions used for defining the multivariate skew t-mixture models in existing proposals. In Section 4, we study some existing implementations of the EM algorithm for obtaining ML estimates for the restricted MST distribution. In Section 5, we examine the EM algorithm for fitting the unrestricted MST distribution, and outline corrections to the two recent papers on this topic. We also present a fast implementation of the EM algorithm for the fitting of the unrestricted FM-MST model. In Section 6, we present some applications of the proposed methodology and comparisons with other alternative implementations. Finally, in section 7, we conclude with a discussion on the computational aspects of the algorithms for mixtures of the uMST distributions.

2 Multivariate Skew-normal distributions

We begin by defining the restricted multivariate skew-normal (rMSN) distribution and briefly describing some related properties. Some alternative parameterizations of the distribution are also discussed. Next, we give the definition of the unrestricted multivariate skew-normal (uMSN) distribution.

2.1 The restricted multivariate skew-normal distribution

Both the restricted and unrestricted skew-normal distribution belong to the class of fundamental skew-normal (FUSN) distributions (Arellano-Valle and Genton, 2005). The restricted case is one of the simplest special cases of FUSN. In particular, the density of the rMSN distribution can be expressed as a product of a multivariate normal density and a *univariate* normal distribution function; that is, the skewing function is of dimension one. Accordingly, the skew-normal distribution used by Pyne et al. (2009), Frühwirth-Schnatter and Pyne (2010), and Cabral et al. (2012) in constructing their finite mixture models is the rMSN distribution, or a reparameterization of it.

2.1.1 The classic skew-normal distribution

The classic multivariate skew-normal distribution refers to one of the early multivariate generalizations of the univariate skew-normal density introduced by Azzalini (1985). Following Azzalini and Dalla Valle (1996), a p-dimensional random vector $\boldsymbol{Y}$ has a (classic) skew-normal distribution, denoted by $\boldsymbol{Y} \sim SN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\delta})$ with $p \times 1$ location vector $\boldsymbol{\mu}$, $p \times p$ scale matrix $\boldsymbol{\Sigma}$, and $p \times 1$ skewness vector $\boldsymbol{\delta}$ if its density is given by

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

where $\lambda^2 = 1 - \delta^T\boldsymbol{R}^{-1}\boldsymbol{\delta}$, $\boldsymbol{R} = \boldsymbol{\omega}^{-1}\boldsymbol{\Sigma}\boldsymbol{\omega}^{-1}$ is the correlation matrix, and $\boldsymbol{\omega}$ is the diagonal matrix created by extracting the main diagonal elements of $\boldsymbol{\Sigma}$, that is, $\boldsymbol{\omega} = \text{DIAG}(\sqrt{\Sigma_{11}}, \ldots, \sqrt{\Sigma_{pp}})$, where Σ_{ij} denotes the ijth entry of $\boldsymbol{\Sigma}$. Here, the operator $\text{DIAG}(\boldsymbol{\delta})$ denotes a diagonal matrix with diagonal elements specified by the vector $\boldsymbol{\delta}$. Also, we let $\phi_p(.; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the p-dimensional normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and $\Phi_1(.)$ is the (cumulative) distribution function of a standard (univariate) normal random variable. Note that when $\boldsymbol{\delta} = 0$, (1) reduces to the normal density $\phi_p(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$.

The SN distribution (1) can be obtained by several stochastic mechanisms (Azzalini, 2005); for example, via the conditioning approach as follows. Let $\boldsymbol{U}_1$ and U_0 be random variables with dimensions p and 1, respectively. Then $\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\omega}(\boldsymbol{U}_1 \mid U_0 > 0)$ has the density (1), where

$$\begin{bmatrix} U_0 \\ \boldsymbol{U}_1 \end{bmatrix} \sim N_{1+p}\left(\begin{bmatrix} 0 \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} 1 & \boldsymbol{\delta}^T \\ \boldsymbol{\delta} & \boldsymbol{R} \end{bmatrix}\right), \quad (2)$$

and where $\boldsymbol{0}$ denotes the zero vector of appropriate dimension. In the above, the notation $\boldsymbol{U} = (\boldsymbol{U}_1 \mid u_0 > 0)$ is taken to imply that $\boldsymbol{U} = \boldsymbol{U}_1$ if the constraint $u_0 > 0$ is satisfied, otherwise $\boldsymbol{U} = -\boldsymbol{U}_1$. The skew component in the skew-normal mixture model of Frühwirth-Schnatter and Pyne (2010) uses this characterization for its component densities.

2.1.2 The restricted skew-normal distribution

In Pyne et al. (2009), the authors proposed a simplified version of the skew-normal density given by Sahu

et al. (2003). This characterization arises from a simple convolution-type stochastic mechanism. Specifically, the random vector $\boldsymbol{Y} = \boldsymbol{\mu}+\boldsymbol{\delta}|U_0|+\boldsymbol{U}_1$ has a restricted skew normal density, where

$$\begin{bmatrix} U_0 \\ \boldsymbol{U}_1 \end{bmatrix} \sim N_{1+p}\left(\begin{bmatrix} 0 \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} 1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma} \end{bmatrix}\right). \tag{3}$$

In this characterization, the scale matrix $\boldsymbol{\Sigma}$ is not factored into $\boldsymbol{\omega R \omega}$, leading to a slightly simpler expression for the density, given by

$$f(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\delta}) = 2\phi_p(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Omega})\,\Phi_1\left(\lambda^{-1}\boldsymbol{\delta}^T\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right), \tag{4}$$

where $\boldsymbol{\Omega} = \boldsymbol{\Sigma} + \boldsymbol{\delta}\boldsymbol{\delta}^T$ and $\lambda^2 = 1 - \boldsymbol{\delta}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\delta}$. We shall adopt the notation $\boldsymbol{Y} \sim rSN_p(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\delta})$. It can be shown that after some suitable reparameteization, the classic SN and the rMSN are equivalent. This characterization of the skew-normal distribution was adopted in the work of Pyne et al. (2009) when formulating finite mixtures of skew-normal distributions.

2.1.3 The skew-normal/independent skew-normal distribution

The skew-normal/independent (SNI) family are scale mixtures of skew-normal distribution (Cabral et al., 2012), much similar to the skew-elliptical class of Branco and Dey (2001). In their definition, the skew-normal density, hereafter SNI-SN, is given by

$$f(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\delta}) = 2\phi_p(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma})\Phi_1\left(\lambda^{-1}\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{y}-\boldsymbol{\mu})\right), \tag{5}$$

where $\lambda^2 = 1 - \boldsymbol{\delta}^T\boldsymbol{\delta}$. This definition of the SN distribution has a corresponding convolution-type stochastic representation given by $\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\delta}|U_0| + (\boldsymbol{I}_p - \boldsymbol{\delta}\boldsymbol{\delta}^T)^{1/2}\boldsymbol{U}_1$, where

$$\begin{bmatrix} U_0 \\ \boldsymbol{U}_1 \end{bmatrix} \sim N_{1+p}\left(\begin{bmatrix} 0 \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} 1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma} \end{bmatrix}\right). \tag{6}$$

It can be observed from (5) and (6) that the SNI-SN distribution is equivalent to the rMSN distribution (4) by replacing $\boldsymbol{\delta}$ and $\boldsymbol{\Omega}$ with $\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\delta}$ and $\boldsymbol{\Sigma}$, respectively. ML estimation for the SNI-SN distribution and a mixture case of these distributions is studied in Cabral et al. (2012).

2.2 The unrestricted multivariate skew-normal distribution

The unrestricted multivariate skew-normal (uMSN) distribution can be viewed as a simple extension of the rMSN distribution in which the univariate latent variable U_0 is replaced by a multivariate analogue, that is, $\boldsymbol{U}_0$. This type of MSN distribution was studied in Sahu et al. (2003), and in the mixture case by Lin (2009). Suppose $\boldsymbol{U}_0$ and $\boldsymbol{U}_1$ are jointly normally distributed as

$$\begin{bmatrix} \boldsymbol{U}_0 \\ \boldsymbol{U}_1 \end{bmatrix} \sim N_{2p}\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{I}_p & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma} \end{bmatrix}\right), \tag{7}$$

where $\boldsymbol{\Delta} = \mathrm{DIAG}(\boldsymbol{\delta})$. Then $\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\Delta}|\boldsymbol{U}_0| + \boldsymbol{U}_1$ defines the unrestricted multivariate skew-normal distribution, and its density is given by

$$f(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\delta}) = 2^p\phi_p(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma})\,\Phi_p\left(\boldsymbol{\Delta}\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu});\boldsymbol{0},\boldsymbol{\Lambda}\right), \tag{8}$$

where $\boldsymbol{\Omega} = \boldsymbol{\Sigma}+\boldsymbol{\Delta}^2$, $\boldsymbol{\Lambda} = \boldsymbol{I}_p - \boldsymbol{\Delta}\boldsymbol{\Omega}^{-1}\boldsymbol{\Delta}$, and $\Phi_p(.;\boldsymbol{\mu},\Sigma)$ denotes the distribution function of a $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ random variable. We shall adopt the notation $\boldsymbol{Y} \sim uSN_p(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\delta})$ if $\boldsymbol{Y}$ has the uMSN density. Observe that with this characterization of the MSN distribution, each element of $\boldsymbol{\delta}$, δ_i $(i = 1,\ldots,p)$, is allowed to have a different random coefficient, namely $|U_{0i}|$, whereas with the restricted case they share the same (scalar) coefficient $|U_0|$. Note that when $\boldsymbol{\delta} = \boldsymbol{0}$, the second part of (8) evaluates to 2^{-p}, and we again recover the multivariate normal density $\phi_p(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma})$.

2.3 Other multivariate skew-normal distributions

Many other variants and extensions of multivariate skew-normal distributions have appeared in recent years. For a comprehensive survey on this topic, see, for example, the review papers by Azzalini (2005) and Arellano-Valle and Azzalini (2006). One important extension is the incorporation of an extension parameter, leading to the 'extended' skew-normal (eMSN) distributions; see, for example, the formulations by Arnold and Beaver (2002), González-Farás et al. (2004), Liseo and Loperfido (2003), and Arellano-Valle and Azzalini (2006). This type of formulation features the property of closure under conditioning, which does not hold for the restricted and unrestricted subfamilies. Another more general extension includes a relaxation of the requirements on the distribution of $\boldsymbol{U}_1$ and $\boldsymbol{U}_0$, which can be crudely termed as a 'generalized' skew-normal (gMSN) distribution. The skewing function in a gMSN distribution need not be a normal distribution function.

3 Multivariate skew t-distributions

The connections between the various multivariate skew t-distributions are analogues to those between the skew-normal distributions discussed in the previous section. In this section, we give the formal definition of the restricted and unrestricted MST distribution, and briefly outline their properties and relationships.

3.1 The restricted multivariate skew t-distribution

As formulated in Pyne et al. (2009), a p-dimensional random vector has a restricted multivariate skew t (rMST) distribution with $p \times 1$ location vector $\boldsymbol{\mu}$, $p \times p$ scale matrix $\boldsymbol{\Sigma}$, $p \times 1$ skewness vector $\boldsymbol{\delta}$, and scalar degrees of freedom ν, denoted by $\boldsymbol{Y} \sim rMST(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\delta}, \nu)$, if its density is given by

$$f(\boldsymbol{y}) = 2t_{p,\nu}\left(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Omega}\right) T_{1,\nu+p}\left(\frac{y_1}{\lambda}; 0, 1\right), \tag{9}$$

where

$$\begin{aligned}
\boldsymbol{\Omega} &= \boldsymbol{\Sigma} + \boldsymbol{\delta}\boldsymbol{\delta}^T, \\
d(\boldsymbol{y}) &= (\boldsymbol{y}-\boldsymbol{\mu})^T \boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}), \\
q &= \boldsymbol{\delta}^T \boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}), \\
y_1 &= q\sqrt{\frac{\nu+p}{\nu+d(\boldsymbol{y})}}, \\
\lambda^2 &= 1 - \boldsymbol{\delta}^T \boldsymbol{\Omega}^{-1}\boldsymbol{\delta},
\end{aligned}$$

Here, we let $t_{p,\nu}(.; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the p-dimensional t-distribution with location vector $\boldsymbol{\mu}$, scale matrix $\boldsymbol{\Sigma}$, and degrees of freedom ν, and $T_{1,\nu}(.; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the distribution function of a standard (univariate) t-random vector with ν degrees of freedom. This formulation of the MST distribution was adopted by Pyne et al. (2009), Wang et al. (2009), and Virbik and McNicholas (2012) in their construction of the MST mixture model. Note that when $\boldsymbol{\delta} = \boldsymbol{0}$, (9) reduces to the t-density $t_{p,\nu}(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$. Also, when $\nu \to \infty$, we obtain the restricted skew normal distribution.

The rMST distribution (9) corresponds to the density of $\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\delta}|U_0| + \boldsymbol{U}_1$, where conditional on a gamma random variable W, the joint distribution of U_0 and $\boldsymbol{U}_1$ is given by

$$\begin{bmatrix} U_0 \\ \boldsymbol{U}_1 \end{bmatrix} \mid W \sim N_{1+p}\left(\begin{bmatrix} 0 \\ \boldsymbol{0} \end{bmatrix}, \frac{1}{w}\begin{bmatrix} 1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma} \end{bmatrix}\right). \tag{10}$$

It follows that the rMST distribution (9) admits a convenient hierarchical characterization, given by

$$\begin{aligned}
\boldsymbol{Y} \mid \boldsymbol{u}, w &\sim N_p\left(\boldsymbol{\mu} + \boldsymbol{\Delta}\boldsymbol{u}, \tfrac{1}{w}\boldsymbol{\Sigma}\right), \\
U \mid w &\sim HN\left(0, \tfrac{1}{w}\right), \\
W &\sim \text{gamma}\left(\tfrac{\nu}{2}, \tfrac{\nu}{2}\right),
\end{aligned} \tag{11}$$

where $HN(0, \sigma^2)$ represents the univariate half-normal distribution with mean 0 and scale parameter σ^2, and gamma(α, β) is the gamma distribution with mean α/β.

The SNI-ST distribution (Cabral et al., 2012) is given by $\boldsymbol{Y} = \mu + \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\delta}|U_0| + (\boldsymbol{I}_p - \boldsymbol{\delta}\boldsymbol{\delta}^T)^{\frac{1}{2}}\boldsymbol{U}_1$, where

$$\begin{bmatrix} U_0 \\ \boldsymbol{U}_1 \end{bmatrix} \mid W \sim N_{1+p}\left(\begin{bmatrix} 0 \\ \boldsymbol{0} \end{bmatrix}, \frac{1}{w}\begin{bmatrix} 1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma} \end{bmatrix}\right). \tag{12}$$

Then the density of $\boldsymbol{Y}$ is given by

$$f(\boldsymbol{y}) = 2t_{p,\nu}(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)\, T_{1,\nu+p}\left(\frac{y_1^*}{\lambda^*}; 0, 1\right), \tag{13}$$

where

$$\begin{aligned}
y_1^* &= q^*\sqrt{\frac{\nu+p}{\nu+d^*(\boldsymbol{y})}}, \\
q^* &= \boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{y}-\boldsymbol{\mu}), \\
d^*(\boldsymbol{y}) &= (\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}), \\
\lambda^{*^2} &= 1 - \boldsymbol{\delta}^T\boldsymbol{\delta}.
\end{aligned}$$

Again, it can be observed from (12) and (13) that by replacing $\boldsymbol{\delta}$ and $\boldsymbol{\Omega}$ in (9) with $\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\delta}$ and $\boldsymbol{\Sigma}$, respectively, we obtain the SNI-ST density (13). The aforementioned rMST and the SNI-ST formulation are, in turn, equivalent (after suitable reparameterization) to the multivariate skew t-distributions of Azzalini and Capitanio (2003) and Gupta (2003).

3.2 The unrestricted multivariate skew t-distribution

Analogous to the skew-normal case, the unrestricted multivariate skew t (uMST) distribution has a similar stochastic representation to the rMST distribution, except that the latent variable U_0 is replaced by a multivariate variable. Specifically, $\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\Delta}|\boldsymbol{U}_0| + \boldsymbol{U}_1$ has the uMST distribution, where conditional on the gamma variable W,

$$\begin{bmatrix} \boldsymbol{U}_0 \\ \boldsymbol{U}_1 \end{bmatrix} \sim N_{2p}\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \frac{1}{w}\begin{bmatrix} \boldsymbol{I}_p & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma} \end{bmatrix}\right), \tag{14}$$

and where $\boldsymbol{I}_p$ denotes the $p \times p$ identity matrix, and $\boldsymbol{U}_0$ and $\boldsymbol{U}_1$ are p-dimensional random vectors. In the above, $|\boldsymbol{U}|$ denotes the vector whose ith element is the magnitude of the ith element of the vector $\boldsymbol{U}$.

It follows that the density of $\boldsymbol{Y}$ is given by

$$f_p(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\delta}, \nu) = 2^p t_{p,\nu}\left(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Omega}\right) T_{p,\nu+p}\left(\boldsymbol{y}_1; \boldsymbol{0}, \boldsymbol{\Lambda}\right), \tag{15}$$

where

$$\begin{aligned}
\boldsymbol{\Delta} &= \mathrm{DIAG}(\boldsymbol{\delta}),\\
\boldsymbol{\Omega} &= \boldsymbol{\Sigma} + \boldsymbol{\Delta}^2,\\
\boldsymbol{y}_1 &= \boldsymbol{q}\sqrt{\frac{\nu+p}{\nu+d(\boldsymbol{y})}},\\
\boldsymbol{q} &= \boldsymbol{\Delta}\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}),\\
d(\boldsymbol{y}) &= (\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}),\\
\boldsymbol{\Lambda} &= \boldsymbol{I}_p - \boldsymbol{\Delta}\boldsymbol{\Omega}^{-1}\boldsymbol{\Delta}.
\end{aligned}$$

Here, we let $T_{p,\nu}(.;\boldsymbol{\mu},\boldsymbol{\Sigma})$ be the (cumulative) distribution function of $t_{p,\nu}(.;\boldsymbol{\mu},\boldsymbol{\Sigma})$. The notation $\boldsymbol{Y} \sim \mathrm{ST}_{p,\nu}(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\delta})$ will be used. Note that the MST density (15) is expressed as the product of a multivariate t-density function (the symmetric part) and a multivariate t-distribution function (the skewing part). The symmetric part of the uMST distribution is identical to the rMST distribution, but the skewing part of the rMST distribution is univariate.

Similar to the restricted version, the uMST distribution admits a convenient hierarchical representation,

$$\begin{aligned}
\boldsymbol{Y} \mid \boldsymbol{u}, w &\sim N_p\left(\boldsymbol{\mu}+\boldsymbol{\Delta}\boldsymbol{u}, \tfrac{1}{w}\boldsymbol{\Sigma}\right),\\
\boldsymbol{U} \mid w &\sim HN_p\left(\boldsymbol{0}, \tfrac{1}{w}\boldsymbol{I}_p\right),\\
W &\sim \mathrm{gamma}\left(\tfrac{\nu}{2}, \tfrac{\nu}{2}\right),
\end{aligned} \tag{16}$$

where $HN_p(\boldsymbol{0},\boldsymbol{\Sigma})$ represents the p-dimensional half-normal distribution with mean $\boldsymbol{0}$ and scale matrix $\boldsymbol{\Sigma}$.

It is worth stressing again that, although also known as the multivariate skew t-distribution, the versions considered by Azzalini and Capitanio (2003), Gupta (2003), and Lachos et al. (2010), among others, are different from (15). These versions are simpler in that the skew t-density is defined in terms involving only the *univariate* t-distribution function instead of the multivariate form of the latter as used in (15).

One major advantage of having simplified forms like (9) is that calculations on the E-step can be expressed in closed form and can be evaluated faster than the uMST analogue. However, the form of skewness is limited in these characterizations. In Section 5, we study an extension of their approach to the more general form of the skew t-density as proposed by Sahu et al. (2003).

4 ML estimation for the restricted MST mixture model

The first multivariate rMST mixture model appeared in Pyne et al. (2009), and a closed-form EM implementation was obtained. An alternative exact implementation of the same model was presented in Virbik and McNicholas (2012).

4.1 The FM-rMST distribution

With reference to (9), the probability density function (pdf) of a finite mixture of g multivariate (restricted) skew t-components is given by

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

where $f(\boldsymbol{y};\boldsymbol{\mu}_h,\boldsymbol{\Sigma}_h,\boldsymbol{\delta}_h,\nu_h)$ denotes the hth rMST component of the mixture model as specified by (9), with location parameter $\boldsymbol{\mu}_h$, scale matrix $\boldsymbol{\Sigma}_h$, skew parameter $\boldsymbol{\delta}_h$, and degrees of freedom ν_h. The mixing proportions π_h satisfy $\pi_h \geq 0$ $h = 1,\ldots,g$ and $\sum_{h=1}^{g}\pi_h = 1$. The vector of unknown parameters $\boldsymbol{\Psi}$ contains $(\pi_1,\ldots,\pi_{g-1}, \boldsymbol{\theta}_1^T,\ldots,\boldsymbol{\theta}_g^T)$, where $\boldsymbol{\theta}_h$ consists of the unknown parameters of the hth component, namely, the elements of $\boldsymbol{\mu}_h$ and $\boldsymbol{\delta}_h$, the distinct elements of $\boldsymbol{\Sigma}_h$, and ν_h. We shall denote this model by FM-rMST.

4.2 ML estimation via the EM algorithm

Under the EM framework, the observed data vector $\boldsymbol{y} = (\boldsymbol{y}_1^T,\ldots,\boldsymbol{y}_n^T)$ is viewed as incomplete, and the associated vector of latent component labels $\boldsymbol{z}_j = (z_{1j},\ldots,z_{gj})^T$ is introduced, where each element z_{hj} is a binary variable defined as

$$z_{hj} = \begin{cases} 1, \text{ if } \boldsymbol{y}_j \text{ belongs to component } i,\\ 0, \qquad\qquad \text{otherwise,}\end{cases} \tag{18}$$

and $\sum_{h=1}^{g} z_{hj} = 1$ $(j = 1,\ldots,n)$. Hence, the random vector $\boldsymbol{Z}_j$ corresponding to $\boldsymbol{z}_j$ follows a multinomial distribution with one trial and cell probabilities $\pi_1,\ldots,\pi_g$; that is, $\boldsymbol{Z}_j \sim \mathrm{Mult}_g(1;\pi_1,\ldots,\pi_g)$.

It follows that the FM-rMST model can be represented in the hierarchical form given by

$$\begin{aligned}
\boldsymbol{Y}_j \mid \boldsymbol{u}_j, w_j, z_{hj}=1 &\sim N_p\left(\boldsymbol{\mu}_h+\boldsymbol{\delta}_h u_j, \frac{1}{w_j}\boldsymbol{\Sigma}_h\right),\\
U_j \mid w_j, z_{hj}=1 &\sim HN\left(0, \frac{1}{w_j}\right),\\
W_j \mid z_{hj}=1 &\sim \mathrm{gamma}\left(\frac{\nu_h}{2}, \frac{\nu_h}{2}\right),\\
\boldsymbol{Z}_j &\sim \mathrm{Mult}_g(1,\boldsymbol{\pi}),
\end{aligned} \tag{19}$$

where $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_g)^T$. This leads to the complete-data log likelihood function given in Pyne et al. (2009), Wang et al. (2009), and Virbik and McNicholas (2012).

The implementation of the EM algorithm requires alternating repeatedly the E- and M-steps until convergence in the case where the sequence of the log likelihood values $L(\boldsymbol{\theta}^{(k)})$ is bounded above. Here $\boldsymbol{\theta}^{(k)}$ denotes the value of $\boldsymbol{\theta}$ after the kth iteration.

At the E-step for the $(k+1)$th EM iteration, the following conditional expectations are required:

$$\tau_{hj}^{(k)} = E_{\Psi^{(k)}}\left\{\boldsymbol{Z}_{hj} \mid \boldsymbol{y}_j\right\}, \tag{20}$$

$$e_{1,hj}^{(k)} = E_{\Psi^{(k)}}\left\{\log(W_j) \mid \boldsymbol{y}_j, z_{hj}=1\right\}, \tag{21}$$

$$e_{2,hj}^{(k)} = E_{\Psi^{(k)}}\left\{W_j \mid, \boldsymbol{y}_j, z_{hj}=1\right\}, \tag{22}$$

$$e_{3,hj}^{(k)} = E_{\Psi^{(k)}}\left\{W_j U_j \mid \boldsymbol{y}_j, z_{hj}=1\right\}, \tag{23}$$

$$e_{4,hj}^{(k)} = E_{\Psi^{(k)}}\left\{W_j U_j^2 \mid \boldsymbol{y}_j, z_{hj}=1\right\}, \tag{24}$$

where $E_{\boldsymbol{\Psi}^{(k)}}$ denotes the expectation operator using $\boldsymbol{\Psi}^{(k)}$ for $\boldsymbol{\Psi}$.

The posterior probability of membership of the hth component by $\boldsymbol{y}_j$, using the current estimate $\boldsymbol{\Psi}^{(k)}$ for $\boldsymbol{\Psi}$, is given using Bayes' Theorem by

$$\tau_{hj}^{(k)} = \frac{\pi_h^{(k)} f\left(\boldsymbol{y}_j; \boldsymbol{\mu}_h^{(k)}, \boldsymbol{\Sigma}_h^{(k)}, \boldsymbol{\delta}_h^{(k)}, \nu_h^{(k)}\right)}{\sum_{h=1}^{g} \pi_h^{(k)} f\left(\boldsymbol{y}_j; \boldsymbol{\mu}_h^{(k)}, \boldsymbol{\Sigma}_h^{(k)}, \boldsymbol{\delta}_h^{(k)}, \nu_h^{(k)}\right)}. \tag{25}$$

The expressions for (21)-(23) first appeared in Pyne et al. (2009), given by (with some reparameterization):

$$e_{1,hj}^{(k)} = e_{2,hj}^{(k)} - \log\left(\frac{\nu_h^{(k)} + d_h^{(k)}(\boldsymbol{y}_j)}{2}\right) - \psi\left(\frac{\nu_h^{(k)}+p}{2}\right) - 1, \tag{26}$$

$$e_{2,hj}^{(k)} = \left(\frac{\nu_h^{(k)}+p}{\nu_h^{(k)} + d_h^{(k)}(\boldsymbol{y}_j)}\right) \frac{T_{1,\nu_h^{(k)}+p+2}\left(\frac{y_{2hj}^{(k)}}{\lambda_h^{(k)}}\right)}{T_{1,\nu_h^{(k)}+p}\left(\frac{y_{1hj}^{(k)}}{\lambda_h^{(k)}}\right)}, \tag{27}$$

$$e_{3,hj}^{(k)} = \lambda_h^{(k)} \sqrt{\frac{\nu_h^{(k)}+p}{\nu_h^{(k)} + d_h^{(k)}(\boldsymbol{y}_j)}} \frac{t_{1,\nu_h^{(k)}+p}\left(\frac{y_{1hj}^{(k)}}{\lambda_h^{(k)}}\right)}{T_{1,\nu_h^{(k)}+p}\left(\frac{y_{1hj}^{(k)}}{\lambda_h^{(k)}}\right)} + e_{1,hj}^{(k)} q_{hj}^{(k)}, \tag{28}$$

$$e_{4,hj}^{(k)} = {\lambda_h^{(k)}}^2 + q_{hj}^{(k)} e_{3,hj}^{(k)}, \tag{29}$$

where

$$\begin{aligned}
q_{hj}^{(k)} &= {\boldsymbol{\delta}_h^{(k)}}^T {\boldsymbol{\Omega}_h^{(k)}}^{-1} (\boldsymbol{y}_j - \boldsymbol{\mu}_h^{(k)}),\\
{\lambda_h^{(k)}}^2 &= 1 - {\boldsymbol{\delta}_h^{(k)}}^T {\boldsymbol{\Omega}_h^{(k)}}^{-1} \boldsymbol{\delta}_h^{(k)},\\
y_{1hj}^{(k)} &= q_{hj}^{(k)} \sqrt{\frac{\nu_h^{(k)}+p}{\nu_h^{(k)} + d_h^{(k)}(\boldsymbol{y}_j)}},\\
y_{2,hj}^{(k)} &= q_{hj}^{(k)} \sqrt{\frac{\nu_h^{(k)}+p+2}{\nu_h^{(k)} + d_h^{(k)}(\boldsymbol{y}_j)}},
\end{aligned} \tag{30}$$

and where $\psi(.)$ denotes the digamma function.

Virbik and McNicholas (2012) subsequently presented alternative expressions for the conditional expectations (21)-(24) in terms of hypergeometric functions,

$$e_{1,hj}^{(k)} = \psi\left(\frac{\nu_h^{(k)}+p+1}{2}\right) - \log\left(\frac{\nu_h^{(k)}}{2}\right) - \frac{I_3\left(\frac{\nu_h^{(k)}+p+1}{2}, \frac{q_{hj(k)}}{\lambda_h^{(k)}\sqrt{\nu_h^{(k)}+d_h^{(k)}(\boldsymbol{y}_j)}}\right)}{I_1\left(\frac{\nu_h^{(k)}+p+1}{2}, \frac{q_{hj(k)}}{\lambda_h^{(k)}\sqrt{\nu_h^{(k)}+d_h^{(k)}(\boldsymbol{y}_j)}}\right)} \tag{31}$$

$$e_{2,hj}^{(k)} = \left(\frac{\nu_h^{(k)}+p+1}{\nu_h^{(k)} + d_h^{(k)}(\boldsymbol{y}_j)}\right) I_1\left(\frac{\nu_h^{(k)}+p+1}{2}+1, \frac{-q_{hj}^{(k)}}{\lambda_h^{(k)}\sqrt{\nu_h^{(k)} + d_h^{(k)}(\boldsymbol{u}_j)}}\right), \tag{32}$$

$$e_{3,hj}^{(k)} = \frac{\lambda_h^{(k)}(\nu_h^{(k)}+p+1)}{\sqrt{\nu_h^{(k)} + d_h^{(k)}(\boldsymbol{y}_j)}} + q_{hj}^{(k)} e_{2,hj}^{(k)} \frac{I_2\left(\frac{\nu_h^{(k)}+p+1}{2}+1, \frac{-q_{hj}^{(k)}}{\lambda_h^{(k)}\sqrt{\nu_h^{(k)}+d_h^{(k)}(\boldsymbol{u}_j)}}\right)}{I_1\left(\frac{\nu_h^{(k)}+p+1}{2}, \frac{-q_{hj}^{(k)}}{\lambda_h^{(k)}\sqrt{\nu_h^{(k)}+d_h^{(k)}(\boldsymbol{u}_j)}}\right)}, \tag{33}$$

$$e_{4,hj}^{(k)} = {\lambda_h^{(k)}}^2\left(\nu_h^{(k)}+p+1\right) - \frac{q_{hj}^{(k)}}{2} e_{3,hj}^{(k)} - {\lambda_h^{(k)}}^2\left(\nu_h^{(k)} + d_h^{(k)}(\boldsymbol{y}_j) + \left(\frac{q_{hj}^{(k)}}{\lambda_h^{(k)}}\right)^2\right) e_{2,hj}^{(k)}, \tag{34}$$

where

$$I_1(\alpha,\beta) = \frac{\pi\Gamma(2\alpha-1)}{\Gamma(\alpha)^2 2^{2\alpha-1}} - \beta\ {}_2F_1\left(\frac{1}{2}, \alpha; \frac{3}{2}; -\beta^2\right), \tag{35}$$

$$I_2(\alpha,\beta) = -\frac{(1+\beta^2)^{1-\alpha}}{2(1-\alpha)}, \tag{36}$$

$$I_3(\alpha,\beta) = -\log(1+\beta^2)\left[\frac{(1+\beta^2)^{\frac{1}{2}-\alpha}}{1-2\alpha} H(\alpha,\beta,1) - \frac{(1+\beta^2)^{\frac{1}{2}-\alpha}}{(2\alpha+1)^2} H(\alpha,\beta,3)\right] + \frac{2(1+\beta^2)^{\frac{1}{2}-\alpha}}{(1-2\alpha)^2} H(\alpha,\beta,1), \tag{37}$$

$$H(\alpha,\beta,\gamma) = {}_3F_2\left(\frac{\gamma}{2}, \alpha+\frac{\gamma}{2}-1, \alpha+\frac{\gamma}{2}-1; \alpha+\frac{\gamma}{2}, \alpha+\frac{\gamma}{2}; \frac{1}{1+\beta^2}\right),$$

and where $\Gamma(.)$ denotes the gamma function, and ${}_pF_q$ denotes the generalized hypergeometric function.

It can be shown that, after some algebraic manipulations, (32)-(34) is identical to (27)-(29). For example, by rewriting (35) as

$$I_1(\alpha,\beta) = \frac{\sqrt{\pi}\Gamma(\alpha-\frac{1}{2})}{\Gamma(\alpha)}\left[1 - T_{1,2\alpha-1}(\beta\sqrt{2\alpha-1})\right],$$

we obtain (27) from (32). One reason for the preference of (26)-(29) is, perhaps, this forms appears more natural in the MST mixture context, and can be extended to the unrestricted case (see Section 5). Also, current routines for the calculation of $T_1(.)$ tends to be (computationally) more efficient than those for calculation of the hypergeometric functions. For the illustrations in Section 6, the expressions (26)-(29) are adopted when fitting FM-rMST models.

For the SNI-ST mixture model, very similar expressions for the conditional expectations (20) and (22) -(24) can be obtained, as presented in Cabral et al. (2012).

As pointed out in Cabral et al. (2012), the computation of $e_{1,hj}^{(k)}$ can be avoided if we consider the ECME extension of the EM-algorithm. In which case, the parameter representing the degrees of freedom is updated by maximizing the actual marginal log likelihood function on the M-step.

5 ML estimation for the unrestricted MST mixture model

Although it may appear that the extension of the restricted model to the unrestricted case is quite straightforward, the resulting estimation problem becomes quite complex. The first implementation of the EM algorithm for the ML estimation of finite mixture of the unrestricted MST distributions, as presented in Lin (2010), utilized Monte Carlo (MC) integration on the E-step to calculate the intractable conditional expectations. Later, Lee and McLachlan (2011) and Ho et al. (2012a) recognized that two of the these expectations can be expressed in terms of moments of a truncated multivariate t-random variate, for which closed-form expressions can be derived. Also, Lee and McLachlan (2011) applied a OSL approach to achieve a fast and simple closed-form implementation. This greatly reduced the computational burden associated with the alternative MC implementation.

We begin with a discussion of the estimation for a single uMST distribution in Section 5.1. The extension to the mixture case is presented in the next subsection.

5.1 ML estimation for the unrestricted MST distribution

To apply the EM algorithm, the observed data vector $\boldsymbol{y}_T = \left(\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T\right)^T$ is regarded as incomplete, and we introduce two latent variables denoted by $\boldsymbol{u}_T = \left(\boldsymbol{u}_1^T, \ldots, \boldsymbol{u}_n^T\right)^T$ and $\boldsymbol{w}_T = (w_1, \ldots, w_n)^T$, as defined by (16). We let $\boldsymbol{\theta}$ be the parameter containing the elements of the location parameter $\boldsymbol{\mu}$, the distinct elements of the scale matrix $\boldsymbol{\Sigma}$, the elements of the skew parameter $\boldsymbol{\delta}$, and the degrees of freedom ν. It follows that the complete-data log likelihood function for $\boldsymbol{\theta}$ is given by

$$\begin{aligned}
&\log L_c(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{u}, w) \\
&= K - \tfrac{1}{2}n\log|\boldsymbol{\Sigma}| - n\log\Gamma\left(\tfrac{1}{2}\nu\right) + \tfrac{1}{2}n\nu\log\left(\tfrac{1}{2}\nu\right) \\
&\quad - \sum_{j=1}^{n}\tfrac{1}{2}w_j\left[d\left(\boldsymbol{y}_j\right) + \left(\boldsymbol{u}_j - \boldsymbol{q}_j\right)^T\boldsymbol{\Lambda}^{-1}\left(\boldsymbol{u}_j - \boldsymbol{q}_j\right)\right] \\
&\quad + \left(\tfrac{1}{2}\nu + p - 1\right)\sum_{j=1}^{n}\log(w_j), \qquad (38)
\end{aligned}$$

where $\boldsymbol{q}_j = \boldsymbol{\Delta}\boldsymbol{\Omega}^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu})$, and K is a constant that does not depend on $\boldsymbol{\theta}$.

5.1.1 E-step

On the $(k+1)$th iteration, the E-step requires the calculation of the conditional expectation of the complete-data log likelihood given the observed data $\boldsymbol{y}_T$, using the current estimate $\boldsymbol{\theta}^{(k)}$ for $\boldsymbol{\theta}$. That is, we have to calculate the so-called Q-function defined by

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) = E_{\boldsymbol{\theta}^{(k)}}\left\{\log L_c(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{u}, w) \mid \boldsymbol{y}_T\right\}, \qquad (39)$$

where $E_{\boldsymbol{\theta}^{(k)}}$ denotes the expectation operator, using $\boldsymbol{\theta}^{(k)}$ for $\boldsymbol{\theta}$. This, in effect, requires the calculation of the conditional expectations

$$e_{1,j}^{(k)} = E_{\boldsymbol{\theta}^{(k)}}\left\{\log(W_j) \mid \boldsymbol{y}_j\right\}, \qquad (40)$$

$$e_{2,j}^{(k)} = E_{\boldsymbol{\theta}^{(k)}}\left\{W_j \mid \boldsymbol{y}_j\right\}, \qquad (41)$$

$$\boldsymbol{e}_{3,j}^{(k)} = E_{\boldsymbol{\theta}^{(k)}}\left\{W_j\boldsymbol{U}_j \mid \boldsymbol{y}_j\right\}, \qquad (42)$$

$$\boldsymbol{e}_{4,j}^{(k)} = E_{\boldsymbol{\theta}^{(k)}}\left\{W_j\boldsymbol{U}_j\boldsymbol{U}_j^T \mid \boldsymbol{y}_j\right\}. \qquad (43)$$

Note that the Q-function (39) does not admit a closed form expression for this problem, due to the conditional expectations $e_{1,j}^{(k)}$, $\boldsymbol{e}_{3,j}^{(k)}$, and $\boldsymbol{e}_{4,j}^{(k)}$ not being able to be evaluated in closed form. However, as recognized by Lee and McLachlan (2011) and Ho et al. (2012a), $\boldsymbol{e}_{3,j}^{(k)}$ and $\boldsymbol{e}_{4,j}^{(k)}$ can be expressed in terms of the first and second moment of a truncated t-variate, respectively. They in turn can be evaluated precisely and swiftly using the closed-form expressions presented in Lee and

McLachlan (2011) and Ho et al. (2012a), the former based on work of O'Hagan (1976); see Appendix A for further details on the calculation of the moments of the truncated multivariate t-distributions. Specifically, the two conditional expectations are given by

$$\boldsymbol{e}_{3j}^{(k)} = e_{2,j}^{(k)} E\left\{\boldsymbol{X}_j \mid \boldsymbol{y}_j\right\}, \tag{44}$$

$$\boldsymbol{e}_{4j}^{(k)} = e_{2,j}^{(k)} E\left\{\boldsymbol{X}_j \boldsymbol{X}_j^T \mid \boldsymbol{y}_j\right\}, \tag{45}$$

where $\boldsymbol{X}_j$ is a p-dimensional t-variate truncated to the positive hyperplane $\mathbb{R}^+$, which (conditional on $\boldsymbol{y}_j$) is distributed as

$$\boldsymbol{X}_j \mid \boldsymbol{y}_j \sim tt_{p,\nu^{(k)}+p+2}\left(\boldsymbol{q}_j^{(k)}, \left(\frac{\nu^{(k)}+d^{(k)}(\boldsymbol{y}_j)}{\nu^{(k)}+p+2}\right)\boldsymbol{\Lambda}^{(k)}; \mathbb{R}^+\right); \tag{46}$$

see Appendix B for further details on the derivation. We note that Ho et al. (2012a) obtained very similar expressions for $\boldsymbol{e}_{3,j}^{(k)}$ and $\boldsymbol{e}_{4,j}^{(k)}$, but their result corresponding to (44) and (45) is incorrect. Specifically, the random truncated t-variable involved in these two conditional expectations should have distribution (46) rather than the conditional distribution of $\boldsymbol{u}_j$ given $\boldsymbol{y}_j$ as reported in Ho et al. (2012a).

The integral $\boldsymbol{e}_{3,j}^{(k)}$ can be written as

$$\begin{aligned}
\boldsymbol{e}_{3,j}^{(k)} &= E\left\{W_j \boldsymbol{U}_j \mid \boldsymbol{y}_j\right\}, \\
&= E\left\{E\left\{W_j \boldsymbol{U}_j \mid w_j, \boldsymbol{y}_j\right\} \mid \boldsymbol{y}_j\right\}, && (47)\\
&= E\left\{W_j E\left\{\boldsymbol{U}_j \mid w_j, \boldsymbol{y}_j\right\} \mid \boldsymbol{y}_j\right\}, && (48)\\
&= E\left\{W_j \mid \boldsymbol{y}_j\right\} E\left\{\boldsymbol{X}_j \mid \boldsymbol{y}_j\right\}, && (49)
\end{aligned}$$

where (49) follows from (48) after some calculations; see equation (80) in Appendix B.2. The error in Ho et al. (2012a) occurs in going from (47) to (49). Using their results, $\boldsymbol{e}_{3,j}^{(k)}$ would be expressed as

$$\boldsymbol{e}_{3,j}^{(k)} = E\left\{W_j \mid \boldsymbol{y}_j\right\} E\left\{\boldsymbol{U}_j \mid \boldsymbol{y}_j\right\},$$

where

$$\boldsymbol{U}_j \mid \boldsymbol{y}_j \sim tt_{p,\nu^{(k)}+p}\left(\boldsymbol{q}_j^{(k)}, \sqrt{\frac{\nu^{(k)}+d^{(k)}(\boldsymbol{y}_j)}{\nu^{(k)}+p}}\boldsymbol{\Lambda}^{(k)}; \mathbb{R}^p\right). \tag{50}$$

Thus on comparing (46) and (50), it can be seen that the difference relates to the degrees of freedom $\nu+p$ in (50) being replaced by $\nu+p+2$ in (46). To demonstrate the effect this apparently small difference can have in practice, we have plotted the log likelihood function for a real data set to be analyzed in more detail in the Section 6. It can be seen that the result is in Ho et al. (2012a) for $\boldsymbol{e}_{3j}$ and $\boldsymbol{e}_{4j}$ would result in the EM failing to maintain monotonic convergence (Figure 1).

Also, there is a misprint in equation (20) of Lee and McLachlan (2011), where the square root term in the location parameter should appear as a coefficient of the scale parameter, as in their expression for the mixture case on page 12.

It can be easily shown that $e_{2,j}^{(k)}$ can be written in closed form as

$$e_{2,j}^{(k)} = \left(\frac{\nu^{(k)}+p}{\nu^{(k)}+d^{(k)}(\boldsymbol{y}_j)}\right)\frac{T_{p,\nu^{(k)}+p+2}\left(\boldsymbol{y}_{2j}^{(k)}; \boldsymbol{0}, \boldsymbol{\Lambda}^{(k)}\right)}{T_{p,\nu^{(k)}+p}\left(\boldsymbol{y}_{1j}^{(k)}; \boldsymbol{0}, \boldsymbol{\Lambda}^{(k)}\right)}. \tag{51}$$

The integral $e_{1,j}^{(k)}$ can be evaluated numerically in a straightforward manner, for example, using the `cubature` package in R. Also, if we were to use a one-step late (OSL) approach (Green, 1990) for the implementation of the EM algorithm here (see Lee and McLachlan (2011)), $e_{1,j}^{(k)}$ can be evaluated by a closed form expression given by

$$\begin{aligned}
e_{1,j}^{(k)} &= e_{2,j}^{(k)} - \log\left(\frac{\nu^{(k)}+d^{(k)}(\boldsymbol{y}_j)}{2}\right) \\
&\quad - \left(\frac{\nu^{(k)}+p}{\nu^{(k)}+d^{(k)}(\boldsymbol{y}_j)}\right) + \psi\left(\frac{\nu^{(k)}+p}{2}\right). && (52)
\end{aligned}$$

It is worth noting that the OSL expression for $e_{1,j}^{(k)}$ is equivalent to the exact expression, except for the term $S_{1,j}^{(k)}$ (see Appendix B), which is practically zero. We note that the observed difference between the two approaches is negligible in practice. In the examples to be presented, we used the OSL approach for the calculation of $e_{1,j}^{(k)}$.

5.1.2 M-step

On the $(k+1)$th iteration, the M-step consists of the maximization of the Q-function (39) with respect to $\boldsymbol{\theta}$. For easier computation, we employ the ECM extension of the EM algorithm, where the M-step is replaced by four conditional–maximization (CM)-steps, corresponding to the decomposition of $\boldsymbol{\theta}$ into four subvectors, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T, \boldsymbol{\theta}_3^T, \theta_4)^T$, where $\boldsymbol{\theta}_1 = \boldsymbol{\mu}$, $\boldsymbol{\theta}_2 = \boldsymbol{\delta}$, $\boldsymbol{\theta}_3$ is the vector containing the distinct elements of $\boldsymbol{\Sigma}$, and $\theta_4 = \nu$. To compute $\boldsymbol{\mu}^{(k+1)}$, we maximize $Q(\boldsymbol{\mu}, \boldsymbol{\theta}_2^{(k)}, \boldsymbol{\theta}_3^{(k)}, \theta_4^{(k)}; \boldsymbol{\theta}^{(k)})$ with respect to $\boldsymbol{\mu}$, and to compute $\boldsymbol{\delta}^{(k+1)}$, we first update $\boldsymbol{\mu}$ to $\boldsymbol{\mu}^{(k+1)}$ and then maximize $Q(\boldsymbol{\mu}^{(k+1)}, \boldsymbol{\delta}, \boldsymbol{\theta}_3^{(k)}, \theta_4^{(k)}; \boldsymbol{\theta}^{(k)})$ with respect to $\boldsymbol{\delta}$, and so on.

We let $\mathrm{diag}(\boldsymbol{A})$ denote the operator that produces a vector by extracting the diagonal elements of the matrix $\boldsymbol{A}$. Straightforward algebraic manipulations lead to the

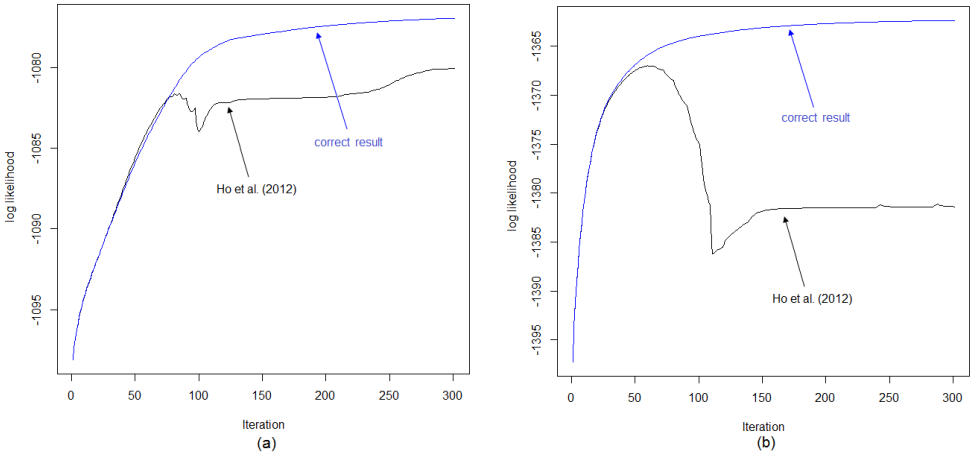

Fig. 1 Log likelihood function for fitting a two-component unrestricted skew t-mixture (FM-uMST) model to the AIS data. (a) Log likelihood curve for fitting FM-uMST on BMI and Bfat. Black line: using (50); blue line: using (46). (b) Results on the bivariate subset of LBM and Bfat. The result using (50) does not maintains monotonic convergence.

following closed form expressions for $\boldsymbol{\mu}^{(k+1)}$, $\boldsymbol{\Sigma}^{(k+1)}$, and $\boldsymbol{\delta}^{(k+1)}$,

$$\boldsymbol{\mu}^{(k+1)} = \frac{\sum_{j=1}^{n}\left[e_{2,j}^{(k)}\boldsymbol{y}_j - \boldsymbol{\Delta}^{(k)}\boldsymbol{e}_{3,j}^{(k)}\right]}{\sum_{j=1}^{n} e_{2,j}^{(k)}}, \tag{53}$$

$$\begin{aligned}\boldsymbol{\delta}^{(k+1)} = &\left(\boldsymbol{\Sigma}^{(k)^{-1}} \circ \sum_{j=1}^{n}\boldsymbol{e}_{4,j}^{(k)}\right)^{-1} \\ &\operatorname{diag}\left[\boldsymbol{\Sigma}^{(k)^{-1}}\sum_{j=1}^{n}\left(\boldsymbol{y}_j - \boldsymbol{\mu}^{(k+1)}\right)\boldsymbol{e}_{3,j}^{(k)^T}\right],\end{aligned} \tag{54}$$

$$\begin{aligned}\boldsymbol{\Sigma}^{(k+1)} = \frac{1}{n}\sum_{j=1}^{n}\Big[&\boldsymbol{\Delta}^{(k+1)}\boldsymbol{e}_{4,j}^{(k)^T}\boldsymbol{\Delta}^{(k+1)^T} \\ &- \left(\boldsymbol{y}_j - \boldsymbol{\mu}^{(k+1)}\right)\boldsymbol{e}_{3,j}^{(k)^T}\boldsymbol{\Delta}^{(k+1)} \\ &+ \left(\boldsymbol{y}_j - \boldsymbol{\mu}^{(k+1)}\right)\left(\boldsymbol{y}_j - \boldsymbol{\mu}^{(k+1)}\right)^T e_{2,j}^{(k)} \\ &-\boldsymbol{\Delta}^{(k+1)}\boldsymbol{e}_{3,j}^{(k)}\left(\boldsymbol{y}_j - \boldsymbol{\mu}^{(k+1)}\right)^T\Big],\end{aligned} \tag{55}$$

where $\boldsymbol{\Delta}^{(k+1)} = \operatorname{diag}\left(\boldsymbol{\delta}^{(k+1)}\right)$, and $\circ$ denotes the Hadamard or elementwise product. We note that the expression for $\boldsymbol{\delta}^{(k+1)}$ is incorrectly written in Lee and McLachlan (2011).

An updated estimate of the degrees of freedom $\nu^{(k+1)}$ is obtained by solving the equation

$$\log\left(\frac{\nu^{(k+1)}}{2}\right) - \psi\left(\frac{\nu^{(k+1)}}{2}\right) + 1 = \frac{1}{n}\sum_{j=1}^{n}\left(e_{2,j}^{(k)} - e_{1,j}^{(k)}\right), \tag{56}$$

where $\psi(\cdot)$ is the Digamma function.

Note that the computation of $e_{1,j}^{(k)}$ can be avoided if we adopt the ECME extension of the EM algorithm. The expectation-conditional maximization either (ECME) algorithm (Liu and Rubin, 1994) proceeds by replacing some of the CM-steps with CML-steps that conditionally maximize the actual log likelihood function. An exact EM-type algorithm for the fitting of mixtures of rMST distributions can be implemented by replacing (56) with the following CML-step:

$$\nu^{(k+1)} = \underset{\nu}{\operatorname{argmax}}\sum_{j=1}^{n}\log f(\boldsymbol{y}_j \mid \boldsymbol{\mu}^{(k+1)}, \boldsymbol{\Sigma}^{(k+1)}, \boldsymbol{\delta}^{(k+1)}, \nu). \tag{57}$$

The EM-type algorithm proceeds as follows on the $(k+1)$th iteration:

E-step: Given $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$, compute the four conditional expectations $e_{1,j}^{(k)}$, $e_{2,j}^{(k)}$, $\boldsymbol{e}_{3,j}^{(k)}$ and $\boldsymbol{e}_{4,j}^{(k)}$ by using (52), (51), (44), and (45), respectively, for $j = 1, \ldots, n$.

M-step: Update $\boldsymbol{\mu}^{(k+1)}$, $\boldsymbol{\delta}^{(k+1)}$, and $\boldsymbol{\Sigma}^{(k+1)}$ by using (53), (54), and (55). Calculate $\nu^{(k+1)}$ by solving (56) or (57).

5.1.3 Starting values

Initial parameter values can be specified based on the sample mean, sample covariance matrix, and sample skewness of the given data. A simple strategy is described in Lin (2010), and is reported here for completeness. Let $\bar{\boldsymbol{y}}$, $\boldsymbol{S}$, and $\boldsymbol{\gamma}$ denotes the sample mean,

sample covariance matrix, and sample skewness, respectively. Here, the ith element of the sample skewness $\boldsymbol{\gamma}$ is given by

$$\gamma_i = \frac{n^{-1}\sum_{j=1}^{n}(y_{ij}-\mu_i)^3}{\left(n^{-1}\sum_{j=1}^{n}(y_{ij}-\mu_i)^2\right)^{\frac{3}{2}}} \quad (i=1,\ldots,p),$$

where y_{ji} denotes the ith element of the jth observation, and μ_i is the ith element of $\boldsymbol{\mu}$. Then the parameters can be initialized using

$$\begin{aligned}
\boldsymbol{\Sigma}^{(0)} &= \boldsymbol{S} + (a-1)\,\mathrm{DIAG}\left(\mathrm{diag}(\boldsymbol{S})\right),\\
\boldsymbol{\delta}^{(0)} &= \mathrm{sign}(\boldsymbol{\gamma})\sqrt{\frac{\pi(1-a)}{\pi-2}}\boldsymbol{s}^{*},\\
\boldsymbol{\mu}^{(0)} &= \bar{\boldsymbol{y}} - \sqrt{\frac{2}{\pi}}\,\boldsymbol{\delta}^{(0)},\\
\nu^{(0)} &= 40,
\end{aligned}$$

where $\boldsymbol{s}^{*}$ is the vector created by taking the square root of the diagonal elements of $\boldsymbol{S}$, and a is a scalar constant between 0 and 1. As the EM algorithm is sensitive to the starting value, a poor choice of initial values may lead to convergence to a local maxima. It is thus highly recommended to apply a wide range of different initializations to avoid being trapped at a local maxima of the likelihood function; see, for example, Karlis and Xekalaki (2003) and O'Hagan et al. (2012) for discussions on these issues. In practice, a number of starting values for the above algorithm can be generated by varying a systematically across $(0,1)$. Then the set of parameters yielding the highest value of the likelihood is selected to start the EM algorithm.

5.2 The FM-uMST distribution

The probability density function (pdf) of a finite mixture of g multivariate skew t-components, using the notation above, is given by

$$f(\boldsymbol{y};\boldsymbol{\Psi}) = \sum_{h=1}^{g}\pi_h f_p(\boldsymbol{y};\boldsymbol{\mu}_h,\boldsymbol{\Sigma}_h,\boldsymbol{\delta}_h,\nu_h), \tag{58}$$

where $f_p(\boldsymbol{y};\boldsymbol{\mu}_h,\boldsymbol{\Sigma}_h,\boldsymbol{\delta}_h,\nu_h)$ denotes the hth MST component of the mixture model as defined by (15), with location parameter $\boldsymbol{\mu}_h$, scale matrix $\boldsymbol{\Sigma}_h$, skew parameter $\boldsymbol{\delta}_h$, and degrees of freedom ν_h. The mixing proportions π_h satisfy $\pi_h \geq 0$ $(h=1,\ldots,g)$ and $\sum_{h=1}^{g}\pi_h = 1$. We shall denote the model defined by (58) by FM-uMST (finite mixture of uMST) distributions. Let $\boldsymbol{\Psi}$ contain all the unknown parameters of the FM-uMST model; that is, $\boldsymbol{\Psi} = \left(\pi_1,\ldots,\pi_{g-1},\boldsymbol{\theta}_1^T,\ldots,\boldsymbol{\theta}_g^T\right)^T$ where now $\boldsymbol{\theta}_h$ consists of the unknown parameters of the hth component density function.

To formulate the estimation of the unknown parameters in the FM-uMST model as an incomplete-data problem in the EM framework, a set of latent component labels $\boldsymbol{z}_j = (z_{1j},\ldots,z_{gj})^T$ $(j=1,\ldots,n)$ is introduced, where again each element z_{hj} is a binary variable defined by (18). Hence, the random vector $\boldsymbol{Z}_j$ corresponding to $\boldsymbol{z}_j$ follows a multinomial distribution $\boldsymbol{Z}_j \sim \mathrm{Mult}_g(1;\pi_1,\ldots,\pi_g)$. It follows that the FM-uMST model can be represented in the hierarchical form given by

$$\begin{aligned}
\boldsymbol{Y}_j \mid \boldsymbol{u}_j, w_j, z_{hj}=1 &\sim N_p\left(\boldsymbol{\mu}_h + \boldsymbol{\Delta}_h\boldsymbol{u}_j, \frac{1}{w_j}\boldsymbol{\Sigma}_h\right),\\
\boldsymbol{U}_j \mid w_j, z_{hj}=1 &\sim HN_p\left(\boldsymbol{0}, \frac{1}{w_j}\boldsymbol{I}_p\right),\\
W_j \mid z_{hj}=1 &\sim \mathrm{gamma}\left(\frac{\nu_h}{2},\frac{\nu_h}{2}\right),\\
\boldsymbol{Z}_j &\sim \mathrm{Mult}_g(1,\boldsymbol{\pi}),
\end{aligned} \tag{59}$$

where $\boldsymbol{\Delta}_h = \mathrm{DIAG}(\boldsymbol{\delta}_h)$ and $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_g)^T$.

5.3 ML estimation for FM-uMST distributions

From the hierarchical characterization (59) of the FM-uMST distributions, the complete-data log likelihood function is given by

$$\log L_c(\boldsymbol{\Psi}) = \log L_{1c}(\boldsymbol{\Psi}) + \log L_{2c}(\boldsymbol{\Psi}) + \log L_{3c}(\boldsymbol{\Psi}), \tag{60}$$

where

$$\begin{aligned}
\log L_{1c}(\boldsymbol{\Psi}) &= \sum_{h=1}^{g}\sum_{j=1}^{n} z_{hj}\log(\pi_h),\\
\log L_{2c}(\boldsymbol{\Psi}) &= \sum_{h=1}^{g}\sum_{j=1}^{n} z_{hj}\left[\left(\frac{\nu_h}{2}\right)\log\left(\frac{\nu_h}{2}\right) - \log\Gamma\left(\frac{\nu_h}{2}\right)\right.\\
&\quad \left. + \left(\frac{\nu_h}{2}+p-1\right)\log(w_j) - \left(\frac{w_j}{2}\right)\nu_h\right],\\
\log L_{3c}(\boldsymbol{\Psi}) &= \sum_{h=1}^{g}\sum_{j=1}^{n} z_{hj}\left\{-p\log(2\pi) - \tfrac{1}{2}\log|\boldsymbol{\Sigma}_h|\right.\\
&\quad - \frac{w_j}{2}\left[(\boldsymbol{u}_j-\boldsymbol{q}_{hj})^T\boldsymbol{\Lambda}_h^{-1}(\boldsymbol{u}_j-\boldsymbol{q}_{hj})\right.\\
&\quad \left.\left. + d_h(\boldsymbol{y}_j)\right]\right\},
\end{aligned} \tag{61}$$

and where

$$\begin{aligned}
d_h(\boldsymbol{y}_j) &= (\boldsymbol{y}_j-\boldsymbol{\mu}_h)^T\boldsymbol{\Omega}_h^{-1}(\boldsymbol{y}_j-\boldsymbol{\mu}_h),\\
\boldsymbol{q}_{hj} &= \boldsymbol{\Delta}_h^T\boldsymbol{\Omega}_h^{-1}(\boldsymbol{y}_j-\boldsymbol{\mu}_h),\\
\boldsymbol{\Lambda}_h &= \boldsymbol{I}_p - \boldsymbol{\Delta}_h^T\boldsymbol{\Omega}_h^{-1}\boldsymbol{\Delta}_h,\\
\boldsymbol{\Omega}_h &= \boldsymbol{\Sigma}_h + \boldsymbol{\Delta}_h\boldsymbol{\Delta}_h^T.
\end{aligned}$$

The Q-function corresponding to the complete-data log likelihood (60) is given by

$$Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)}) = E_{\boldsymbol{\Psi}^{(k)}}\left\{\log L_c\left(\boldsymbol{\Psi}\right) \mid \boldsymbol{y}_T\right\},$$
$$= \sum_{l=1}^{3} Q_l(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)}), \quad (62)$$

where

$$Q_l(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)}) = E_{\boldsymbol{\Psi}^{(k)}}\left\{\log L_{lc}(\boldsymbol{\Psi}) \mid \boldsymbol{y}_T\right\} \quad (l = 1,\, 2,\, 3).$$

It is clear from (62) that maximization of the Q-function on each M-step only involves the maximization of each function $Q_l(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)})$ considered separately, ($l = 1,\, 2,\, 3$).

The necessary conditional expectations required in computing the functions $Q_l(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)})$ are, namely,

$$\begin{aligned}
\tau_{hj}^{(k)} &= E_{\boldsymbol{\Psi}^{(k)}}\{Z_{hj} \mid \boldsymbol{y}_j\},\\
e_{1,hj}^{(k)} &= E_{\boldsymbol{\Psi}^{(k)}}\{\log(W_j) \mid \boldsymbol{y}_j, z_{hj} = 1\},\\
e_{2,hj}^{(k)} &= E_{\boldsymbol{\Psi}^{(k)}}\{W_j \mid \boldsymbol{y}_j, z_{hj} = 1\},\\
\boldsymbol{e}_{3,hj}^{(k)} &= E_{\boldsymbol{\Psi}^{(k)}}\{W_j\boldsymbol{U}_j \mid \boldsymbol{y}_j, z_{hj} = 1\},\\
\boldsymbol{e}_{4,hj}^{(k)} &= E_{\boldsymbol{\Psi}^{(k)}}\{W_j\boldsymbol{U}_j\boldsymbol{U}_j^T \mid \boldsymbol{y}_j, z_{hj} = 1\}. \quad (63)
\end{aligned}$$

Again, the posterior probability of membership of the hth component by $\boldsymbol{y}_j$, using the current estimate $\boldsymbol{\Psi}^{(k)}$ for $\boldsymbol{\Psi}$, is given by (25), except that the density $f\left(\boldsymbol{y}_j;\boldsymbol{\mu}_h^{(k)},\boldsymbol{\Sigma}_h^{(k)},\boldsymbol{\delta}_h^{(k)},\nu_h^{(k)}\right)$ now refers to the unrestricted MST density.

The other four expectations have analogous expressions to their one-component counterpart given in Section 5.1; see Appendix Appendix B for further details.

The ECM algorithm is implemented as follows on the $(k+1)$th iteration:

E-step: Compute $\tau_{hj}^{(k)}$ using (25), and $e_{1,hj}^{(k)}$, $e_{2,hj}^{(k)}$, $\boldsymbol{e}_{3,hj}^{(k)}$, and $\boldsymbol{e}_{4,hj}^{(k)}$ as given by (83), (84), (85), and (86) respectively, for $h = 1, \ldots, g$ and $j = 1, \ldots, n$.

M-step: Update the estimate of $\boldsymbol{\Psi}$ by calculating the following for $h = 1, \ldots, g$,

$$\boldsymbol{\mu}_h^{(k)} = \frac{\sum_{j=1}^n \tau_{hj}^{(k)}\left[e_{2,hj}^{(k)}\boldsymbol{y}_j - \boldsymbol{\Delta}_h^{(k)}\boldsymbol{e}_{3,hj}^{(k)}\right]}{\sum_{j=1}^n \tau_{hj}^{(k)} e_{2,hj}^{(k)}},$$

$$\boldsymbol{\delta}_h^{(k+1)} = \left(\boldsymbol{\Sigma}_h^{(k)^{-1}} \circ \sum_{j=1}^n \tau_{hj}^{(k)}\boldsymbol{e}_{4hj}^{(k)}\right)^{-1}$$
$$\mathrm{DIAG}\left[\boldsymbol{\Sigma}_h^{(k)^{-1}}\sum_{j=1}^n \tau_{hj}^{(k)}\left(\boldsymbol{y}_j - \boldsymbol{\mu}_h^{(k+1)}\right)\boldsymbol{e}_{3,hj}^{(k)^T}\right],$$

and

$$\begin{aligned}
\boldsymbol{\Sigma}_h^{(k+1)} = \frac{1}{\sum_{j=1}^n \tau_{hj}^{(k)}}\sum_{j=1}^n \tau_{hj}^{(k)}\Big[&\boldsymbol{\Delta}_h^{(k+1)}\boldsymbol{e}_{4,hj}^{(k)^T}\boldsymbol{\Delta}_h^{(k+1)^T}\\
&- \left(\boldsymbol{y}_j - \boldsymbol{\mu}_h^{(k+1)}\right)\boldsymbol{e}_{3,hj}^{(k)^T}\boldsymbol{\Delta}_h^{(k+1)}\\
&- \boldsymbol{\Delta}_h^{(k+1)}\boldsymbol{e}_{3,hj}^{(k)}\left(\boldsymbol{y}_j - \boldsymbol{\mu}_h^{(k+1)}\right)^T\\
&+ \left(\boldsymbol{y}_j - \boldsymbol{\mu}_h^{(k+1)}\right)\left(\boldsymbol{y}_j - \boldsymbol{\mu}_h^{(k+1)}\right)^T e_{2,hj}^{(k)}\Big].
\end{aligned}$$

It should be pointed out that the approach can be extended in a straightforward manner for other structures on the component-covariance matrices than the general structure considered above. For example, when homoscedasticity is assumed, that is, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 = \ldots = \boldsymbol{\Sigma}_g$, the following expression is used to update $\boldsymbol{\Sigma}^{(k+1)}$:

$$\begin{aligned}
\boldsymbol{\Sigma}^{(k+1)} = \frac{1}{n}\sum_{j=1}^n\sum_{h=1}^g \tau_{hj}^{(k)}\Big[&\boldsymbol{\Delta}_h^{(k+1)}\boldsymbol{e}_{4hj}^{(k)^T}\boldsymbol{\Delta}_h^{(k+1)}\\
&- \left(\boldsymbol{y}_j - \boldsymbol{\mu}_h^{(k+1)}\right)\boldsymbol{e}_{3hj}^{(k)^T}\boldsymbol{\Delta}_h^{(k+1)}\\
&-\boldsymbol{\Delta}_h^{(k+1)}\boldsymbol{e}_{3hj}^{(k)}\left(\boldsymbol{y}_j - \boldsymbol{\mu}_h^{(k+1)}\right)^T\\
&+e_{2hj}^{(k)}\left(\boldsymbol{y}_j - \boldsymbol{\mu}_h^{(k+1)}\right)\left(\boldsymbol{y}_j - \boldsymbol{\mu}_h^{(k+1)}\right)^T\Big]. \quad (64)
\end{aligned}$$

In the case where $\boldsymbol{\Sigma}_h$ is restricted to be diagonal, that is, $\boldsymbol{\Sigma}_h = \mathrm{diag}\,(\boldsymbol{\sigma}_h)$, an update estimate of $\boldsymbol{\Sigma}_h$ is given by

$$\begin{aligned}
\boldsymbol{\sigma}_h^{(k+1)} = \frac{1}{\sum_{j=1}^n \tau_{hj}^{(k)}}\sum_{j=1}^n \tau_{hj}^{(k)}\Big[&\mathrm{DIAG}\left(\boldsymbol{\Delta}_h^{(k+1)}\boldsymbol{e}_{4hj}^{(k)}\boldsymbol{\Delta}_h^{(k+1)}\right)\\
&-2\left(\boldsymbol{y}_j - \boldsymbol{\mu}_h^{(k+1)}\right)\circ\left(\boldsymbol{\Delta}_h^{(k)}\boldsymbol{e}_{3hj}^{(k)}\right)\\
&+ e_{2hj}^{(k)}\left(\boldsymbol{y}_j - \boldsymbol{\mu}_h^{(k+1)}\right)\circ\left(\boldsymbol{y}_j - \boldsymbol{\mu}_h^{(k+1)}\right)\Big]. \quad (65)
\end{aligned}$$

For $\boldsymbol{\Sigma}_h$ restricted to the form $\boldsymbol{\Sigma}_h = \sigma_h^2\boldsymbol{I}_p$, (55) is replaced by the following expression:

$$\begin{aligned}
\sigma_h^2 = \frac{1}{\sum_{j=1}^n \tau_{hj}^{(k)}}\sum_{j=1}^n \tau_{hj}^{(k)}\Big[&tr\left(\boldsymbol{\Delta}_h^{(k+1)}\boldsymbol{e}_{4hj}^{(k)}\boldsymbol{\Delta}_h^{(k+1)}\right)\\
&-2\,\boldsymbol{e}_{3hj}^{(k)^T}\boldsymbol{\Delta}_h^{(k)}\left(\boldsymbol{y}_j - \boldsymbol{\mu}_h^{(k+1)}\right)\\
&+ e_{2hj}^{(k)}\left(\boldsymbol{y}_j - \boldsymbol{\mu}_h^{(k+1)}\right)^T\left(\boldsymbol{y}_j - \boldsymbol{\mu}_h^{(k+1)}\right)\Big]. \quad (66)
\end{aligned}$$

Finally, an update $\nu_h^{(k+1)}$ of the degrees of freedom is obtained by solving iteratively the equation

$$\log\left(\frac{\nu_h^{(k+1)}}{2}\right) - \psi\left(\frac{\nu_h^{(k+1)}}{2}\right) + 1$$
$$= \frac{\sum_{j=1}^n \tau_{hj}^{(k)}\left(e_{2,hj}^{(k)} - e_{1,hj}^{(k)}\right)}{\sum_{j=1}^n \tau_{hj}^{(k)}}.$$

When the degrees of freedom are assumed to be the same, that is, $\nu = \nu_1 = \ldots = \nu_g$, the updated estimate of ν is obtained by maximizing the actual marginal log likelihood function, as given by (67), namely

$$\nu^{(k+1)} = \underset{\nu}{\operatorname{argmax}} \sum_{j=1}^{n} \log \left[\sum_{h=1}^{g} \pi_h^{(k+1)} f\left(\boldsymbol{y}_j; \boldsymbol{\mu}_h^{(k+1)}, \boldsymbol{\Sigma}_h^{(k+1)}, \boldsymbol{\delta}_h^{(k+1)}, \nu\right)\right]. \quad (67)$$

As mentioned previously, this alternative provides an exact ECME implementation, and monotone convergence is guaranteed.

5.4 Initial values and stopping criteria

For starting values, the following procedure can be used. The component labels $\boldsymbol{z}_j^{(0)}$ $(j = 1, \ldots, n)$ can be initialized randomly, or according to some clustering algorithms such as k-means. Initial values for the other parameters can be set as follows:

$$\begin{aligned}
\pi_h^{(0)} &= \frac{1}{n}\sum_{j=1}^{n} z_{hj}^{(0)}, \\
\boldsymbol{\mu}_h^{(0)} &= \frac{\sum_{j=1^n} z_{hj}^{(0)} \boldsymbol{y}_j}{\sum_{j=1}^{n} z_{hj}^{(0)}}, \\
\boldsymbol{\Sigma}_h^{(0)} &= \frac{\sum_{j=1}^{n} z_{hj}^{(0)} \left(\boldsymbol{y}_j - \boldsymbol{\mu}_h^{(0)}\right)\left(\boldsymbol{y}_j - \boldsymbol{\mu}_h^{(0)}\right)^T}{\sum_{j=1}^{n} z_{hj}^{(0)}}, \\
\nu_h^{(0)} &= 4, \qquad (68)
\end{aligned}$$

and each element of $\boldsymbol{\delta}_h^{(0)}$ is initialized with ± 5 if $|\boldsymbol{\gamma}_h| > 0.1$, and zero otherwise, where the sign of each element depends on the sign of the associated element of $\boldsymbol{\gamma}_h$. Here, $\boldsymbol{\gamma}_h$ denotes the sample skewness for the hth component.

As the log likelihood function may exhibit multiple local maxima, it is useful to try several different starting values using various methods, and compare their relative log likelihood values. A convenient way to generate different initial values has been described in Section 5.1.3. In the examples to follow, we first select an initial clustering corresponding to the highest log likelihood value based on (68). Then the method described in Section 5.1.3 is employed to obtain a set of initial values for the model parameters.

Our stopping criterion terminates the algorithm when the relative difference in the log likelihood values between two successive iterations is less than the desired tolerance. For the examples in Section 6, the tolerance $\epsilon = 10^{-6}$ is used.

Table 1 BIC values of the fitted models with one, two, and three components for the reduced Lymphoma data.

Model	$g = 1$	$g = 2$	$g = 3$
FM-SNI-ST	2638.424	**2616.4**	2632.031
FM-rMST	2610.419	**2478.487**	2485.693
FM-uMST	2548.217	**2449.221**	2493.077

6 Illustrations

In this section, we fit three of the multivariate skew t-mixture models (namely FM-SNI-ST, FM-rMST, and FM-uMST) to real datasets to demonstrate their usefulness in analyzing and clustering multivariate skewed data. In the first example, we focus on the flexibility of the FM-uMST model in capturing the asymmetric shape of flow cytometric data. The next example illustrates the clustering capability of the unrestricted model in comparison with the restricted model. In the final example, we demonstrate the computational efficiency of the closed-form implementation of EM algorithm for the FM-uMST model.

6.1 Lymphoma data

We consider a subset of the T-cell phosphorylation data collected by Maier et al. (2007). In the original data, blood samples from 30 subjects were stained with four fluorophore-labeled antibodies against CD4, CD45RA, SLP76(pY128), and ZAP70(pY292) before and after an anti-CD3 stimulation. In this example, we focus on a reduced subset of the data in two variables – CD4 and ZAP70. This bivariate sample (Figure 2) appears to be bimodal and exhibits asymmetric pattern. We fit a two-component FM-uMST model to the data. More specifically, the fitted model can be written as

$$\begin{aligned}
f_2\left(\boldsymbol{y}_j; \boldsymbol{\Psi}\right) = \pi_1 f_2\left(\boldsymbol{y}_j; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\delta}_1, \nu_1\right) \\
+ (1-\pi_1) f_2\left(\boldsymbol{y}_j; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, \boldsymbol{\delta}_2, \nu_2\right),
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{\mu}_i &= (\mu_{i,1}, \mu_{i,2})^T, \\
\boldsymbol{\Sigma}_i &= \begin{pmatrix} \sigma_{i,11} & \sigma_{i,12} \\ \sigma_{i,12} & \sigma_{i,22} \end{pmatrix}, \\
\boldsymbol{\delta}_i &= (\delta_{i,1}, \delta_{i,2})^T \quad (i = 1, 2).
\end{aligned}$$

For comparison, we include the fitting of mixtures of skew t-distributions from the Skew-normal Independent (SNI) family (Lachos, Ghosh, and Arellano-Valle, 2010), hereafter named the FM-SNI-ST model, and mixtures of restricted skew t-distribution FM-rMST as described in Section 5.1. The estimated FM-SNI-ST density can be computed from the R package `mixsmsn` (Cabral, Lachos, and Prates, 2012), and the FM-rMST

model is implemented in the R package `emmix` (Wang, 2009). Note again that the FM-SNI-ST and FM-rMST models are different to the FM-uMST distribution, since the skewing function of the latter is not of dimension one. Moreover, under the FM-SNI-ST and FM-rMST settings, the correlation structure of $\boldsymbol{Y}$ will also be dependent on the skewness parameters, whereas for the FM-uMST distributions the covariance structure is not affected by $\boldsymbol{\delta}$.

The BIC values corresponding to the fitted models with one, two and three components are given in Table 1. For each of the three models, the solution with the smallest BIC value is chosen, that is, the two-component model. Initial values for the FM-uMST model are obtained from the procedure described in Section 5.4. Initialization for the FM-SNI-ST and FM-rMST models are according to the built-in procedure from the `mixsmsn` and `emmix` packages, respectively.

The contours of the fitted FM-SNI-ST, FM-rMST and FM-uMST component densities are depicted in Fig 2(b) to (d). To better visualize the shape of the fitted models, we display the estimated densities of each component instead of the mixture contours, and superimposed on the intensity plot of the dataset. It can be seen that the FM-uMST model provides a noticeably better fit than the other candidates. From a clustering point of view, the FM-uMST model also shows better performance as it is able to separate the two clusters correctly, whereas the restricted models tends to find it challenging. Moreover, the unrestricted model adapts to the asymmetric shape of each cluster more adequately. Although the fitted FM-rMST model shows an improvement from the FM-SNI-ST result, it did not capture the shape of each cluster as well as the FM-uMST solution. Most noticeably, the triangular shape of the second cluster (blue contours) from the FM-uMST model provides a close fit to the lower group of cells in the data. Thus the superiority of FM-uMST model is evident in dealing with asymmetric and heavily tailed data in this dataset.

6.2 GvHD data

Our second example concerns a dataset collected by Brinkman et al. (2007), where peripheral blood samples were collected weekly from patients following blood and bone marrow transplant. The original goal was to identify cellular signatures that can predict or assist in early detection of Graft versus Host Disease (GvHD), a common post-transplantation complication in which the recipient's bone marrow was attacked by the new donor material. Samples were stained with four fluorescence reagents: CD4 FITC, CD8β PE, CD3 PerCP, and CD8 APC. Hence we fit a 4-variate FM-uMST model to a case sample with a population of 13773 cells. The dataset is shown in the left panel of Figure 3, where cells are displayed in five different colours according to a manual expert clustering into five clusters. In addition, we include the results for the FM-SNI-ST model and the FM-rMST model.

We compare the performance of the three models FM-uMST, FM-SNI-ST, and FM-rMST in assigning cells to the expert clusters. Manual gating suggests there are five clusters in this case sample. Hence we applied the algorithm for the fitting of each model with g predefined as 5. For a fair comparison, we started the three algorithms using the same initial values. The initial clustering is based on k-means. The degrees of freedom are set to be identical for all components for the first iteration and assigned a relatively large value.

To assess the performance of these three algorithms, we take the manual expert clustering as being the 'true' class membership and we calculated the error rate of classification against this benchmark result with dead cells removed, measured by choosing among the possible permutations of class labels the one that gives the best value.

A summary of the results are listed in Table 2. Also reported there are the values of the log likelihood, the Akaike information criterion (AIC) (Akaike, 1974) and the Bayesian information criterion (BIC) (Schwarz, 1978) defined by

$$\text{AIC} = 2m - 2\log L \text{ and } \text{BIC} = m\log n - 2\log L,$$

respectively, where $\log L$ is the value of log likelihood value, m is the number of free parameters, and n is the sample size. Models with smaller AIC and BIC values are preferred when comparing different fitted results. The right panel of Figure 3 shows the classification results of FM-uMST against expert clustering, where incorrect allocations of cluster is indicated by the red dots.

As anticipated, a better clustering result was given by the FM-uMST model. It achieved the lowest misclassification rate and AIC and BIC values. The FM-rMST model has a disappointing performance in terms of clustering for this dataset, having the highest number of misallocations. Although the AIC and BIC values favour the FM-rMST model than the FM-SNI-ST in the restricted case, the latter gave a lower misallocation rate. It is evident that these two restricted models have inferior performance. This reveals some evidence of the extra flexibility offered by the more general FM-uMST model.

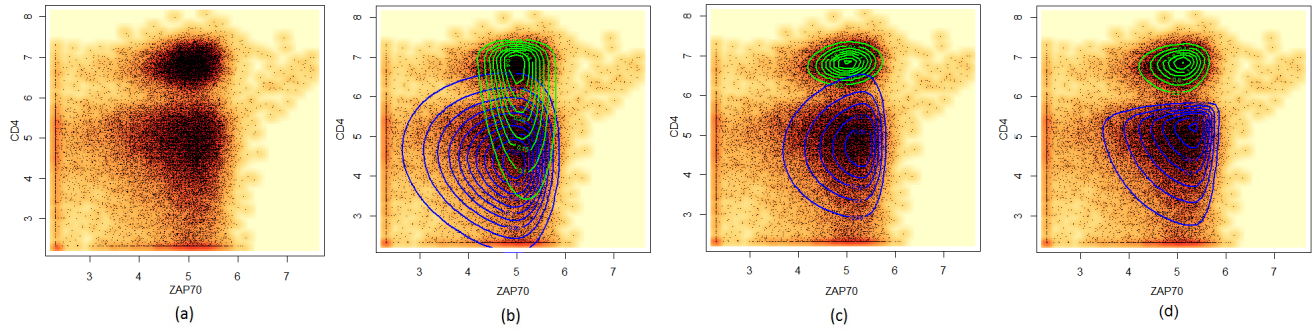

Fig. 2 *Mixture modelling of a reduced subset of prephosphorylation T cell population. Bivariate skew t-mixtures were fitted to a subset of the data and restricted in two dimensions CD45 and ZAP70. (a) Hue intensity plot of the Lymphoma dataset; (b) the contours of the component densities in the fitted two-component skew t-mixture model FM-SNI-ST using the R package* `mixsmsn`*; (c) the component density contours of the fitted two-component restricted skew t-mixture model FM-rMST using the R package* `emmix`*; (d) the fitted component contours of the two-component FM-uMST model.*

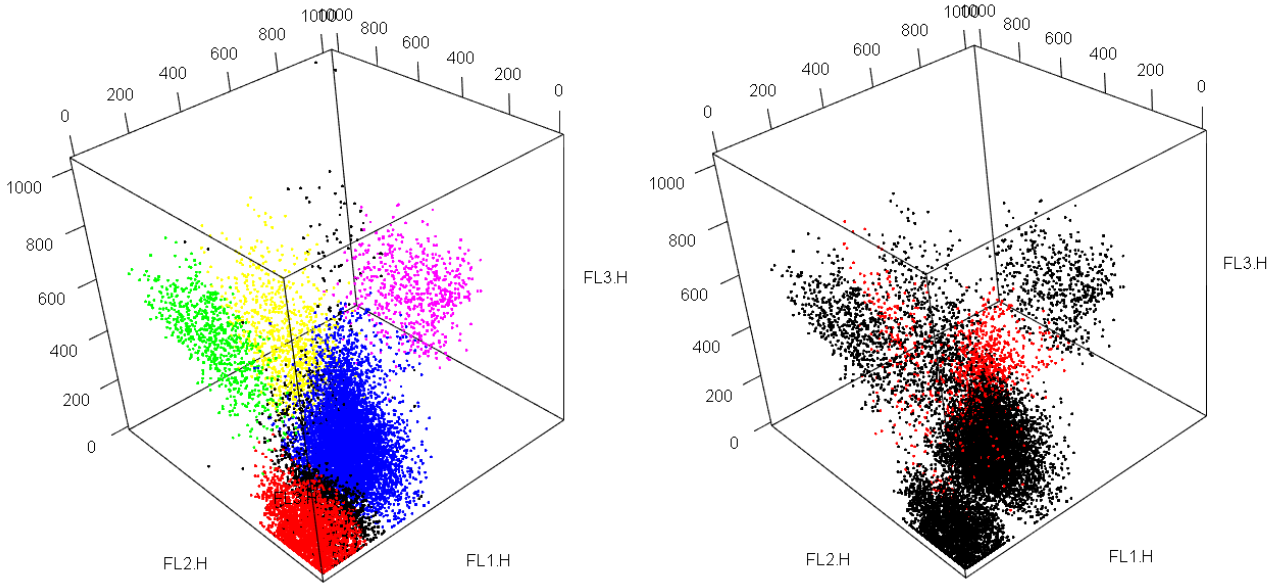

Fig. 3 GvHD dataset: *Clustering of a population of 13773 cells stained with four fluorescence reagents – CD4 FITC (FL1-H), CD8β PE (FL2.H), CD3 PerCP (FL3.H), and CD8 APC (FL4.H). Left panel: manual expert clustering of the GvHD data into five groups; Right panel: clustering result of FM-uMST, where black dots represents correct assignment of cluster labels and red dots indicates incorrect classification.*

Table 2 Misclassification rates for various multivariate skew t mixture models on the GvHD dataset.

Model	Misclassification rate	log L	AIC	BIC
FM-uMST	**0.0875**	-335561.7	671321.4	672006.9
FM-SNI-ST	0.1308	-339823.7	679837.5	680552.9
FM-rMST	0.2070	-336595.1	673388.2	674133.8

6.3 AIS data

We now illustrate the computational efficiency of the closed-form implementation of the EM algorithm as described in Section 5. We denote this version of the EM algorithm with an exact E-step as EM-exact. In addition, we consider the EM alternative with a Monte Carlo (MC) E-step as given by Lin (2010), which is denoted by EM-MC. Since both models are based on the same characterization of the multivariate skew t-distribution defined by Sahu et al. (2003), it is appropriate to compare their computation time directly. We assess their time performance on the well-analyzed Australian Institute of Sport (AIS) data, which consists of $p = 13$ measurements made on $n = 202$ athletes. We limit this illustration to a bivariate subset of two variables – Height (Ht) and the percentage of body fat (Bfat). These data are apparently bimodal with asymmetric pattern (Figure 4(a)); hence a two-component mixture model is fitted to the dataset.

A summary of the results is listed in Table 3. The contours of the fitted mixture density is depicted in Figure 4(b), where misallocations are indicted by black crosses. For this illustration, the EM-MC E-step is undertaken with 50 random draws for iterations $1-19$, 100 draws for iterations $20-39$, and 5000 draws for subsequent iterations, as recommended by Lin (2010). The same starting values and stopping rule were applied to both EM-exact and EM-MC.

The gender of each individual in this dataset is recorded, thus enabling us to evaluate the misclassification rate of the binary classification for the two methods. Not surprisingly, EM-exact obtained a lower misclassification rate. Not only did it achieve a higher log likelihood value, the computation time is remarkably lower than its competitor.

Figure 4(c) shows a plot of the likelihood values for the first 85 iterations, where the dashed curve represents the EM-MC result and the solid line indicates the EM-exact result. It can observed that the EM-exact curve is smooth and maintains stable monotonic convergence. On the other hand, the EM-MC curve tends to be slightly jagged, although it roughly resembles the likelihood curve given by EM-exact. This is a typical phenomenon for EM-MC in our experiments, as there is no guarantee that the log likelihood increases at each iteration.

7 Computational time and accuracy for the FM-uMST model

We now proceed to two informative experiments for evaluating the computational cost and accuracy of using the EM-exact and EM-MC algorithms on higher dimensional data. As pointed out previously, the main computational cost for EM-exact is evaluating the multivariate t-distribution function. Calculation of the first two moments of a p-variate truncated t-distribution requires the evaluation of two $T_{p,\nu}(\cdot)$ functions, p evaluations of $T_{p-1,\nu}(\cdot)$, and $\frac{1}{2}p(p-1)$ evaluations of $T_{p-2,\nu}(\cdot)$. Hence, the computation time will increase substantially with the number of dimensions. However, with the EM-exact algorithm, accuracy can be compromised for time.

We sampled 100 data points randomly from a Brain Tumour dataset supplied by Geoff Osborne from the Queensland Brain Institute at the University of Queensland. In both experiments we varied the dimension p of the sample. The graph in Figure 5(a) shows the typical CPU time per each E-step iteration for various dimensions p of the data; EM-MC(m) represents the EM-MC algorithm with m random draws using the Gibbs sampling approach described in Lin (2010). It is worth noting that in both experiments EM-exact is evaluated with a default tolerance of at least 10^{-6}. As seen in Figure 5(a), EM-exact is the fastest among the four versions of the E-step for low dimensions. For example, at $p = 2$, EM-exact is at least 25 times faster than EM-MC(50). It is important to note that although EM-MC(50) is slightly faster than EM-exact at higher dimensions, EM-exact produces results to a significantly higher accuracy (see Figure 5(b)), while EM-MC requires a large number of draws to achieve comparable results. We note that in our simulations, for example, at $p = 7$, 50 draws are insufficient to achieve acceptable estimates. Preliminary results suggests that at least 500 draws is required to generate reasonable approximations when p is greater than 6. In this case, EM-exact is at least 10 times quicker. Furthermore, EM-exact also has an additional advantage over the EM-MC alternative in that its results are deterministic.

To compare the accuracy of the EM-exact and EM-MC algorithms, we compute the total absolute error against a baseline result with minimum tolerance of 10^{-18}. Here, the total absolute error refers to the sum of absolute difference between the estimates and baseline result for all the conditional expectations involved in the E-step. Specifically, this is calculated by $\sum_{r=1}^{4} |e_{r,hj} - \tilde{e}_{r,hj}|$, where $e_{r,hj}$ denotes the baseline result and $\tilde{e}_{r,hj}$ represents the corresponding estimated value. For each of the EM-MC(m) algorithms, the average total absolute error of 100 trials is used. For EM-exact, the default tolerance is set to 10^{-6}. The results are shown in Figure 5(b). Not surprisingly, the absolute error of the EM-MC algorithm is significantly higher than that of the EM-exact algorithm. It can be observed that the absolute error is very high even for EM-

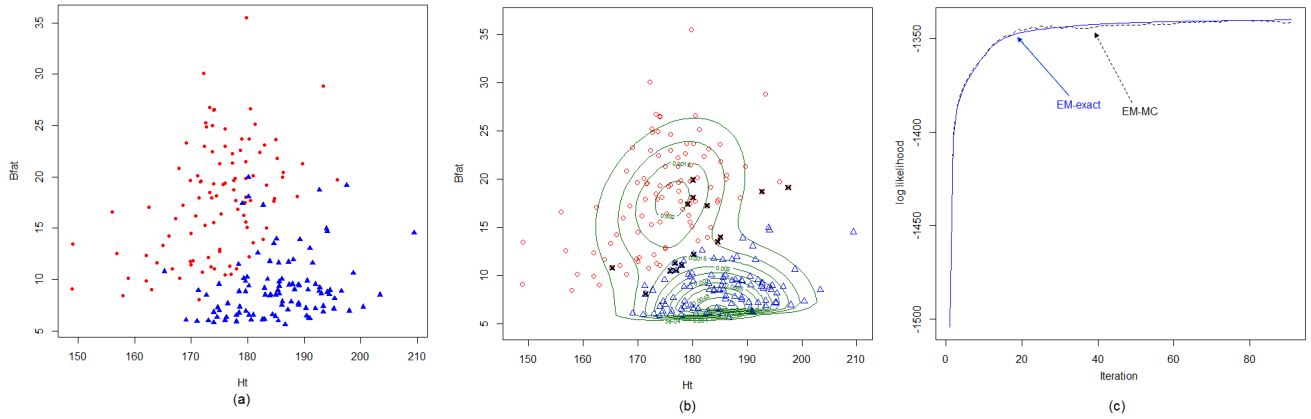

Fig. 4 AIS dataset: *Contour plots of fitted two-component FM-uMST model on Bfat and Ht. (a) Scatter plot of Bfat and Ht in two colours, red dots for female and blue triangles for male; (b) the fitted mixture contour of FM-uMST, where black crosses indicate misclassified observation. (c) Log likelihood curve for the AIS dataset given by EM-MC (dashed line) and EM-exact (solid line).*

Table 3 Computation time and misclassification rates for two different implementations of the EM algorithm for the multivariate skew t mixture models on the AIS dataset. For EM-exact, the E-step is implemented exactly as described in Section 6. As an alternative, the EM algorithm was implemented with a Monte Carlo E-step, EM-MC, as in (Lin, 2010). Time refers to the average CPU time per iteration, measured in seconds.

Model	EM-exact		EM-MC	
Component	1	2	1	2
π	0.53	0.47	0.54	0.45
μ_{i1}	179.11	182.04	178.04	182.94.
μ_{i2}	19.10	5.94	17.22	6.04
$\Sigma_{i,11}$	59.46	59.79	68.57	58.83
$\Sigma_{i,12}$	12.97	2.09	13.51	2.56
$\Sigma_{i,22}$	25.04	0.12	27.25	0.24
δ_{i1}	-3.90	3.42	-2.58	2.74
δ_{i2}	-0.23	3.28	0.49	2.97
ν	15.40	21.14	30.41	20.89
$L(\boldsymbol{\Psi})$	**-1340.95**		-1342.22	
misclassification rate	**0.0743**		0.0842	
average time per iteration	**0.71**		476.46	
total time	46.15		30970	

MC(500). At $p = 10$, for example, EM-exact is at least 30000 times more accurate and takes less than half the time required for EM-MC(500).

It is important to emphasize that as the dimension p of the data increases, EM-MC requires considerably more draws to provide a comparable (and acceptable) level of accuracy as EM-exact, which can be computationally intensive. Hence we advocate the use of EM-exact for fitting FM-uMST, which greatly improves the speed and accuracy of parameter estimation, especially for applications involving high dimensional data.

Finally, although a comparison of the computation time between the FM-uMST and FM-rMST model is not of primary interest here, we remark that the FM-rMST model requires significantly less computation time, primarily due to the simpler expressions that need to be calculated. ML estimation for the unrestricted model requires multiple evaluations of the multivariate t-distribution function, while the restricted model requires only evaluations of the univariate t-distribution function, regardless of the dimension of the observed data.

8 Concluding Remarks

The multivariate skew t-mixture model has emerged as a flexible and robust alternative to the skew-normal mixture model. This paper provides an up-to-date overview of recent developments in mixtures of multivariate skew t-distributions. We provide descriptions on various characterizations of the MST distribution used in different proposals of MST mixture models, and examine several existing EM algorithms for evaluating the parameters of the restricted and unrestricted multivariate skew t-mixture models. The uMST model has a

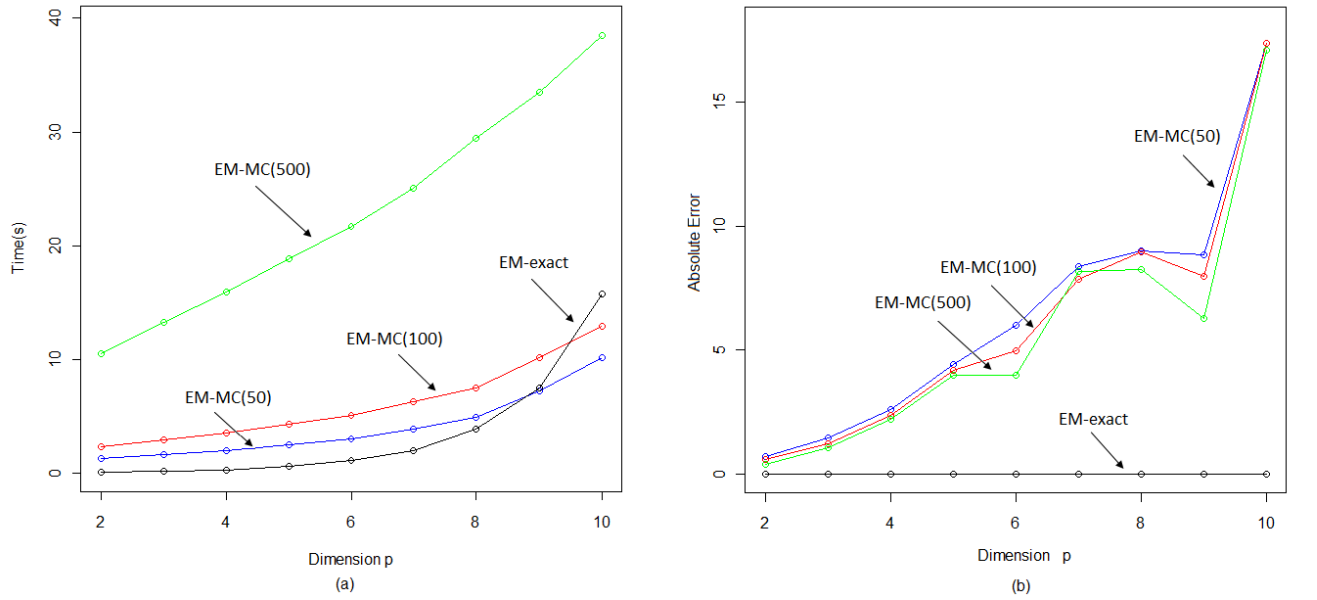

Fig. 5 *Comparison of performance of the EM-MC and EM-exact methods on a subset of 100 samples from the Brain Tumour data. Green line: EM-MC with 500 draws, red line: EM-MC with 100 draws, blue line: EM-MC with 50 draws, black line: EM-exact. (a) Typical computation time for E-step on a sample of 100 data in various dimensions. (b) Typical total absolute error of E-step per data point.*

more general characterization than various alternative 'restricted' versions of the skew t-distribution available in the literature and hence offers greater flexibility in capturing the asymmetric shape of skewed data, which can benefit applications in various scientific fields.

Examples on several real datasets shows that the unrestricted model is capable of achieving better clustering than the restricted models. Furthermore, the recently proposed closed-form fitting algorithms for the unrestricted model have been demonstrated in several real datasets to have a marked advantage over the EM algorithm with a Monte Carlo E-step. To achieve comparable accuracy to that of the EM algorithm with the E-step implemented using the above approach, the version of the algorithm with a Monte Carlo E-step would require a large number of draws, which would be computationally expensive. While significant advancement on the ML parameter estimation for the unrestricted model have been made in two recent papers, further development aimed at reducing the computational cost would be necessary to make it a fully viable tool in analyzing high dimensional data.

Appendix A The truncated multivariate t-distribution

In this appendix, we briefly describe the truncated multivariate t-distribution and provide some formulas for computing its moments (Lee and McLachlan, 2011). These expressions are crucial for the swift evaluation of the conditional expectations on the E-step of the FM-uMST model discussed in Section 5. We follow the approach of Lee and McLachlan (2011). A alternative description is given by Ho et al. (2012a), which provides equivalent expressions for the doubly truncated case.

Let $\boldsymbol{X}$ be a p-dimensional random variable having a multivariate t-distribution with location vector $\boldsymbol{\mu}$, scale matrix $\boldsymbol{\Sigma}$, and ν degrees of freedom. Truncating $\boldsymbol{x}$ to the hyperplane region $\mathbb{A} = \{\boldsymbol{x} \geq \boldsymbol{a},\ \boldsymbol{a} \in \mathbb{R}^p\}$, where $\boldsymbol{x} \geq \boldsymbol{a}$ means each element $x_i = (\boldsymbol{x})_i$ is greater than or equal to $a_i = (\boldsymbol{a})_i$ for $i = 1, \ldots, p$, results in a left-truncated t-distribution whose density is given by

$$f_{\mathbb{A}}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = T_{p,\nu}^{-1}(\boldsymbol{a}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\, t_{p,\nu}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{x} \in \mathbb{A}. \tag{69}$$

For a random vector $\boldsymbol{X}$ with density (69), we write $\boldsymbol{X} \sim tt_{p,\nu}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbb{A})$. For our purposes, we will be concerned with the first two moments of $\boldsymbol{X}$, specifically $E(\boldsymbol{X})$ and $E(\boldsymbol{X}\boldsymbol{X}^T)$. Explicit formulas for the truncated central t-distribution in the univariate case $tt_{1,\nu}(0, \sigma^2; \mathbb{A})$ were provided by O'Hagan (1973), who expressed the moments in terms of the non-truncated t-distribution. The multivariate case was studied in O'Hagan (1976), but still considering the central case only. Here we describe a generalization of the results in O'Hagan (1976) to the multivariate non-central case and express them in a form suitable for undertaking the E-step in the direct application of the EM algorithm to the fitting of mixtures of MST distributions.

Before presenting the expressions, it will be convenient to introduce some notation. Let $\boldsymbol{x}$ be a vector, then

x_i denotes the ith element,
$\boldsymbol{x}_{ij}$ is a two-dimensional vector with elements x_i and x_j,
$\boldsymbol{x}_{-i}$ represents the $(p-1)$-dimensional vector with the ith element removed, and

$\boldsymbol{x}_{-ij}$ represents the $(p-2)$-dimensional vector with the ith and jth elements removed.

For a matrix $\boldsymbol{X}$, let

x_{ij} denote the ijth element,
$\boldsymbol{X}_{ij}$ defines the 2×2 matrix consisting of the elements x_{ii}, x_{ij}, x_{ji} and x_{jj},
$\boldsymbol{X}_{-i}$ be created by removing the ith row and column from $\boldsymbol{X}$,
$\boldsymbol{X}_{-ij}$ be the $(p-2)$ square matrix resulting from the removal of the ith and jth row and column from $\boldsymbol{X}$, and
$\boldsymbol{X}_{(ij)}$ be the ith and jth column of $\boldsymbol{X}$ with the elements of $\boldsymbol{X}_{ij}$ removed, yielding a $(p-2) \times 2$ matrix.

We now proceed to the expressions for the first two moments of $\boldsymbol{X}$.

One can show that the first moment of (69) is

$$E(\boldsymbol{X}) = \boldsymbol{\mu} + \boldsymbol{\epsilon}, \tag{70}$$

where $\boldsymbol{\epsilon} = c^{-1}\boldsymbol{\Sigma}\boldsymbol{\xi}$ and $c = T_{p,\nu}(\boldsymbol{\mu} - \boldsymbol{a}; \boldsymbol{0}, \boldsymbol{\Sigma})$, and $\boldsymbol{\xi}$ is a $p \times 1$ vector with elements

$$\xi_i = (2\pi\sigma_{ii})^{-\frac{1}{2}} \left(\frac{\nu}{\nu + \sigma_{ii}^{-1}(\mu_i - a_i)^2}\right)^{\left(\frac{\nu-1}{2}\right)} \sqrt{\frac{\nu}{2}} \frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} T_{p,\nu-1}(a^*; \boldsymbol{0}, \boldsymbol{\Sigma}^*),$$

for $i = 1, \ldots, p$, and where

$$\boldsymbol{a}^* = (\boldsymbol{\mu}_{-i} - \boldsymbol{a}_{-i}) - (\boldsymbol{\mu}_{-i} - \boldsymbol{a}_{-i})\sigma_{ii}^{-1}, \boldsymbol{\Sigma}_{(i)}$$
$$\boldsymbol{\Sigma}^* = \left(\frac{\nu + \sigma_{ii}^{-1}(\mu_i - a_i)^2}{\nu - 1}\right)\left(\boldsymbol{\Sigma}_{-i} - \frac{1}{\sigma_{ii}}\boldsymbol{\Sigma}_{(i)}\boldsymbol{\Sigma}_{(i)}^T\right).$$

The second moment is given by

$$\begin{aligned} &E(\boldsymbol{X}\boldsymbol{X}^T) \\ &= \boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\mu}\boldsymbol{\epsilon}^T + \boldsymbol{\epsilon}\boldsymbol{\mu}^T - c^{-1}\boldsymbol{\Sigma}\boldsymbol{H}\boldsymbol{\Sigma} \\ &\quad + c^{-1}\left(\tfrac{\nu}{\nu-2}\right) T_{p,\nu-2}\left(\boldsymbol{\mu} - \boldsymbol{a}; \boldsymbol{0}, \left(\tfrac{\nu}{\nu-2}\right)\boldsymbol{\Sigma}\right)\boldsymbol{\Sigma}, \end{aligned} \tag{71}$$

where $\boldsymbol{H}$ is a $p \times p$ matrix with off-diagonal elements

$$h_{ij} = \frac{1}{2\pi\sqrt{\sigma_{ii}\sigma_{jj} - \sigma_{ij}^2}} \left(\frac{\nu}{\nu-2}\right)\left(\frac{\nu}{\nu^*}\right)^{\frac{\nu}{2}-1} T_{p-2,\nu-2}(\boldsymbol{a}^{**}; \boldsymbol{0}, \boldsymbol{\Sigma}^{**}), \quad i \neq j,$$

and diagonal elements,

$$\begin{aligned} h_{ii} &= \sigma_{ii}^{-1}(\mu_i - a_i)\xi_i - \sigma_{ii}^{-1}\sum_{j \neq i}\sigma_{ij}h_{ij}, \\ \nu^* &= \nu + (\boldsymbol{\mu}_{ij} - \boldsymbol{a}_{ij})^T \boldsymbol{\Sigma}_{ij}^{-1}(\boldsymbol{\mu}_{ij} - \boldsymbol{a}_{ij}), \\ \boldsymbol{a}^{**} &= (\boldsymbol{\mu}_{-ij} - \boldsymbol{a}_{-ij}) - \boldsymbol{\Sigma}_{(ij)}\boldsymbol{\Sigma}_{ij}^{-1}(\boldsymbol{\mu}_{ij} - \boldsymbol{a}_{ij}), \\ \boldsymbol{\Sigma}^{**} &= \frac{\nu^*}{\nu - 2}\left(\boldsymbol{\Sigma}_{-ij} - \boldsymbol{\Sigma}_{(ij)}\boldsymbol{\Sigma}_{ij}^{-1}\boldsymbol{\Sigma}_{(ij)}^T\right). \end{aligned}$$

It is worth noting that evaluation of the expressions (70) and (71) rely on algorithms for computing the multivariate central t-distribution function for which highly efficient procedures are readily available in many statistical packages. For example, an implementation of Genz's algorithm (Genz and Bretz, 2002, Kotz and Nadarajah, 2004) is provided by the `mvtnorm` package available from the R website.

Appendix B E-step for uMST

Derivations of $e_{1,j}^{(k)}$, $\boldsymbol{e}_{3,j}^{(k)}$ and $\boldsymbol{e}_{4,j}^{(k)}$ are detailed as follows.

Appendix B.1 Calculation of $e_{1,j}^{(k)}$

Concerning the calculation of the expectation $e_{1,j}^{(k)}$, the conditional density of W_j given $\boldsymbol{y}_j$, is given by

$$\begin{aligned} &f(w_j \mid \boldsymbol{y}_j) \\ &= \frac{\Gamma\left(w_j; \frac{\nu^{(k)}+p}{2}, \frac{\nu^{(k)}+d^{(k)}(\boldsymbol{y}_j)}{2}\right)\Phi_p\left(\boldsymbol{q}_j^{(k)}\sqrt{w_j}; \boldsymbol{0}, \boldsymbol{\Lambda}^{(k)}\right)}{T_{p,\nu^{(k)}+p}\left(\boldsymbol{y}_{1j}^{(k)}; \boldsymbol{0}, \boldsymbol{\Lambda}^{(k)}\right)}, \end{aligned} \tag{72}$$

where

$$\begin{aligned} \boldsymbol{y}_{1j}^{(k)} &= \boldsymbol{q}_j^{(k)}\sqrt{\frac{\nu^{(k)}+p}{\nu^{(k)}+d^{(k)}(\boldsymbol{y}_j)}}, \\ \boldsymbol{q}_j^{(k)} &= \boldsymbol{\Delta}^{(k)^T}\boldsymbol{\Omega}^{(k)^{-1}}\left(\boldsymbol{y}_j - \boldsymbol{\mu}^{(k)}\right), \\ d^{(k)}(\boldsymbol{y}_j) &= \left(\boldsymbol{y}_j - \boldsymbol{\mu}^{(k)}\right)^T\boldsymbol{\Omega}^{(k)^{-1}}\left(\boldsymbol{y}_j - \boldsymbol{\mu}^{(k)}\right), \end{aligned}$$

and $\boldsymbol{0}$ is the zero vector of appropriate dimension.

The conditional expectation $E_{\boldsymbol{\theta}^{(k)}}\{\log(W_j) \mid \boldsymbol{y}_j\}$ can be reduced to

$$\begin{aligned} e_{1,j}^{(k)} &= \left(\frac{\nu^{(k)}+p}{\nu^{(k)}+d^{(k)}(\boldsymbol{y}_j)}\right) \frac{T_{p,\nu^{(k)}+p+2}\left(\boldsymbol{y}_{2j}^{(k)}; \boldsymbol{0}, \boldsymbol{\Lambda}^{(k)}\right)}{T_{p,\nu^{(k)}+p}\left(\boldsymbol{y}_{1j}^{(k)}; \boldsymbol{0}, \boldsymbol{\Lambda}^{(k)}\right)} \\ &\quad - \log\left(\frac{\nu^{(k)}+d^{(k)}(\boldsymbol{y}_j)}{2}\right) - \left(\frac{\nu^{(k)}+p}{\nu^{(k)}+d^{(k)}(\boldsymbol{y}_j)}\right) \\ &\quad + \psi\left(\frac{\nu^{(k)}+p}{2}\right) + S_j^{(k)}, \end{aligned} \tag{73}$$

where

$$\boldsymbol{y}_{2j}^{(k)} = \boldsymbol{q}_j^{(k)}\sqrt{\frac{\nu^{(k)}+p+2}{\nu^{(k)}+d^{(k)}(\boldsymbol{y}_j)}},$$

and where the last term S is given by

$$\begin{aligned} S_j^{(k)} &= \psi\left(\frac{\nu^{(k)}}{2} + p\right) - \psi\left(\frac{\nu^{(k)}+p}{2}\right) + \left(\frac{\nu^{(k)}+p}{\nu^{(k)}+d^{(k)}(\boldsymbol{y}_j)}\right) \\ &\quad - \left(\frac{\nu^{(k)}+p}{\nu^{(k)}+d^{(k)}(\boldsymbol{y}_j)}\right)\frac{T_{p,\nu^{(k)}+p+2}\left(\boldsymbol{y}_{2j}^{(k)}; \boldsymbol{0}, \boldsymbol{\Lambda}^{(k)}\right)}{T_{p,\nu^{(k)}+p}\left(\boldsymbol{y}_{1j}^{(k)}; \boldsymbol{0}, \boldsymbol{\Lambda}^{(k)}\right)} \\ &\quad - \frac{\left[\pi\left(\nu^{(k)}+p\right)\right]^{-\frac{p}{2}}|\boldsymbol{\Lambda}|^{-\frac{1}{2}}}{T_{p,\nu^{(k)}+p}\left(\boldsymbol{y}_{1j}^{(k)}; \boldsymbol{0}, \boldsymbol{\Lambda}^{(k)}\right)}\frac{\Gamma\left(\frac{\nu^{(k)}}{2}+p\right)}{\Gamma\left(\frac{\nu^{(k)}+p}{2}\right)}S_{1,j}^{(k)}, \end{aligned} \tag{74}$$

and $S_{1,j}^{(k)}$ is an integral given by

$$\begin{aligned} S_{1,j}^{(k)} &= \int_{-\infty}^{\left[\boldsymbol{q}_j^{(k)}\right]_1}\int_{-\infty}^{\left[\boldsymbol{q}_j^{(k)}\right]_2}\cdots\int_{-\infty}^{\left[\boldsymbol{q}_j^{(k)}\right]_p} log\left(1 + \frac{\boldsymbol{s}^T\boldsymbol{\Lambda}^{(k)^{-1}}\boldsymbol{s}}{\nu^{(k)}+d^{(k)}(\boldsymbol{y}_j)}\right) \\ &\quad \left[1 + \frac{\boldsymbol{s}^T\boldsymbol{\Lambda}^{(k)^{-1}}\boldsymbol{s}}{\nu^{(k)}+d^{(k)}(\boldsymbol{y}_j)}\right]^{-\left(\frac{\nu^{(k)}}{2}+p\right)} ds_1 ds_2 \ldots ds_p. \end{aligned} \tag{75}$$

Combining (73) and (74), $e_{1,j}^{(k)}$ can be reduced to

$$e_{1,j}^{(k)} = \psi\left(\frac{\nu^{(k)}}{2}+p\right) - \log\left(\frac{\nu^{(k)}+d^{(k)}(\boldsymbol{y}_j)}{2}\right) - T_{p,\nu^{(k)}+p}^{-1}\left(\boldsymbol{y}_{1j}^{(k)};\boldsymbol{0},\boldsymbol{\Lambda}^{(k)}\right) S_{1,j}^{(k)}. \tag{76}$$

We note that the term $S_{1,j}^{(k)}$ will be very small in practice since it would be zero if we adopted an OSL EM algorithm. In which case, there would be no need to calculate the multiple integral $S_{1,j}^{(k)}$ in (74). Hence then, $e_{1,j}^{(k)}$ can be reduced to

$$e_{1,j}^{(k)} = 2_{2,j}^{(k)} - \log\left(\frac{\nu^{(k)}+d^{(k)}(\boldsymbol{y}_j)}{2}\right) - \left(\frac{\nu^{(k)}+p}{\nu^{(k)}+d^{(k)}(\boldsymbol{y}_j)}\right) + \psi\left(\frac{\nu^{(k)}+p}{2}\right). \tag{77}$$

Appendix B.2 Calculation of $\boldsymbol{e}_{3,j}^{(k)}$ and $\boldsymbol{e}_{4,j}^{(k)}$

To obtain $\boldsymbol{e}_{3,j}^{(k)}$ and $\boldsymbol{e}_{4,j}^{(k)}$, first note that the joint density of $\boldsymbol{y}_j$, $\boldsymbol{u}_j$, and w_j is given by

$$f(\boldsymbol{y}_j,\boldsymbol{u}_j,w_j) = \pi^{-p}\Gamma\left(\frac{\nu^{(k)}}{2}\right)^{-1}\left(\frac{\nu^{(k)}}{2}\right)^{\left(\frac{\nu^{(k)}}{2}\right)} w_j^{\left(\frac{\nu^{(k)}}{2}+p-1\right)} e^{-\frac{w_j}{2}\left[\nu^{(k)}+d^{(k)}(\boldsymbol{y}_j)+\left(\boldsymbol{u}_j-\boldsymbol{q}_j^{(k)}\right)^T\boldsymbol{\Lambda}^{(k)^{-1}}\left(\boldsymbol{u}_j-\boldsymbol{q}_j^{(k)}\right)\right]}. \tag{78}$$

Using Bayes' rule, the conditional density of $\boldsymbol{u}_j$ and w_j given $\boldsymbol{y}_j$ can be written as

$$f(\boldsymbol{u}_j,w_j \mid \boldsymbol{y}_j) = \frac{(2\pi)^{-\frac{p}{2}}\left|\boldsymbol{\Lambda}^{(k)}\right|^{-\frac{1}{2}} w_j^{\frac{p}{2}}\Gamma\left(w_j;\frac{\nu^{(k)}+p}{2},\frac{d^{(k)}(\boldsymbol{y}_j)}{2}\right)}{T_{p,\nu^{(k)}u+p}\left(\boldsymbol{y}_{1j}^{(k)};\boldsymbol{0},\boldsymbol{\Lambda}^{(k)}\right)} e^{-\frac{w_j}{2}\left(\boldsymbol{u}_j-\boldsymbol{q}_j^{(k)}\right)^T\boldsymbol{\Lambda}^{(k)^{-1}}\left(\boldsymbol{u}_j-\boldsymbol{q}_j^{(k)}\right)}. \tag{79}$$

From (79), standard conditional expectation calculations yield

$$\begin{aligned}
\boldsymbol{e}_{3,j}^{(k)} &= E(W_j\boldsymbol{U}_j \mid \boldsymbol{y}_j)\\
&= \int_{\boldsymbol{0}}^{\boldsymbol{\infty}}\int_0^{\infty}\boldsymbol{u}_j w_j f(\boldsymbol{u}_j,w_j \mid \boldsymbol{y}_j)\,dw_j d\boldsymbol{u}_j\\
&= T_{p,\nu^{(k)}+p}^{-1}\left(\boldsymbol{y}_{1j}^{(k)};\boldsymbol{0},\boldsymbol{\Lambda}\right)\\
&\quad \int_{\boldsymbol{0}}^{\boldsymbol{\infty}}\boldsymbol{u}_j\int_0^{\infty}\left(\tfrac{\nu^{(k)}+p}{\nu^{(k)}+d^{(k)}(\boldsymbol{y}_j)}\right)\phi\left(\boldsymbol{u}_j;\boldsymbol{q}_j,\tfrac{1}{w_j}\boldsymbol{\Lambda}\right)\\
&\quad \Gamma\left(w_j;\tfrac{\nu^{(k)}+p+2}{2},\tfrac{\nu^{(k)}+d^{(k)}(\boldsymbol{y}_j)}{2}\right)dw_j d\boldsymbol{u}_j\\
&= \left(\tfrac{\nu^{(k)}+p}{\nu^{(k)}+d^{(k)}(\boldsymbol{y}_j)}\right)T_{p,\nu^{(k)}+p}^{-1}\left(\boldsymbol{y}_{1j}^{(k)};\boldsymbol{0},\boldsymbol{\Lambda}\right)\\
&\quad \int_{\boldsymbol{0}}^{\boldsymbol{\infty}}E\left[\phi_p\left(\boldsymbol{u}_j;\boldsymbol{q}_j,\tfrac{1}{w_j}\boldsymbol{\Lambda}\right)\right]d\boldsymbol{u}_j\\
&= \left(\tfrac{\nu^{(k)}+p}{\nu^{(k)}+d^{(k)}(\boldsymbol{y}_j)}\right)T_{p,\nu^{(k)}+p}^{-1}\left(\boldsymbol{y}_{1j}^{(k)};\boldsymbol{0},\boldsymbol{\Lambda}\right)\\
&\quad \int_{\boldsymbol{0}}^{\boldsymbol{\infty}}\boldsymbol{u}_j t_{p,\nu^{(k)}+p+2}\left(\boldsymbol{u}_j;\boldsymbol{q}_j,\left(\tfrac{\nu_k+d^{(k)}(\boldsymbol{y}_j)}{\nu^{(k)}+p+2}\right)\boldsymbol{\Lambda}\right)d\boldsymbol{u}_j\\
&= \left(\tfrac{\nu^{(k)}+p}{\nu^{(k)}+d^{(k)}(\boldsymbol{y}_j)}\right)\\
&\quad \frac{T_{\nu^{(k)}+p+2}\left(\boldsymbol{q}_j^{(k)};\boldsymbol{0},\left(\frac{\nu^{(k)}+d^{(k)}(\boldsymbol{y}_j)}{\nu^{(k)}+p+2}\right)\boldsymbol{\Lambda}^{(k)}\right)}{T_{p,\nu^{(k)}+p}\left(\boldsymbol{y}_{1,j}^{(k)};\boldsymbol{0},\boldsymbol{\Lambda}^{(k)}\right)}E(\boldsymbol{X}_j),\\
&= e_{2,j}^{(k)}\,E(\boldsymbol{X}_j \mid \boldsymbol{y}_j),
\end{aligned} \tag{80}$$

where $\boldsymbol{X}_j$ is a p-dimensional t-variate truncated to the positive hyperplane $\mathbb{R}^+$, which is conditionally distributed as

$$\boldsymbol{X}_j \mid \boldsymbol{y}_j \sim tt_{p,\nu^{(k)}+p+2}\left(\boldsymbol{q}_j^{(k)},\left(\frac{\nu^{(k)}+d^{(k)}(\boldsymbol{y}_j)}{\nu^{(k)}+p+2}\right)\boldsymbol{\Lambda}^{(k)};\mathbb{R}^+\right). \tag{81}$$

Analogously, $\boldsymbol{e}_{4,j}^{(k)}$ can be reduced to

$$\boldsymbol{e}_{4,j}^{(k)} = e_{2,j}^{(k)}E(\boldsymbol{X}_j\boldsymbol{X}_j^T \mid \boldsymbol{y}_j). \tag{82}$$

The truncated moments $E(\boldsymbol{X}_j \mid \boldsymbol{y}_j)$ and $E(\boldsymbol{X}_j\boldsymbol{X}_j^T \mid \boldsymbol{y}_j)$ can be swiftly evaluated using the expressions (70) and (71) in Section 3.2.

Appendix C E-step for FM-uMST

The four conditional expectations $e_{1,hj}^{(k)}$, $e_{2,hj}^{(k)}$, $\boldsymbol{e}_{3,hj}^{(k)}$, and $\boldsymbol{e}_{4,hj}^{(k)}$ involved in the E-step are given by

$$e_{1,hj}^{(k)} = \psi\left(\frac{\nu_h^{(k)}}{2}+p\right) - \log\left(\frac{\nu_h^{(k)}+d_h^{(k)}(\boldsymbol{y}_j)}{2}\right) - T_{p,\nu_h^{(k)}+p}^{-1}\left(\boldsymbol{y}_{1,hj}^{(k)};\boldsymbol{0},\boldsymbol{\Lambda}_h^{(k)}\right)S_{1,hj}^{(k)}, \tag{83}$$

$$e_{2,hj}^{(k)} = \left(\frac{\nu_h^{(k)}+p}{\nu_h^{(k)}+d_h^{(k)}(\boldsymbol{y}_j)}\right)\frac{T_{p,\nu_h^{(k)}+p+2}\left(\boldsymbol{y}_{2,hj}^{(k)};\boldsymbol{0},\boldsymbol{\Lambda}_h^{(k)}\right)}{T_{p,\nu_h^{(k)}+p}\left(\boldsymbol{y}_{1,hj}^{(k)};\boldsymbol{0},\boldsymbol{\Lambda}_h^{(k)}\right)}, \tag{84}$$

$$\boldsymbol{e}_{3,hj}^{(k)} = e_{2,hj}^{(k)}E(\boldsymbol{X}_{hj} \mid \boldsymbol{y}_j), \tag{85}$$

$$\boldsymbol{e}_{4,hj}^{(k)} = e_{2,hj}^{(k)}E(\boldsymbol{X}_{hj}\boldsymbol{X}_{hj}^T \mid \boldsymbol{y}_j), \tag{86}$$

where $S_{1,hj}^{(k)}$ is a scalar defined by

$$S_{1,hj}^{(k)} = \int_{-\infty}^{\left[\boldsymbol{q}_{hj}^{(k)}\right]_1}\int_{-\infty}^{\left[\boldsymbol{q}_{hj}^{(k)}\right]_2}\cdots\int_{-\infty}^{\left[\boldsymbol{q}_{hj}^{(k)}\right]_p} \log\left(1+\frac{\boldsymbol{s}^T{\boldsymbol{\Lambda}_h}^{-1}\boldsymbol{s}}{\nu_h^{(k)}+d_h^{(k)}(\boldsymbol{y}_j)}\right) \left[1+\frac{\boldsymbol{s}^T{\boldsymbol{\Lambda}_h}^{-1}\boldsymbol{s}}{\nu_h^{(k)}+d_h^{(k)}(\boldsymbol{y}_j)}\right]^{-\left(\frac{\nu_h^{(k)}}{2}+p\right)}d\boldsymbol{u}, \tag{87}$$

and $\boldsymbol{X}_{hj}$ is a truncated p-dimensional t-variate given by

$$\boldsymbol{X}_{hj} \mid \boldsymbol{y}_j \sim tt_{p,\nu_h+p+2}\left(\boldsymbol{q}_{hj}^{(k)},\left(\frac{\nu_h^{(k)}+d_h^{(k)}(\boldsymbol{y}_j)}{\nu_h^{(k)}+p+2}\right)\boldsymbol{\Lambda}_h^{(k)},\mathbb{R}^+\right).$$

The first two moments of $\boldsymbol{X}_{hj}$ can be implicitly expressed in terms of the parameters $\boldsymbol{q}_{hj}^{(k)}$, $d_h^{(k)}(\boldsymbol{y}_j)$, $\boldsymbol{\Lambda}_h^{(k)}$, $\nu_h^{(k)}$ using results (70) and (71). It is worth emphasizing that computation of $\boldsymbol{e}_{3hj}^{(k)}$ and $\boldsymbol{e}_{4hj}^{(k)}$ depends on algorithms for evaluating the multivariate t-distribution function, for which fast procedures are available.

Acknowledgements We would like to thank Professor Seung-Gu Kim for comments and corrections, and Drs Kui (Sam) Wang and Saumyadipta Pyne for their helpful discussions on this topic.

References

Akaike H (1974) A new look at the statistical model identification. Automatic Control 19:716–723

Arellano-Valle RB, Azzalini A (2006) On the unification of fmailies 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

Arnold BC, Beaver RJ (2002) Skewed multivariate models related to hidden truncation and/or selective reporting. Test 11:7–54

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

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

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):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

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, Discase Mapping and Others. Chapman and Hall, New York

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

Brinkman R, Gaspareto M, Lee SJ, Ribickas A, Perkins J, Janssen W, Smiley R, Smith C (2007) High content flow cytometry and temporal data analysis for defining a cellular signature of graft versus host disease. Biological of Blood and Marrow Transplantation 13:691–700

Cabral C, Bolfarine H, Pereira J (2008) Bayesian density estimation using skew student-t-normal mixtures. Computational Statistics and Data Analysis 52:5075–5090

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

Dempster A, Laird NM, Rubin D (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

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

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

Genz A, Bretz F (2002) Methods for the computation of multivariate t-probabilities. Journal of Computational and Graphical Statistics 11:950–971

Gómez H, Venegas O, Bolfarine H (2007) Skew-symmetric distributions generated by the distribution function of the normal distribution. Environmetrics 18:395–407

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

Green PJ (1990) On use of the em algorithm for penalized likelihood estimation. Journal of the Royal Statistical Society B 52:443–452

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

Ho H, Lin T, Chen H, Wang W (2012a) Some results on the truncated multivariate t distribution. Journal of Statistical Planning and Inference 142:25–40

Ho H, Pyne S, Lin T (2012b) Maximum likleihood inference for mixtures of skew student-t-normal distributions through practical em-type algorithms. Statistics and Computing 22:287–299

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, Nadarajah S (2004) Multivariate t Distributions and Their Applications. Cambridge University Press, Cambridge

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 G (2011) On the fitting of mixtures of multivariate skew t-distributions via the em algorithm. arXiv:11094706 [statME]

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, Lee JC, Hsieh WJ (2007a) Robust mixture modeling using the skew-t distribution. Statistics and Computing 17:81–92

Lin TI, Lee JC, Yen SY (2007b) Finite mixture modelling using the skew normal distribution. Statistica Sinica 17:909–927

Lindsay BG (1995) Mixture Models: Theory, Geometry, and Applications. NSF-CBMS Regional Conference Series in probability and Statistics, Volume 5, Institute of Mathematical Statistics, Hayward, CA

Liseo B, Loperfido N (2003) A bayesian interpretation of the multivariate skew-normal distribution. Statistics & Probability Letters 61:395–401

Liu C, Rubin D (1994) The ecme algorithm: a simple extension of the em and ecm with faster monotone convergence. Biometricka 81:633–648

Maier LM, Anderson DE, De Jager PL, Wicker L, Hafler DA (2007) Allelic variant in ctla4 alters t cell phosphorylation patterns. Proceedings of the National Academy of Sciences of the United States of America 104:18,607–18,612

McLachlan G, 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, vol 1451, Springer-Verlag, Berlin, pp 658–666

McLachlan GJ, Basford KE (1988) Mixture Models: Inference and Applications. Marcel Dekker, New York

McLachlan GJ, Peel D (2000) Finite Mixture Models. Wiley Series in Probability and Statistics

O'Hagan A (1973) Bayes estimation of a convex quadratic. Biometrika 60:565–571

O'Hagan A (1976) Moments of the truncated multivariate-t distribution. http://www.tonyohagan.co.uk/academic/pdf/trunc_multi_t.PDF

O'Hagan A, Murphy T, Gormley I (2012) Computational aspects of fitting mixture models via the expectation-maximization algorithm. Computational Statistics and Data Analysis TBA:TBA

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

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 (2009) Automated high-dimensional flow cytometric data analysis. Proceedings of the National Academy of Sciences USA 106:8519–8524

Sahu S, Dey D, Branco M (2003) A new class of multivariate skew distributions with applications to bayesian regression models. The Canadian Journal of Statistics 31:129–150

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

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

Virbik I, McNicholas P (2012) Analytic calculations for the em algorithm for multivariate skew t-mixture models. Statistics and Probability Letters 82:1169–1174

Wang K (2009) `EMMIX-skew`: EM Algorithm for Mixture of Multivariate Skew Normal/t Distributioins. URL http://www.maths.uq.edu.au/ gjm/mix_soft/EMMIX-skew, R package version 1.0-12

Wang K, Ng SK, McLachlan GJ (2009) Multivariate skew t mixture models: applications : applications to fluorescence-activated cell sorting data. In: Shi H, Zhang Y, Botema M, Lovell B, Maoder A (eds) DICTA 2009 (Conference of Digital Image Computing: Techniques and Applications, Melbourne), IEEE Computer Society, Los Alamitos, California, pp 526–531