On the fitting of mixtures of multivariate skew $t$-distributions via the EM algorithm

Sharon Lee*, Geoffrey J. McLachlan *,†
Department of mathematics, the University of Queensland, Brisbane, Australia * and Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia †

Abstract

We show how the expectation-maximization (EM) algorithm can be applied exactly for the fitting of mixtures of general multivariate skew $t$ (MST) distributions, eliminating the need for computationally expensive Monte Carlo estimation. Finite mixtures of 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. However, without restrictions on the characterizations of the component skew $t$-distributions, Monte Carlo methods have been used to fit these models. In this paper, we show how the EM algorithm can be implemented for the iterative computation of the maximum likelihood estimates of the model parameters without resorting to Monte Carlo methods for mixtures with unrestricted MST components. The fast calculation of semi-infinite integrals on the E-step of the EM algorithm is effected by noting that they can be put in the form of moments of the truncated multivariate non-central $t$-distribution, 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 demonstrate the usefulness of the proposed methodology by some applications to three real data sets.

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, Smith, and Markov (1985), Lindsay (1995), McLachlan and Basford (1988), and McLachlan and Peel (2000), among others; 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. 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 symmetry in data sets. Following the development of the SN and skew t-mixture models by Lin, Lee, and Yen (2007), and Lin, Lee, and Hsieh (2007), respectively, Basso et al. (2010) studied a class of mixture models where the components densities are scale mixtures of skew-normal distributions introduced by Branco and Dey (2001), 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.

In a study of automated flow cytometry analysis, Pyne et al. (2009) proposed a finite mixture of multivariate skew t-distributions based on a ‘restricted’ variant of the skew t-distribution introduced by Sahu, Dey, and Branco (2003). Lin (2010) considered a similar approach, but working with the original (unrestricted) characterization by Sahu et al. (2003). However, with this more general formulation, maximum likelihood (ML) estimation via the EM algorithm (Dempster, Laird, and Rubin, 1977) 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 potential drawback of this approach is that the model fitting procedure relies on MC estimates which can be computationally expensive.

In this paper, we show how the EM algorithm can be implemented exactly to calculate the ML estimates of the parameters for the (unrestricted) multivariate skew t-mixture model, based on 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 show that the proposed algorithm is highly efficient compared to the version with a MC E-step. It produces highly accurate results for which, if MC were to achieve comparable accuracy, a large number of draws would be necessary.

The remainder of the paper is organized as follows. In Section 2, for the sake of completeness, we include a brief description of the multivariate skew t-distribution (MST) used for defining the multivariate skew t-mixture model. We also describe the truncated t-distribution in the multivariate case, critical for the swift evaluation of the integrals on the E-step occurring in the calculation of some of the conditional expectations. Section 3 presents the development of an EM algorithm for obtaining ML estimates for the MST distribution. In the following section, the finite mixture of MST (FM-MST) distributions is defined. Section 5 presents an implementation of the EM algorithm to the fitting of the FM-MST model. An approximation to the observed information matrix is discussed in Section 6. Finally, we present some applications of the proposed methodology in Section 7.

2 Preliminaries

We begin by defining the multivariate skew t-distribution and briefly describing some related properties. Some alternative versions of the distribution are also discussed. Next, we introduce the truncated multivariate t-distribution and provide some formulas for computing its moments. These expressions are crucial for the swift evaluation of the conditional expectations on the E-step to be discussed in the next section.
2.1 The Multivariate Skew t-Distribution

Following Sahu et al. (2003), a random vector $Y$ is said to follow a $p$-dimensional skew $t$-distribution with $p \times 1$ location vector $\mu$, $p \times p$ scale matrix $\Sigma$, $p \times 1$ skewness vector $\delta$, and scalar degrees of freedom $\nu$, if its density is given by

$$f_p(y; \mu, \Sigma, \delta, \nu) = 2^p t_{p,\nu}(y; \mu, \Omega) T_{p,\nu+p}(y^*; 0, \Lambda),$$

(1)

where

$$\Delta = \text{diag}(\delta),$$
$$\Omega = \Sigma + \Delta \Delta^T,$$
$$y^* = q \sqrt{\frac{\nu + p}{\nu + d(y)}},$$
$$q = \Delta^T \Omega^{-1} (y - \mu),$$
$$d(y) = (y - \mu)^T \Omega^{-1} (y - \mu),$$
$$\Lambda = I_p - \Delta^T \Omega^{-1} \Delta.$$

Here the operator $\text{diag}(\delta)$ denotes a diagonal matrix with diagonal elements specified by the vector $\delta$. Also, we let $t_{p,\nu}(.; \mu, \Sigma)$ be the $p$-dimensional $t$-distribution with location vector $\mu$, scale matrix $\Sigma$, and degrees of freedom $\nu$, and $T_{p,\nu}(.; \mu, \Sigma)$ the corresponding (cumulative) distribution function. The notation $Y \sim \text{ST}_{p,\nu}(\mu, \Sigma, \delta)$ will be used. Note that when $\delta = 0$, (1) reduces to the symmetric $t$-density $t_{p,\nu}(y; \mu, \Sigma)$. Also, when $\nu \to \infty$, we obtain the skew normal distribution.

The MST distribution admits a convenient hierarchical form,

$$Y | u, w \sim N_p (\mu + \Delta u, \frac{1}{w} \Sigma),$$
$$U | w \sim HN_p (0, \frac{1}{w} I_p),$$
$$W \sim \text{gamma} \left( \frac{\nu}{2}, \frac{\nu}{2} \right),$$

(2)

where $I_p$ is the $p \times p$ identity matrix, $N_k(\mu, \Sigma)$ denotes the multivariate normal distribution with mean $\mu$ and covariance matrix $\Sigma$, $HN_p(0, \Sigma)$ represents the $p$-dimensional half-normal distribution with mean $0$ and scale matrix $\Sigma$, and $\text{gamma}(\alpha, \beta)$ is the Gamma distribution with mean $\alpha/\beta$.

We observe from (2) that the MST distribution (1) has the following stochastic representation. Suppose that conditional on the value $w$ of the gamma random variable $W$,

$$\left( \begin{array}{c} \mathbf{U}_0 \\ \mathbf{U} \end{array} \right) \sim N \left( \left( \begin{array}{c} \mu \\ 0 \end{array} \right), \left( \begin{array}{cc} \Sigma/w & 0 \\ 0 & I_p/w \end{array} \right) \right),$$

(3)

where $I_p$ denotes the $p \times p$ identity matrix, $0$ denotes the zero vector of appropriate dimension, and $\mathbf{U}_0$ is a $p$-dimensional random vector. Then

$$Y = \Delta |\mathbf{U}| + U_0$$

(4)

has the multivariate skew $t$-distribution density (1). In the above, $|\mathbf{U}|$ denotes the vector whose $i$th element is the magnitude of the $i$th element of the vector $\mathbf{U}$. It is worth noting
that, although also known as the multivariate skew $t$-distribution, the versions considered by Azzalini and Dalla Valle (1996), Gupta (2003), and Lachos, Ghosh, and Arellano-Valle (2010), among others, are different from (1). 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 (1). Recently, Pyne et al. (2009) proposed a simplified version of the skew $t$-density given by (1) by replacing the term $\Delta \mathbf{U}$ in (1) by the term $\mathbf{\delta} \mathbf{U}$, where $\mathbf{U}$ is a univariate standard normal random variable, leading to the reduced skew $t$-density:

$$2^{\nu}t_{p,\nu}(\mathbf{y}; \mathbf{\mu}, \mathbf{\Sigma}) T_{1,\nu+p}(\mathbf{y}^*_i; 0, 1),$$

(5)

where $y^*_i = [(\nu + p) / (\nu + d(y))]^{\frac{1}{2}} \mathbf{\delta}^T \mathbf{\Omega}^{-1} \mathbf{\delta}$. One immediate consequence of this type of ‘simplification’ is that the correlation structure of the original symmetric model is affected by the introduction of skewness, whereas for (1) the correlation structure remains the same, as noted in Arellano-Valle, Bolfarine, and Lachos (2007). Nevertheless, one major advantage of having simplified forms like (5) is that calculations on the E-step can be expressed in closed form.

2.2 The truncated multivariate $t$-distribution

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

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

(6)

For a random vector $\mathbf{X}$ with density (6), we write $\mathbf{X} \sim tt_{p,\nu}(\mathbf{\mu}, \mathbf{\Sigma}; \mathcal{A})$. For our purposes, we will be concerned with the first two moments of $\mathbf{X}$, specifically $E(\mathbf{X})$ and $E(\mathbf{X}\mathbf{X}^T)$. Explicit formulas for the truncated central $t$-distribution in the univariate case $tt_{1,\nu}(0, \sigma^2; \mathcal{A})$ 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. We will generalize these results to the multivariate non-central case here.

Before presenting the expressions, it will be convenient to introduce some notation. Let $\mathbf{x}$ be a vector, where $x_i$ denotes the $i$th element and $\mathbf{x}_{ij}$ is a two-dimensional vector with elements $x_i$ and $x_j$. Also, $\mathbf{x}_{-i}$ and $\mathbf{x}_{-ij}$ represents the $(p-1)$ and $(p-2)$-dimensional vector, respectively, with the corresponding elements removed. For a matrix $\mathbf{X}$, $\mathbf{x}_{ij}$ denotes the $ij$th element, and $\mathbf{X}_{ij}$ defines the $2 \times 2$ matrix consisting of the elements $x_{ii}, x_{ij}, x_{ji}$ and $x_{jj}$. $\mathbf{X}_{-i}$ is created by removing the $i$th row and column from $\mathbf{X}$. Similarly, $\mathbf{X}_{-ij}$ is the $(p-2)$ square matrix resulting from the removal of the $i$th and $j$th row and column from $\mathbf{X}$. Lastly, $\mathbf{X}_{(ij)}$ is the $i$th and $j$th column of $\mathbf{X}$ with the elements of $\mathbf{X}_{ij}$ removed, yielding a $(p-2) \times 2$ matrix. We now proceed to the expressions for the first two moments of $\mathbf{X}$.

With some effort, one can show that the first moment of (6) is

$$E(\mathbf{X}) = \mathbf{\mu} - c^{-1}\mathbf{\Sigma}\mathbf{\xi} = \mathbf{\mu}^*,$$

(7)

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

$$\xi_i = (2\pi\sigma_i)^{-\frac{1}{2}} \left( \frac{\nu}{\nu + \sigma_i^{-1}(a_i - \mu_i)^2} \right) \left( \frac{\nu - 1}{2} \right) \frac{\Gamma\left(\frac{\nu - 1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \sqrt{\frac{\nu}{2}} T_{p,\nu-1}(a^*_i; 0, \mathbf{\Sigma}^*)$$
for \( i = 1, \ldots, p \), and where
\[
\alpha^* = (a_{-i} - \mu_{-i}) - (a_{-i} - \mu_{-i}) \sigma_{ii}^{-1} \Sigma(i)
\]
and
\[
\Sigma^* = \left( \frac{\nu + \sigma_{ii}^{-1} (a_i - \mu_i)^2}{\nu - 1} \right) \left( \Sigma_{-i} - \frac{1}{\sigma_{ii}} \Sigma(i) \right).
\]

The second moment is given by
\[
E \left( XX^T \right) = \mu\mu^T + \mu\mu^{*T} + \mu^*\mu^T - c^{-1} \Sigma H \Sigma
+ c^{-1} \left( \frac{\nu}{\nu - 2} \right) T_{p,\nu - 2} (a - \mu; 0, \left( \frac{\nu}{\nu - 2} \right) \Sigma) \Sigma,
\]
where \( H \) is a \( p \times p \) matrix with off-diagonal elements
\[
h_{ij} = \frac{1}{2\pi \sqrt{\sigma_{ii} \sigma_{jj} - \sigma_{ij}^2}} \left( \frac{\nu}{\nu - 2} \right)^{\frac{\nu}{2} - 1} T_{p - 2,\nu - 2} (a^*; 0, \Sigma^*) , \quad i \neq j,
\]
and diagonal elements,
\[
h_{ii} = \sigma_{ii}^{-1} (a_i - \mu_i) \xi_i - \sigma_{jj}^{-1} \sum_{j \neq i} \sigma_{ij} h_{ij},
\]
and
\[
\nu^* = \nu + \left( a_{ij} - \mu_{ij} \right)^T \Sigma_{ij}^{-1} \left( a_{ij} - \mu_{ij} \right),
\]
\[
\alpha^{**} = (a_{-ij} - \mu_{-ij}) - \Sigma_{(ij)}^{-1} \Sigma_{ij}^{-1} (a_{ij} - \mu_{ij}),
\]
\[
\Sigma^{**} = \frac{\nu^*}{\nu - 2} \left( \Sigma_{-ij} - \Sigma_{(ij)} \Sigma_{ij}^{-1} \Sigma_{(ij)}^T \right).
\]

It is worth noting that evaluation of the expressions \([7]\) and \([8]\) 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 \texttt{mvtnorm} package available from the R website.

### 3 ML Estimation for the MST Distribution

In this section, we describe an EM algorithm for the ML estimation of the MST distribution specified by \([1]\). To apply the EM algorithm, the observed data vector \( y = (y_1^T, \ldots, y_n^T)^T \) is regarded as incomplete, and we introduce two latent variables denoted by \( u \) and \( w \), as defined by \([2]\). We let \( \theta \) be the parameter containing the elements of the location parameter \( \mu \), the distinct elements of the scale matrix \( \Sigma \), the elements of the skew parameter \( \delta \), and the degrees of freedom \( \nu \). It follows that the complete-data log-likelihood function for \( \theta \) is given by
\[
L_c(\theta; y, u, w) = K - \frac{1}{2} n \log |\Sigma| - n \log \Gamma \left( \frac{1}{2} \nu \right) + \frac{1}{2} n \nu \log \left( \frac{1}{2} \nu \right)
- \frac{1}{2} w \left( d(y) + (u - q)^T \Lambda^{-1} (u - q) \right)
+ \left( \frac{1}{2} \nu + p - 1 \right) \log(w),
\]
\[
(9)
\]
where $K$ does not depend on $\theta$.

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 $L(\theta^{(k)})$ is bounded above. Here $\theta^{(k)}$ denotes the value of $\theta$ after the $k$th iteration.

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 $y$, using the current estimate $\theta^{(k)}$ for $\theta$. That is, we have to calculate the so-called $Q$-function defined by

$$Q(\theta; \theta^{(k)}) = E_{\theta^{(k)}} \{ L_c(\theta; y, u, w) | y \}, \quad (10)$$

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

$$e_{1,j}^{(k)} = E_{\theta^{(k)}} \{ \log(W_j) | y_j \},$$
$$e_{2,j}^{(k)} = E_{\theta^{(k)}} (W_j | y_j),$$
$$e_{3,j}^{(k)} = E_{\theta^{(k)}} (W_j U_j | y_j),$$
$$e_{4,j}^{(k)} = E_{\theta^{(k)}} (W_j U_j U_j^T | y_j).$$

Note that the $Q$-function does not admit a closed form expression for this problem, due to the conditional expectations $e_{1,j}^{(k)}$, $e_{3,j}^{(k)}$, and $e_{4,j}^{(k)}$ not being able to be evaluated in closed form.

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

$$f(w_j | y_j) = \frac{\Gamma \left( w_j; \frac{\nu^{(k)}+p}{2}, \frac{\nu^{(k)}+d^{(k)}(y_j)}{2} \right) \phi_p \left( q_j^{(k)} \sqrt{w_j}; 0, \Lambda^{(k)} \right)}{T_{p,\nu^{(k)}+p} \left( y_j^{(k)}; 0, \Lambda^{(k)} \right)}, \quad (11)$$

where

$$y_j^{(k)} = q_j^{(k)} \sqrt{\frac{\nu^{(k)}+p}{\nu^{(k)}+d^{(k)}(y_j)}},$$
$$q_j^{(k)} = \Delta^{(k)} \Omega^{(k)-1} (y_j - \mu^{(k)}),$$
$$d^{(k)}(y_j) = (y_j - \mu^{(k)})^T \Omega^{(k)-1} (y_j - \mu^{(k)}),$$

and $0$ is the zero vector of appropriate dimension.

The conditional expectation $E_{\theta^{(k)}} \{ \log(W_j) | y_j \}$ can be reduced to

$$e_{1,j}^{(k)} = \left( \frac{\nu^{(k)}+p}{\nu^{(k)}+d^{(k)}(y_j)} \right) \frac{T_{p,\nu^{(k)}+p+2} \left( q_j^{(k)} \sqrt{\frac{\nu^{(k)}+p+2}{\nu^{(k)}+d^{(k)}(y_j)}}, 0, \Lambda^{(k)} \right)}{T_{p,\nu^{(k)}+p} \left( y_j^{(k)}; 0, \Lambda^{(k)} \right)}$$
$$- \log \left( \frac{\nu^{(k)}+d^{(k)}(y_j)}{2} \right) - \left( \frac{\nu^{(k)}+p}{\nu^{(k)}+d^{(k)}(y_j)} \right) + \psi \left( \frac{\nu^{(k)}+p}{2} \right) + S, \quad (12)$$
where the last term \( S \) is given by

\[
S = \psi\left(\frac{\nu^{(k)} + p}{2}\right) - \psi\left(\frac{\nu^{(k)} + p}{2}\right) + \left(\frac{\nu^{(k)} + p}{\nu^{(k)} + d^{(k)}(y_j)}\right)
- \left(\frac{\nu^{(k)} + p}{\nu^{(k)} + d^{(k)}(y_j)}\right)^2 \frac{T_{p,\nu^{(k)}+p+2}\left(q_j^{(k)}\sqrt{\nu^{(k)}+p+2}/\nu^{(k)}+d^{(k)}(y_j)\right); 0, \Lambda^{(k)}}{T_{p,\nu^{(k)}+p}\left(q_j^{(k)}\sqrt{\nu^{(k)}+p}/\nu^{(k)}+d^{(k)}(y_j)\right); 0, \Lambda^{(k)}}
- \frac{\pi (\nu^{(k)} + p)}{2} \left[1 - \frac{1}{2} \frac{\Gamma\left(\frac{\nu^{(k)} + p}{2}\right)}{\Gamma\left(\frac{\nu^{(k)} + p}{2}\right)} S_{1,j}^{(k)}\right], \quad (13)
\]

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

\[
S_{1,j}^{(k)} = \int_{-\infty}^{[q_j^{(k)}]} \int_{-\infty}^{[q_j^{(k)}]} \cdots \int_{-\infty}^{[q_j^{(k)}]} \log \left(1 + \frac{s^T \Lambda^{-1} \Lambda^{(k)} s}{\nu^{(k)} + d^{(k)}(y_j)}\right) ds_1ds_2\ldots ds_p, \quad (14)
\]

and \( \psi(\cdot) \) denotes the Digamma function.

Combining (12) and (13), \( e_{1,j}^{(k)} \) can be reduced to

\[
e_{1,j}^{(k)} = \psi\left(\frac{\nu^{(k)} + p}{2}\right) - \log\left(\frac{\nu^{(k)} + d^{(k)}(y_j)}{2}\right)
- T_{p,\nu^{(k)}+p}^{-1}\left(q_j^{(k)}\sqrt{\nu^{(k)}+p}/\nu^{(k)}+d^{(k)}(y_j)\right); 0, \Lambda^{(k)} S_{1,j}^{(k)}. \quad (15)
\]

We note that the term \( S \) will be very small in practice since it would be zero if we adopted a one-step late (OSL) EM algorithm (Green, 1990). In which case, there would be no need to calculate the multiple integral \( S_{1,j}^{(k)} \) in (13). Hence then, \( e_{1,j}^{(k)} \) can be approximated as

\[
e_{1,j}^{(k)} \approx \left(\frac{\nu^{(k)} + p}{\nu^{(k)} + d^{(k)}(y_j)}\right) \frac{T_{p,\nu^{(k)}+p+2}\left(q_j^{(k)}\sqrt{\nu^{(k)}+p+2}/\nu^{(k)}+d^{(k)}(y_j)\right); 0, \Lambda^{(k)}}{T_{p,\nu^{(k)}+p}\left(q_j^{(k)}\sqrt{\nu^{(k)}+p}/\nu^{(k)}+d^{(k)}(y_j)\right); 0, \Lambda^{(k)}}
- \log\left(\frac{\nu^{(k)} + d^{(k)}(y_j)}{2}\right) - \left(\frac{\nu^{(k)} + p}{\nu^{(k)} + d^{(k)}(y_j)}\right) + \psi\left(\frac{\nu^{(k)} + p}{2}\right). \quad (16)
\]

However, in computing \( e_{1,j}^{(k)} \) in the examples to follow, we did not take \( S_{1,j}^{(k)} \) to be zero but calculated it by numerical integration.

It can be easily shown that \( e_{2,j}^{(k)} \) can be written in closed form (see, for example, Lin (2010)), given by

\[
e_{2,j}^{(k)} = \left(\frac{\nu^{(k)} + p}{\nu^{(k)} + d^{(k)}(y_j)}\right) \frac{T_{p,\nu^{(k)}+p+2}\left(q_j^{(k)}\sqrt{\nu^{(k)}+p+2}/\nu^{(k)}+d^{(k)}(y_j)\right); 0, \Lambda^{(k)}}{T_{p,\nu^{(k)}+p}\left(q_j^{(k)}\sqrt{\nu^{(k)}+p}/\nu^{(k)}+d^{(k)}(y_j)\right); 0, \Lambda^{(k)}}. \quad (17)
\]
To obtain $e_{3,j}^{(k)}$ and $e_{4,j}^{(k)}$, first note that the joint density of $y_j$, $u_j$, and $w_j$ is given by

$$f(y_j, u_j, w_j) = \pi^{-p} \Gamma \left( \frac{\nu(k)}{2} - 1 \right) \left( \frac{\nu(k)}{2} \right) w_j^{n_p} \left( \frac{\nu(k)}{2} + p - 1 \right) \left( w_j \frac{\nu(k) + d(k)(y_j)}{2} \right) e^{-\frac{w_j}{2} \left( \frac{\nu(k) + d(k)(y_j)}{2} \right)^{1/2} (u_j - q_j^{(k)})^T \Lambda(k)^{-1} (u_j - q_j^{(k)})}.$$  \hfill (18)

Using Bayes’ rule, the conditional density of $u_j$ and $w_j$ given $y_j$ can be written as

$$f(u_j, w_j | y_j) = \frac{w_j^{\frac{\nu(k) + p}{2}} \Gamma \left( \frac{\nu(k) + p + d(k)(y_j)}{2} \right) e^{-\frac{w_j}{2} \left( \frac{\nu(k) + p + d(k)(y_j)}{2} \right)^{1/2} (u_j - q_j^{(k)})^T \Lambda(k)^{-1} (u_j - q_j^{(k)})}}{(2\pi)^{\frac{p}{2}} \left| \Lambda(k) \right|^{\frac{1}{2}} \Gamma \left( p + d(k)(y_j) \right) \left( \frac{\nu(k) + p + d(k)(y_j)}{2} \right)^{\frac{p}{2}} \left( u_j - q_j^{(k)} \right)^T \Lambda(k)^{-1} \left( u_j - q_j^{(k)} \right)}.$$

From [19], standard conditional expectation calculations yield

$$e_{3,j}^{(k)} = \left( \frac{\nu(k) + p}{\nu(k) + d(k)(y_j)} \right) T_{\nu(k)+p+2} \left( q_j^{(k)} ; 0, \left( \frac{\nu(k) + d(k)(y_j)}{\nu(k) + p + 2} \right) \Lambda(k) \right) S_{2,j}^{(k)}$$

(19)

where $S_{2,j}^{(k)}$ represents the expected value of a truncated $p$-dimensional $t$-variate $X$, which is distributed as,

$$X \sim t_{\nu(k)+p+2} \left( q_j^{(k)} ; 0, \left( \frac{\nu(k) + p + 2}{\nu(k) + d(k)(y_j)} \right) \Lambda(k) \right).$$

(20)

That is, the random vector $X$ is truncated to lie in the positive hyperplane $\mathbb{R}^+$. Analogously, $e_{4,j}^{(k)}$ can be reduced to

$$e_{4,j}^{(k)} = \left( \frac{\nu(k) + p}{\nu(k) + d(k)(y_j)} \right) T_{\nu(k)+p+2} \left( q_j^{(k)} ; 0, \left( \frac{\nu(k) + d(k)(y_j)}{\nu(k) + p + 2} \right) \Lambda(k) \right) S_{3,j}^{(k)}$$

(21)

where $S_{3,j}^{(k)}$ represents the second moment of $X$. The truncated moments $S_{2,j}^{(k)}$ and $S_{3,j}^{(k)}$ can be swiftly evaluated using the expressions (7) and (8) in Section 2.2.

### 3.1 M-step

On the $(k+1)$th iteration, the M-step consists of the maximization of the $Q$-function (10) with respect to $\theta$. 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 $\theta$ into subvectors, $\theta = (\theta_1^T, \theta_2^T, \theta_3^T, \theta_4^T)^T$, where $\theta_1 = \mu$, $\theta_2 = \delta$, $\theta_3$ is the vector containing the distinct elements of $\Sigma$, and $\theta_4 = \nu$. To compute $\mu^{(k+1)}$, we maximize $Q(\mu, \theta_2^{(k)}, \theta_3^{(k)}, \theta_4^{(k)} ; \theta^{(k)})$ with respect to $\mu$, and to compute $\delta^{(k+1)}$, we first update $\mu$ to $\mu^{(k+1)}$ and then maximize $Q(\mu^{(k+1)}, \delta, \theta_3^{(k)}, \theta_4^{(k)} ; \theta^{(k)})$ with respect to $\delta$, and so on.
For clarity and distinction from \( \text{diag} (a) \) introduced earlier, we let \( \text{DIAG}(A) \) denote the operator that produces a vector by extracting the diagonal elements of \( A \). Simple algebraic manipulations lead to the following closed form expressions for \( \mu^{(k+1)} \), \( \Sigma^{(k+1)} \), and \( \Lambda^{(k+1)} \),

\[
\mu^{(k+1)} = \frac{\sum_{j=1}^{n} \left[ e^{(k)}_{2,j} y_j - \Delta^{(k)} e^{(k)}_{3,j} \right]}{\sum_{j=1}^{n} e^{(k)}_{2,j}},
\]

\[
\delta^{(k+1)} = \text{DIAG} \left\{ \left[ \sum_{j=1}^{n} \left( y_j - \mu^{(k+1)} \right) e^{(k)}_{3,j} \right] \left[ \sum_{j=1}^{n} e^{(k)}_{4,j} \right]^{-1} \right\},
\]

and

\[
\Sigma^{(k+1)} = \frac{1}{n} \sum_{j=1}^{n} \left[ \Delta^{(k+1)} e^{(k)} e^{(k)T} \Delta^{(k+1)T} \right.
- (y_j - \mu^{(k+1)}) e^{(k)}_{3,j} \Delta^{(k+1)}
+ (y_j - \mu^{(k+1)}) (y_j - \mu^{(k+1)})^T e^{(k)}_{2,j}
- \Delta^{(k+1)} e^{(k)}_{3,j} (y_j - \mu^{(k+1)})^T \right],
\]

where \( \Delta^{(k+1)} = \text{diag} (\delta^{(k+1)}) \).

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

\[
\log \left( \frac{\nu^{(k+1)}}{2} \right) - \psi \left( \frac{\nu^{(k+1)}}{2} \right) + 1 = \frac{1}{n} \sum_{j=1}^{n} \left( e^{(k)}_{2,j} - e^{(k)}_{1,j} \right),
\]

In summary, the ECM algorithm proceeds as follows on the \( (k+1) \)th iteration:

**E-step:** Given \( \theta = \theta^{(k)} \), compute the four conditional expectations \( e^{(k)}_{1,j} \), \( e^{(k)}_{2,j} \), \( e^{(k)}_{3,j} \) and \( e^{(k)}_{4,j} \) by using (15), (17), (19), and (21), respectively, for \( j = 1, \ldots, n \).

**M-step:** Update \( \mu^{(k+1)} \), \( \delta^{(k+1)} \), \( \Sigma^{(k+1)} \) and by using (22), (23), and (24). Calculate \( \nu^{(k+1)} \) by solving (25).

### 4 The Multivariate Skew \( t \)-Mixture Model

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

\[
f(y; \Psi) = \sum_{h=1}^{g} \pi_h f_p(y; \mu_h, \Sigma_h, \delta_h, \nu_h),
\]

where \( f_p(y; \mu_h, \Sigma_h, \delta_h, \nu_h) \) denotes the \( i \)th MST component of the mixture model as defined by (1), with location parameter \( \mu_h \), scale matrix \( \Sigma_h \), skew parameter \( \delta_h \), and degrees of freedom \( \nu_h \). The mixing proportions \( \pi_h \) satisfy \( \pi_h \geq 0 \) (\( h = 1, \ldots, g \)) and \( \sum_{h=1}^{g} \pi_h = 1 \). We shall denote
the model defined by (26) by FM-MST (finite mixture of MST) distributions. Let $\Psi$ contain all the unknown parameters of the FM-MST model; that is, $\Psi = (\pi_1, \ldots, \pi_{g-1}, \theta_1^T, \ldots, \theta_g^T)^T$ where now $\theta_h$ consists of the unknown parameters of the $i$th component density function.

To formulate the estimation of the unknown parameters in the FM-MST model as an incomplete-data problem in the EM framework, a set of latent component labels $z_j = (z_{1j}, \ldots, z_{gj})^T (j = 1, \ldots, n)$ is introduced, where each element $z_{hj}$ is a binary variable defined as

$$z_{hj} = \begin{cases} 1, & \text{if } y_j \text{, belongs to component } i, \\ 0, & \text{otherwise,} \end{cases}$$

and $\sum_{h=1}^g z_{hj} = 1 (j = 1, \ldots, n)$. Hence, the random vector $Z_j$ corresponding to $z_j$ follows a multinomial distribution with one trial and cell probabilities $\pi_1, \ldots, \pi_g$; that is, $Z_j \sim \text{Mult}_g(1; \pi_1, \ldots, \pi_g)$. It follows that the FM-MST model can be represented in the hierarchical form given by

$$Y_j \mid u_j, w_j, z_{hj} = 1 \sim N_p \left( \mu_h + \Delta_h u_j, \frac{1}{w_j} \Sigma_h \right),$$

$$U_j \mid w_j, z_{hj} = 1 \sim HN_p \left( 0, \frac{1}{w_j} I_p \right),$$

$$W_j \mid z_{hj} = 1 \sim \text{gamma} \left( \nu_h, \frac{\nu_h}{2} \right),$$

$$Z_j \sim \text{Mult}_g (1, \pi),$$

where $\Delta_h = \text{diag} (\delta_h)$ and $\pi = (\pi_1, \ldots, \pi_g)^T$.

5 ML Estimation for FM-MST Distributions

From the hierarchical characterization [28] of the FM-MST distributions, the complete-data log-likelihood function is given by

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

where

$$L_{1c} (\Psi) = \sum_{h=1}^g \sum_{j=1}^n z_{hj} \log (\pi_h),$$

$$L_{2c} (\Psi) = \sum_{h=1}^g \sum_{j=1}^n z_{hj} \left[ \left( \frac{\nu_h}{2} \right) \log \left( \frac{\nu_h}{2} \right) + \left( \frac{\nu_h}{2} + p - 1 \right) \log (w_j) \right. \left. - \log \Gamma \left( \frac{\nu_h}{2} \right) - \frac{w_j}{2} \nu_h \right],$$

$$L_{3c} (\Psi) = \sum_{h=1}^g \sum_{j=1}^n z_{hj} \left\{ -p \log (2\pi) - \frac{1}{2} \log |\Sigma_h| \right. \left. - \frac{w_j}{2} \left[ D_h (y_j) + (u_j - q_{hj})^T A_h^{-1} (u_j - q_{hj}) \right] \right\},$$

and $D_h (y_j)$ is the determinant of the matrix $\Sigma_h$.
and where

\[ d_h(y_j) = (y_j - \mu_h)^T \Omega_h^{-1} (y_j - \mu_h), \]
\[ q_{hj} = \Delta_h^T \Omega_h^{-1} (y_j - \mu_h), \]
\[ \Lambda_h = I_p - \Delta_h^T \Omega_h^{-1} \Delta_h, \]
\[ \Omega_h = \Sigma_h + \Delta_h \Delta_h^T. \]

It is clear from (29) that maximization of the Q-function of the complete-data log likelihood (McLachlan and Krishnan, 2008),

\[ Q(\Psi; \Psi^{(k)}) = E_{\Psi^{(k)}} \{ \log L_c(\Psi) \mid y \}, \]

only requires maximization of the components functions \( L_{hc}(\Psi) \) separately \( (h = 1, 2, 3) \). The necessary conditional expectations involved in computing the Q-function with respect to (30) are, namely,

\[ \tau_{hj}^{(k)} = E_{\Psi^{(k)}} \{ Z_{hj} \mid y_j \}, \]
\[ e_{1,hj}^{(k)} = E_{\Psi^{(k)}} \{ \log(W_j) \mid y_j, z_{hj} = 1 \}, \]
\[ e_{2,hj}^{(k)} = E_{\Psi^{(k)}} \{ W_jU_j \mid y_j, z_{hj} = 1 \}, \]
\[ e_{3,hj}^{(k)} = E_{\Psi^{(k)}} \{ W_jU_j^T \mid y_j, z_{hj} = 1 \}, \]
\[ e_{4,hj}^{(k)} = E_{\Psi^{(k)}} \{ W_jU_jU_j^T \mid y_j, z_{hj} = 1 \}. \] (31)

The posterior probability of membership of the \( h \)th component by \( y_j \), using the current estimate \( \Psi^{(k)} \) for \( \Psi \), is given using Bayes’ Theorem by

\[ \tau_{hj}^{(k)} = \frac{\pi_h^{(k)} f_p \left( y_j; \mu_h^{(k)}, \Sigma_h^{(k)}, \delta_h^{(k)}, \nu_h^{(k)} \right)}{\sum_{h=1}^q \pi_h^{(k)} f_p \left( y_j; \mu_h^{(k)}, \Sigma_h^{(k)}, \delta_h^{(k)}, \nu_h^{(k)} \right)}. \] (32)

Let \( y_{hj}^{(k)} = q_{hj}^{(k)} \sqrt{\nu_h^{(k)} + p \nu_h^{(k)} + p} \). The other four expectations have analogous expressions to their one-component counterpart given in Section 3.

They are given by

\[ e_{1,hj}^{(k)} = \psi \left( \frac{\nu_h^{(k)} + p}{2} \right) - \log \left( \frac{\nu_h^{(k)} + d_{hj}^{(k)}}{2} \right) \] (33)

\[ T_{p,\nu_h^{(k)} + p} \left( q_{hj}^{(k)} \sqrt{\frac{\nu_h^{(k)} + p}{\nu_h^{(k)} + d_{hj}^{(k)}}(y_j)}; 0, \Lambda_h^{(k)} \right) S_{1,hj}^{(k)}, \]

\[ e_{2,hj}^{(k)} = \left( \frac{\nu_h^{(k)} + p}{\nu_h^{(k)} + d_{hj}^{(k)}}(y_j) \right) \frac{T_{p,\nu_h^{(k)} + p + 2} \left( q_{hj}^{(k)} \sqrt{\frac{\nu_h^{(k)} + p + 2}{\nu_h^{(k)} + d_{hj}^{(k)}}(y_j)}; 0, \Lambda_h^{(k)} \right)}{T_{p,\nu_h^{(k)} + p} \left( y_{hj}^{(k)}; 0, \Lambda_h^{(k)} \right)}, \] (34)

\[ e_{3,hj}^{(k)} = \left( \frac{\nu_h^{(k)} + p}{\nu_h^{(k)} + d_{hj}^{(k)}}(y_j) \right) \frac{T_{p,\nu_h^{(k)} + p} \left( y_{hj}^{(k)}; 0, \Lambda_h^{(k)} \right) S_{2,ij}^{(k)}}{T_{p,\nu_h^{(k)} + p} \left( y_{hj}^{(k)}; 0, \Lambda_h^{(k)} \right)}, \] (35)

\[ e_{4,hj}^{(k)} = \left( \frac{\nu_h^{(k)} + p}{\nu_h^{(k)} + d_{hj}^{(k)}}(y_j) \right) \frac{T_{-1}^{-1} \left( y_{hj}^{(k)}; 0, \Lambda_h^{(k)} \right) S_{3,ij}^{(k)}}{T_{p,\nu_h^{(k)} + p} \left( y_{hj}^{(k)}; 0, \Lambda_h^{(k)} \right)}, \] (36)
where $S_{1,hj}^{(k)}$ is a scalar defined by

$$
S_{1,ij}^{(k)} = \int_{-\infty}^{g_{ij}^{(k)}} \int_{-\infty}^{g_{ij}^{(k)}} \cdots \int_{-\infty}^{g_{ij}^{(k)}} \log \left( 1 + \frac{s^T \Lambda_h^{-1} s}{\nu_h^{(k)} + d_h^{(k)}(y_j)} \right) \left[ 1 + \frac{s^T \Lambda_h^{-1} s}{\nu_h^{(k)} + d_h^{(k)}(y_j)} \right]^{-\frac{d_h^{(k)}}{2} + p} du,
$$

(37)

$S_{2,hj}^{(k)}$ is a $p \times 1$ vector whose $r$th element is

$$
\int_0^\infty \int_0^\infty \cdots \int_0^\infty u_r t_{p,\nu_h^{(k)}+p+2} \left( u; q_{hj}^{(k)}, \left( \nu_h^{(k)} + d_h^{(k)}(y_j) \right) \Delta_h^{(k)} \right) du,
$$

(38)

and $S_{3,hj}^{(k)}$ is a $p \times p$ matrix whose $(r,s)$th element is

$$
\int_0^\infty \int_0^\infty \cdots \int_0^\infty u_r u_s t_{p,\nu_h^{(k)}+p+2} \left( u; q_{hj}^{(k)}, \left( \nu_h^{(k)} + d_h^{(k)}(y_j) \right) \Delta_h^{(k)} \right) du,
$$

(39)

where, for convenience of notation, $du$ is used to denote $du_1, du_2, \ldots, du_p$.

It is important to note that $S_{2,hj}^{(k)}$ and $S_{3,hj}^{(k)}$ are related to the first two moments of a truncated $p$-dimensional $t$-variate $X$. More specifically, let

$$
X \sim t_{p,\nu_h+p+2} \left( q_{hj}^{(k)}, \left( \nu_h^{(k)} + d_h^{(k)}(y_j) \right) \Delta_h^{(k)}, \mathbb{R}^+ \right),
$$

the truncated $t$-distribution as defined by (38). Then

$$
S_{2,hj}^{(k)} = T_{p,\nu_h^{(k)}+p+2} \left( q_{hj}^{(k)}, 0, \left( \nu_h^{(k)} + d_h^{(k)}(y_j) \right) \Delta_h^{(k)} \right) E(X),
$$

and

$$
S_{3,hj}^{(k)} = T_{p,\nu_h^{(k)}+p+2} \left( q_{hj}^{(k)}, 0, \left( \nu_h^{(k)} + d_h^{(k)}(y_j) \right) \Delta_h^{(k)} \right) E(XX^T),
$$

which can be implicitly expressed in terms of the parameters $q_{hj}^{(k)}$, $d_h^{(k)}(y_j)$, $\Delta_h^{(k)}$, $\nu_h^{(k)}$ using results (7) and (8). It is worth emphasizing that computation of $e_{3,hj}^{(k)}$ and $e_{4,hj}^{(k)}$ depends on algorithms for evaluating the multivariate $t$-distribution function, for which fast procedures are available.

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

**E-step**: Given $\Psi = \Psi^{(k)}$, compute $r_{hj}^{(k)}$ using (32), and $e_{1,hj}^{(k)}$, $e_{2,hj}^{(k)}$, $e_{3,hj}^{(k)}$, and $e_{4,hj}^{(k)}$ as described by (33), (34), (35), and (36) respectively, for $h = 1, \ldots, g$ and $j = 1, \ldots, n$.

**M-step**: Update the estimate of $\Psi$ by calculating for $h = 1, \ldots, g$, the following estimates of the parameters in $\Psi$,
\[ \mu_h^{(k)} = \frac{\sum_{j=1}^{n} \tau_{hj}^{(k)} \left( e_{2,hj}^{(k)} y_j - \Delta_h^{(k)} e_{3,hj}^{(k)} \right)}{\sum_{j=1}^{n} \tau_{hj}^{(k)} e_{2,hj}^{(k)}}, \]

\[ \delta_h^{(k+1) \pm} = \text{DIAG} \left\{ \left[ \sum_{j=1}^{n} \tau_{hj}^{(k)} \left( y_j - \mu_h^{(k)} \right) e_{3,hj}^{(k)T} \right] \left[ \sum_{j=1}^{n} \tau_{hj}^{(k)} e_{4,hj}^{(k)} \right]^{-1} \right\}, \]

and

\[ \Sigma_h^{(k+1)} = \frac{1}{\sum_{j=1}^{n} \tau_{hj}^{(k)}} \sum_{j=1}^{n} \tau_{hj}^{(k)} \left[ \Delta_h^{(k)} e_{3,hj}^{(k)T} \Delta_h^{(k+1)T} - \left( y_j - \mu_h^{(k+1)} \right) e_{3,hj}^{(k)T} \Delta_h^{(k+1)} - \Delta_h^{(k+1)} e_{3,hj}^{(k)} \left( y_j - \mu_h^{(k+1)} \right)^T + \left( y_j - \mu_h^{(k+1)} \right) \left( y_j - \mu_h^{(k+1)} \right)^T e_{2,hj}^{(k)} \right]. \]

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

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

A program for implementing this EM algorithm has been written in R.

6 The Empirical Information Matrix

We consider an approximation to the asymptotic covariance matrix of the ML estimates using the inverse of the empirical information matrix (Basford et al., 1997). The empirical information matrix is given by

\[ I_e(\Psi; y) = \sum_{j=1}^{n} s \left( y_j; \hat{\Psi} \right) s^T \left( y_j; \hat{\Psi} \right), \]

(40)

where \( s \left( y_j; \hat{\Psi} \right) = E_{\Psi} \left\{ \partial \log L_{cj}(\Psi) / \partial \Psi \mid y_j \right\} \) \((j = 1, \ldots, n)\) are the individual scores, consisting of

\[ (s_j, \pi_1, \ldots, s_j, \pi_g - 1, s_j, \mu_1, \ldots, s_j, \mu_g, \ldots, s_j, \Delta_g, \Sigma_1, \ldots, s_j, \Sigma_g, s_j, \nu_1, \ldots, s_j, \nu_g). \]

We let \( L_{cj}(\Psi) \) denote the complete-data log likelihood formed from the single observation \( y_j \). An estimate of the covariance matrix of \( \hat{\Psi} \) is given by taking the inverse of (40). After some algebraic manipulations, one can show that the elements of \( s \left( y_j; \hat{\Psi} \right) \) are given by the following explicit expressions:

\[ s_j, \pi_n = \frac{\tau_{hj}}{\pi_h} - \frac{\tau_{gj}}{\pi_g}, \]
More specifically, the fitted model can be written as:

\[
s_j,\mu_h = \tau_{hj} \Sigma_h^{-1} \left[ e_{2,ij} (y_j - \bar{\mu}_h) - \Delta_h e_{3,ij} \right],
\]
\[
s_j,\Sigma_h = \frac{1}{2} \tau_{hj} \left[ (y_j - \bar{\mu}_h) (y_j - \bar{\mu}_h)^T - (y_j - \bar{\mu}_h) e_{3,ij}^T \Delta_h - \Delta_h e_{3,hj} (y_j - \bar{\mu}_h) + \Delta_h e_{3,ij} (y_j - \bar{\mu}_h) + \delta_{hj} e_{4,hj}^T \Delta_h \right] \Sigma_h^{-1}
\]
\[
s_j,\Delta_h = \tau_{hj} \Sigma_h^{-1} \text{diag} \left[ (y_j - \bar{\mu}_h) e_{3,hj}^T - \Delta_h e_{4,hj} \right],
\]
\[
s_j,\nu_h = \frac{1}{2} \tau_{hj} \log \left( \frac{1}{2} \nu_h \right) + 1 + e_{1,hj} - \psi \left( \frac{1}{2} \nu_h \right) - e_{2,hj}.
\]

\section{Examples}

In this section, we fit the FM-MST model to three real data sets to demonstrate its usefulness in analyzing and clustering multivariate skewed data. In the first example, we focus on the flexibility of the FM-MST model in capturing the asymmetric shape of flow cytometric data. The next example illustrates the clustering capability of the model. In the final example, we demonstrate the computational efficiency of the proposed algorithm.

\subsection{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 1) is apparently bimodal and exhibits asymmetric pattern. Hence we fit a two-component FM-MST model to the data. More specifically, the fitted model can be written as

\[
f_2 (y_j; \Psi) = \pi_1 f_2 (y_j; \mu_1, \Sigma_1, \delta_1, \nu_1) + (1 - \pi_1) f_2 (y_j; \mu_2, \Sigma_2, \delta_2, \nu_2),
\]

where

\[
\mu_i = (\mu_{i,1}, \mu_{i,2})^T, \quad \Sigma_i = \begin{pmatrix} \sigma_{i,11} & \sigma_{i,12} \\ \sigma_{i,12} & \sigma_{i,22} \end{pmatrix}, \quad \delta_i = (\delta_{i,1}, \delta_{i,2})^T \quad (i = 1, 2).
\]

For comparison, we include the fitting of a two-component mixture of skew t-distributions from the Skew-normal Independent (SNI) family (Lachos, Ghosh, and Arellano-Valle, (2010)), hereafter named the FM-SNI-ST model. The estimated FM-SNI-ST density can be computed from the R package \texttt{mixsmsn} (Cabral, Lachos, and Prates, (2012)). Note that the MST distribution is different to the SNI-ST distribution since the skewing function is not of dimension one. Moreover, under the FM-SNI-ST settings, the correlation structure of \( Y \) will also be dependent on the skewness parameter, whereas for the FM-MST distributions the covariance structure is not affected by \( \delta \). The contours of the fitted SNI-ST and MST component densities are depicted in Fig 1(b) and Fig 1(c), respectively. To better visualize the shape of the fitted models, we display the estimated densities of each component instead of the mixture contours. It can be seen that the FM-MST model provides a noticeably better fit. From a clustering point of view, the FM-MST model also shows better performance as it is able to separate the two clusters correctly. Moreover, it adapts to the asymmetric shape of each cluster more adequately. Thus the superiority of FM-MST model is evident in dealing with asymmetric and heavily tailed data in this data set.
Figure 1: Mixture modelling of a reduced subset of prephosphorylation T cell population. Bi-variate skew t-mixtures were fitted to the data restricted in two dimensions CD45 and ZAP70. (a) Hue intensity plot of the Lymphoma data set; (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 fitted component contours of the two-component FM-MST model.

7.2 GvHD Data

Our second example concerns a data set 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-MST model to a case sample with a population of 13773 cells. The data set is shown in Figure 2, 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 a restricted version of the FM-MST model introduced in Section 2.1 (equation 5), hereafter denoted by FM-RMST.

We compare the performance of the three models FM-MST, 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. A similar strategy was described in Lin (2010).

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 highest value.

As anticipated, the optimal clustering result was given by the FM-MST model. It achieved the lowest misclassification rate. The FM-SNI-ST model has a higher number of misallocations. The FM-RMST model has a disappointing performance in terms of clustering. Its error rate is almost double that of its competitors. It is worth pointing out that both the FM-MST and FM-RMST models have 99 free parameters, while the FM-SNI-ST model has 95 parameters.
Figure 2: GvHD data set: Expert manual 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).

Table 1: Clustering error rates for various multivariate skew t mixture models on the GvHD data set.

| Model     | Error rate | Number of free parameters |
|-----------|------------|----------------------------|
| FM-MST    | 0.10807    | 99                         |
| FM-SNI-ST | 0.13078    | 95                         |
| FM-RMST   | 0.20700    | 99                         |

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-MST model.

7.3 AIS Data

We now illustrate the computational efficiency of our exact implementation of the E-step of the EM algorithm as in Section 5, where functions not in closed form are evaluated numerically. We denote this version of the EM algorithm with the 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. 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. As in Lin (2010), we limit this illustration to a bivariate subset of two variables – body mass index (BMI) and the percentage of body fat (Bfat). As noted by Lin (2010), these data are apparently bimodal. Hence a two-component mixture model is fitted to the data set.

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
Table 2: Computation time and clustering error rates for two different implementations of the EM algorithm for the multivariate skew $t$ mixture models on the AIS data set. For EM-exact, the E-step is implemented exactly as described in Section 5. As an alternative, the EM algorithm was implemented with a Monte Carlo E-step, EM-MC, as in Lin (2010). Time is measured in seconds.

| Component | EM-exact | EM-MC |
|-----------|----------|-------|
|           |          |       |
| $\pi$     | 0.58     | 0.59  |
| $\mu_{11}$| 19.95    | 19.89 |
| $\mu_{21}$| 15.82    | 15.50 |
| $\Sigma_{1,11}$| 3.83     | 2.96  |
| $\Sigma_{1,12}$| 6.19     | 6.17  |
| $\Sigma_{2,22}$| 25.60    | 25.80 |
| $\delta_{1}$| 3.55     | 2.72  |
| $\delta_{2}$| 1.82     | 2.22  |
| $\nu$     | 17.65    | 23.98 |
| $L(\Psi)$ | -1078.937| -1088.066 |
| AIC       | 2191.837 | 2207.956 |
| BIC       | 2248.114 | 2264.197 |
| error rate| 0.08910891 | 0.08910891 |
| time      | 45.49    | 349.9 |

criterion (BIC) (Schwarz, 1978) defined by

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

respectively, where $L(\Psi)$ is the value of the log likelihood at $\Psi$, $m$ 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 best value from each criterion are highlighted in bold font in Table 2. For this illustration, the EM-MC E-step is undertaken with 50 random draws as recommended by Lin (2010). The gender of each individual in this data set is also recorded, thus enabling us to evaluate the error rate of binary classification for the two methods. Both algorithms were terminated after the 23rd iteration, yielding the same misclassification rate of 0.0891.

Not surprisingly, the model selection criteria favour the EM-exact algorithm. Not only did it achieve lower AIC and BIC values, the computation time is remarkably lower than its competitor. It is more than seven times faster than the EM-MC alternative.

8 Computation Time and Accuracy for E-step

We now proceed to two interesting experiments for evaluating the computational cost and accuracy of using the EM-exact and EM-MC algorithms on high-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 $T_p(\cdot)$ functions, $p$ evaluations of $T_{p-1}(\cdot)$, and $\frac{1}{2} p(p - 1)$ evaluations of $T_{p-2}(\cdot)$. 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 from a Brain Tumor dataset supplied by Geoff Osborne from the Queensland Brain Institute at the University of Queensland. In both experiments we varied the dimensionality $p$ of the sample. The graph in Figure 3(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 3, EM-exact is the fastest among the four versions of the E-step for low dimensions. For example, at $p = 2$, EM-exact at least eight 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, 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 is 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 18 times quicker. Furthermore, EM-exact also has an additional advantage over the EM-MC alternative in that its results are reproducible.

To compare the accuracy of the EM-exact and EM-MC algorithms, we compute the total absolute error against the baseline EM-exact with a maximum tolerance of $10^{-18}$. 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 3(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-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, especially for applications involving high dimensional data.

9 Concluding Remarks

We have described an exact EM algorithm for evaluating the parameters of a general multivariate skew t-mixture model. This model has a more general characterization than various alternative versions of the skew t-distribution available in the literature and hence offers greater flexibility in capturing the asymmetric shape of skewed data.

Our proposed method is based on reduced analytical expressions for the E-step conditional expectations, which can be formulated in terms of the first and second moments of a multivariate truncated t-distribution. The latter can then be expressed further in terms of the distribution function of the multivariate central t-distribution for which fast algorithms capable of producing highly accurate results already exist. It is demonstrated 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 numerical approach, the version of the algorithm with a Monte Carlo E-step would require a large number of draws, which would be computationally expensive.

Acknowledgments

This work is supported by a grant from the Australian Research Council. Also, we would like to thank Dr Kui (Sam) Wang for his many helpful discussions on this topic.

References

[1] H. Akaike. A new look at the statistical model identification. *Automatic Control*, 19:716–723, 1974.

[2] R.B. Arellano-Valle, H. Bolvarine, and V.H. Lachos. Bayesian inference for skew-normal linear mixed models. *Journal of Applied Statistics*, 34(6):663–682, 2007.

[3] A. Azzalini. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12:171–178, 1985.

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

[5] J. D. Banfield and A. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, 49:803–821, 1993.

[6] K. E. Basford, D. R. Greenway, G. J. McLachlan, and D. Peel. Standard errors of fitted means under normal mixture. *Computational Statistics*, 12:1–17, 1997.

[7] R. M. Basso, V. H. Lachos, C. R. B. Cabral, and P. Ghosh. Robust mixture modeling based on scale mixtures of skew-normal distributions. *Computational Statistics and Data Analysis*, 54:2926–2941, 2010.
[8] M. D. Branco and D. K. Dey. A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis*, 79:99–113, 2001.

[9] R. Brinkman, M. Gaspareto, S.-J. Lee, A. Ribickas, J. Perkins, W. Janssen, R. Smiley, and C. Smith. 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, 2007.

[10] C.S. Cabral, V.H. Lachos, and M.O. Prates. Multivariate mixture modeling using skew-normal independent distributions. *Computational Statistics and Data Analysis*, 56:126–142, 2012.

[11] A.P. Dempster, N. M. Laird, and D. Rubin. Maximum likelihood from incomplete data via the e.m algorithm. *Journal of Royal Statistical Society, Series B*, 39:1–38, 1977.

[12] B. S. Everitt and D. J. Hand. *Finite Mixture Distributions*. Chapman and Hall, London, 1981.

[13] C. Fraley and A. E. Raftery. How many clusters? which clustering methods? answers via model-based cluster analysis. *Computer Journal*, 41:578–588, 1999.

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

[15] P. J. Green. On use of the e.m algorithm for penalized likelihood estimation. *Journal of the Royal Statistical Society B*, 52:443–452, 1990.

[16] A. K. Gupta. Multivariate skew-t distribution. *Statistics*, 37:359–363, 2003.

[17] S. Kotz and S. Nadarajah. *Multivariate t Distributions and Their Applications*. Cambridge University Press, Cambridge, 2004.

[18] V. H. Lachos, P. Ghosh, and R. B. Arellano-Valle. Likelihood based inference for skew normal independent linear mixed models. *Statistica Sinica*, 20:303–322, 2010.

[19] T. I. Lin. Robust mixture modeling using multivariate skew t distribution. *Statistics and Computing*, 20:343–356, 2010.

[20] T. I. Lin, J. C. Lee, and W. J. Hsieh. Robust mixture modeling using the skew-t distribution. *Statistics and Computing*, 17:81–92, 2007.

[21] T. I. Lin, J. C. Lee, and S. Y. Yen. Finite mixture modelling using the skew normal distribution. *Statistica Sinica*, 17:909–927, 2007.

[22] B. G. Lindsay. *Mixture Models: Theory, Geometry, and Applications*. NSF-CBMS Regional Conference Series in probability and Statistics, Volume 5, Institute of Mathematical Statistics, Hayward, CA, 1995.

[23] L. M. Maier, D. E. Anderson, P. L. De Jager, L.S. Wicker, and D. A. Hafler. Allelic variant in ctla4 alters t cell phosphorylation patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 104:18607–18612, 2007.

[24] G. J. McLachlan and K. E. Basford. *Mixture Models: Inference and Applications*. Marcel Dekker, New York, 1988.
[25] G. J. McLachlan and T. Krishnan. The EM Algorithm and Extensions. Wiley-Interscience, Hokoben, N. J., 2nd edition, 2008.

[26] G. J. McLachlan and D. Peel. Finite Mixture Models. Wiley Series in Probability and Statistics, 2000.

[27] G.J. McLachlan and D. Peel. Robust cluster analysis via mixtures of multivariate $t$-distributions. In A. Amin, D. Dori, P. Pudil, and H. Freeman, editors, Lecture Notes in Computer Science, volume 1451, pages 658–666. Springer-Verlag, Berlin, 1998.

[28] A. O’Hagan. Bayes estimation of a convex quadratic. Biometrika, 60:565–571, 1973.

[29] A. O’Hagan. Moments of the truncated multivariate-$t$ distribution. 1976.

[30] S. Pyne, X. Hu, K. Wang, E. Rossin, T.-I. Lin, L. M. Maier, C. Baecher-Allan, G. J. McLachlan, P. Tamayo, D. A. Hafler, P. L. De Jager, and J. P. Mesirow. Automated high-dimensional flow cytometric data analysis. PNAS, 106:8519–8524, 2009.

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

[32] G. Schwarz. Estimating the dimension of a model. Annals of Statistics, 6:461–464, 1978.

[33] D. M. Titterington, A. F. M. Smith, and U. E. Markov. Statistical Analysis of Finite Mixture Distributions. Wiley, New York, 1985.