Multivariate skew t mixture models: Applications to fluorescence-activated cell sorting data

Kui Wang
Department of Mathematics
University of Queensland
St. Lucia, Q4072, Australia
Email: kwang@maths.uq.edu.au

Shu-Kay Ng
School of Medicine
Griffith University (Logan Campus)
Meadowbrook, Q4131, Australia
Email: s.ng@griffith.edu.au

Geoffrey J McLachlan
Department of Mathematics
University of Queensland
St. Lucia, Q4072, Australia
Email: g.mclachlan@uq.edu.au

***Abstract*—In many applied problems in the context of pattern recognition, the data often involve highly asymmetric observations. Normal mixture models tend to overfit when additional components are included to capture the skewness of the data. Increased number of pseudo-components could lead to difficulties and inefficiencies in computations. Also, the contours of the fitted mixture components may be distorted. In this paper, we propose to adopt mixtures of multivariate skew t distributions to handle highly asymmetric data. The EM algorithm is used to compute the maximum likelihood estimates of model parameters. The method is illustrated using a flurorescence-activated cell sorting data.**

***Keywords*-Asymmetric multivariate data; EM algorithm; fluorescene-activated cell sorting; mixture models; skewed t.**

I. Introduction

Finite mixture models have been extensively developed and widely applied to density estimation and pattern recognition problems [1], [2], [3], [4]. With this approach to pattern recognition, the observed p-dimensional feature vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are assumed to have come from a mixture of a finite number, say g, of groups in some unknown proportions $\pi_1, \ldots, \pi_g$ that sum to one. That is, each feature vector $\boldsymbol{y}_j$ is taken to be a realization of the mixture probability density function (p.d.f.) defined by

$$f(\boldsymbol{y}; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i f(\boldsymbol{y}_j; \boldsymbol{\theta}_i), \tag{1}$$

where $f(\boldsymbol{y}_j; \boldsymbol{\theta}_i)$ denotes the ith component density with unknown parameter vector $\boldsymbol{\theta}_i$ $(i = 1, \ldots, g)$. The component distributions are usually specified to belong to the same parametric family. Here the vector $\boldsymbol{\Psi}$ of unknown parameters consists of the mixing proportions $\pi_1, \ldots, \pi_{g-1}$ and the elements of the $\boldsymbol{\theta}_i$ known *a priori* to be distinct. The fitting of finite mixture models (1) can be obtained by maximum likelihood via the expectation-maximization (EM) algorithm of Dempster, Laird, and Rubin [5]; see also [6]. Frequently, in practice, it is reasonable to consider fitting mixtures of elliptically symmetric component densities. Within this class of component densities, the multivariate normal density is a convenient choice given its computational tractability [1]. In applications where the tails of the normal distribution are shorter than appropriate or the parameter estimates are affected by atypical observations (outliers), the fitting of mixtures of multivariate t-distributions provides a more robust approach to the fitting of normal mixture models [7].

The t component density with location parameter $\boldsymbol{\mu}_i$, positive-definite matrix $\boldsymbol{\Sigma}_i$, and ν_i degrees of freedom is given by

$$t_p(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \nu_i) = \frac{\Gamma(\frac{\nu_i + p}{2})|\boldsymbol{\Sigma}_i|^{-1/2}}{(\pi\nu_i)^{\frac{1}{2}p}\Gamma(\frac{\nu_i}{2})\{1 + \delta(\boldsymbol{y}_j, \boldsymbol{\mu}_i; \boldsymbol{\Sigma}_i)/\nu_i\}^{\frac{1}{2}(\nu_i + p)}}, \tag{2}$$

where

$$\delta(\boldsymbol{y}_j, \boldsymbol{\mu}_i; \boldsymbol{\Sigma}_i) = (\boldsymbol{y}_j - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{y}_j - \boldsymbol{\mu}_i)$$

denotes the Mahalanobis squared distance between $\boldsymbol{y}_j$ and $\boldsymbol{\mu}_i$ (with $\boldsymbol{\Sigma}_i$ as the covariance matrix), and where the superscript T denotes vector transpose. As ν_i tends to infinity, $\boldsymbol{Y}_j$ becomes marginally multivariate normal with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. Therefore, the parameter ν_i can be viewed as a robustness tuning parameter, which can be inferred from the data by computing its maximum likelihood estimate. The application of the EM algorithm for maximum likelihood estimation of a mixture of multivariate t distributions is described in McLachlan and Peel [1] and the references therein.

In many applied problems in the context of pattern recognition, the contours of the fitted mixture models based on symmetric normal or t components are often distorted when the data involve highly asymmetric observations. In particular, the normal (or t) mixture model tends to overfit and produce many spurious clusters when additional components are required to capture the skewness and asymmetry in the feature data [8]. Including such spurious and irrelevant components may induce computational problems and difficulties in interpretation of results, which can further lead to invalid

inferences being made. The multivariate skew normal and skew t distributions have been proposed to fit asymmetric data in various applied problems [9], [10], [11], [12]. However, the extension of these multivariate skew distributions to a mixture model framework is not straightforward because of the complexity involved in the use of the EM algorithm to compute the maximum likelihood estimates of the model parameters. Mixture models of skew distributions have been therefore limited to univariate data [13], [14]. In this paper, we consider the extension to mixtures of multivariate skew t distributions for fitting highly asymmetric multivariate data. A variant of the EM algorithm is developed to compute the maximum likelihood estimates of model parameters. The method is illustrated using a flurorescence-activated cell sorting data.

The paper is organized as follows: Section II introduces the multivariate skew t mixture model and describes the EM algorithm for the iterative computation of maximum likelihood estimates. With multivariate data, singularity problems may occur with the use of EM algorithm. In Section III, we develop a "singularity handling" procedure within the framework of the EM algorithm to handle singularity problems that may exist in the applications. The estimation of the degrees of freedom does not exist in closed form. In Section IV, we consider three different methods and compare their performances via a simulation study. The application of the proposed method to a fluorescence-activated cell sorting dataset is provided in Section V. Section VI ends the paper with some discussion.

II. Multivariate Skew t Mixture Model

A. Multivariate Skew t Distribution

The multivariate skew t distribution as used here can be characterized using a particular form of that given by Sahu, Dey, and Branco [15] for the case of the skew normal distribution. We let $\boldsymbol{D}$ be a p-dimensional vector of skew parameters, and suppose that

$$\begin{pmatrix} \boldsymbol{U}_0 \\ U \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu} \\ 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\Omega} & 0 \\ 0 & 1 \end{pmatrix} \frac{1}{w}\right),$$

where $w \sim \text{gamma}(\nu/2, \nu/2)$; see [1]. Then $\boldsymbol{Y} = \boldsymbol{D}|U| + \boldsymbol{U}_0$ defines a p-dimensional multivariate skew t distribution with its density function as

$$f(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{D}, \nu) = 2t_p(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) T_{p+\nu}\left(\frac{\boldsymbol{\xi}}{\sigma}\sqrt{\frac{\nu+p}{\nu+\delta}}\right), \quad (3)$$

where $\boldsymbol{\Sigma} = \boldsymbol{\Omega} + \boldsymbol{D}\boldsymbol{D}^T$, $\boldsymbol{\xi} = \boldsymbol{D}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$, $\sigma^2 = (1 - \boldsymbol{D}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{D})$, and $\delta = (\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$. Here $T_{p+\nu}(\cdot)$ is the cumulative distribution function of a univariate (central) t random variable with degress of freedom $(p + \nu)$.

For the multivariate skew t distribution (3), the mean and covariance matrix are derived similar to that in [15] as

$$E(\boldsymbol{Y}) = \boldsymbol{\mu} + \sqrt{\frac{\nu}{\pi}}\frac{\Gamma((\nu-1)/2)}{\Gamma(\nu/2)}\boldsymbol{D}$$

and

$$cov(\boldsymbol{Y}) = (\boldsymbol{\Omega} + \boldsymbol{D}\boldsymbol{D}^T)\frac{\nu}{\nu-2} - \frac{\nu}{\pi}\left[\frac{\Gamma((\nu-1)/2)}{\Gamma(\nu/2)}\right]^2 \boldsymbol{D}\boldsymbol{D}^T.$$

B. EM Algorithm for Multivariate Skew t Mixture Model

With reference to (1), the mixture p.d.f. with multivariate skew t component densities is given by

$$f(\boldsymbol{Y}; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i f(\boldsymbol{y}; \boldsymbol{\mu}_i, \boldsymbol{\Omega}_i, \boldsymbol{D}_i, \nu_i), \quad (4)$$

where $f(\boldsymbol{y}; \boldsymbol{\mu}_i, \boldsymbol{\Omega}_i, \boldsymbol{D}_i, \nu_i)$ is speicifed by (3). The vector of unknown parameters $\boldsymbol{\Psi}$ is estimated by maximum likelihood via the EM algorithm. Within the framework of the EM algorithm, the observed feature data vector $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T$ is viewed as being incomplete, as the associated component-label vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, are not available [1]. In this framework, where each $\boldsymbol{y}_j$ is conceptualized as having arisen from one of the components of the mixture model (4) being fitted, $\boldsymbol{z}_j$ is a g-dimensional vector with $z_{ij} = (\boldsymbol{z}_j)_i = 1$ or 0, according to whether $\boldsymbol{y}_j$ did or did not arise from the ith component $(i = 1, \ldots, g;\ j = 1, \ldots, n)$.

In the light of the above characteristics of the skew t distribution (3), it is convenient to view the observed data augmented by the $\boldsymbol{z}_j$ as still being incomplete and introduce the additional missing data, $u_1, \ldots, u_n$ and $w_1, \ldots, w_n$. The complete-data vector is therefore given by

$$\boldsymbol{y}_c = (\boldsymbol{y}_{c1}^T, \ldots, \boldsymbol{y}_{cn}^T)^T,$$

where $\boldsymbol{y}_{c1} = (\boldsymbol{y}_1^T, \boldsymbol{z}_1^T, u_1, w_1)^T$, $\ldots$, and $\boldsymbol{y}_{cn} = (\boldsymbol{y}_n^T, \boldsymbol{z}_n^T, u_n, w_n)^T$ are assumed to be independently and identically distributed with $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ being independent realizations from a multinomial distribution consisting of one draw on g categories with respective probabilities $\pi_1, \ldots, \pi_g$. For this specification, the complete-data log likeihood can be written as

$$\log L_c(\boldsymbol{\Psi}) = \log L_{c1}(\boldsymbol{\pi}) + \log L_{c2}(\boldsymbol{\theta}) + \log L_{c3}(\boldsymbol{\nu}), \quad (5)$$

where

$$\log L_{c1}(\boldsymbol{\pi}) = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij}\log(\pi_i),$$

$$\log L_{c2}(\boldsymbol{\theta}) = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij}\{-\tfrac{1}{2}[p\log(2\pi) + \log|\boldsymbol{\Omega}_i| + w_j(\boldsymbol{y}_j - \boldsymbol{\mu}_i - \boldsymbol{D}_i u_j)^T\boldsymbol{\Omega}_i^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu}_i - \boldsymbol{D}_i u_j)]\},$$

and

$$\log L_{c3}(\boldsymbol{\nu}) = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij}\{-\tfrac{1}{2}[(p-1)\log(w_j)+w_j u_j^2]-$$

$$\tfrac{\nu_i}{2}[w_j - \log(\nu_i/2)] - \log\Gamma(\nu_i/2) + (\nu_i/2-1)\log(w_j)\}.$$

In (5), $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_g)^T$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T,\ldots,\boldsymbol{\theta}_g^T)^T$, and $\boldsymbol{\nu} = (\nu_1,\ldots,\nu_g)^T$, where $\boldsymbol{\theta}_i$ contains the elements of $\boldsymbol{\mu}_i$, the distinct elements of $\boldsymbol{\Omega}_i$ and $\boldsymbol{D}_i$ $(i = 1,\ldots,g)$.

The EM algorithm is a broadly applicable approach to the iterative computation of maximum likelihood estimates [6]. On the $(k+1)$th iteration of the EM algorithm, the E-step computes the conditional expectation of the above complete-data log likelihood $\log L_c(\boldsymbol{\Psi})$ given the observed data and the current estimates. This involves the calculations of the following five conditional expectations:

$$\tau_{ij}^{(k)} = E_{\boldsymbol{\Psi}^{(k)}}(\boldsymbol{Z}_{ij}|\boldsymbol{y}_j),$$

$$e_{1,ij}^{(k)} = E_{\boldsymbol{\Psi}^{(k)}}(W_j|\boldsymbol{y}_j, z_{ij}=1),$$

$$e_{2,ij}^{(k)} = E_{\boldsymbol{\Psi}^{(k)}}(U_j W_j|\boldsymbol{y}_j, z_{ij}=1),$$

$$e_{3,ij}^{(k)} = E_{\boldsymbol{\Psi}^{(k)}}(U_j^2 W_j|\boldsymbol{y}_j, z_{ij}=1),$$

and

$$e_{4,ij}^{(k)} = E_{\boldsymbol{\Psi}^{(k)}}(\log W_j|\boldsymbol{y}_j, z_{ij}=1),$$

where the expectations are based on the current value $\boldsymbol{\Psi}^{(k)}$ for $\boldsymbol{\Psi}$. In particular,

$$\tau_{ij}^{(k)} = \frac{\pi_i^{(k)} f(\boldsymbol{y}_j;\boldsymbol{\mu}_i^{(k)},\boldsymbol{\Omega}_i^{(k)},\boldsymbol{D}_i^{(k)})}{\sum_h^g \pi_h^{(k)} f(\boldsymbol{y}_j;\boldsymbol{\mu}_h^{(k)},\boldsymbol{\Omega}_h^{(k)},\boldsymbol{D}_h^{(k)})}$$

is the posterior probability that the jth feature vector $\boldsymbol{y}_j$ belongs to the ith component of the mixture (4). An outright partition of feature data into g non-overlapping clusters is achieved by assigning each feature vector to the component to which it has the highest estimated posterior probability of belonging [1]. The other four conditional expectations can be obtained according to [8].

On the M-step at the $(k+1)$th iteration of the EM algorithm, it follows from (5) that $\boldsymbol{\pi}^{(k+1)}$, $\boldsymbol{\theta}^{(k+1)}$, and $\boldsymbol{\nu}^{(k+1)}$ can be computed independently of each other. The solutions for $\pi_i^{(k+1)}$ and $\boldsymbol{\theta}_i^{(k+1)}$ exist in closed form. Only the update $\nu_i^{(k+1)}$ for the degrees of freedom ν_i need to be computed iteratively. That is,

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} \tau_{ij}^{(k)}/n, \tag{6}$$

$$\boldsymbol{\mu}_i^{(k+1)} = \sum_{j=1}^{n} \tau_{ij}^{(k)}(\boldsymbol{y}_j e_{1,ij}^{(k)} - \boldsymbol{D}_i^{(k)} e_{2,ij}^{(k)})/\sum_{j=1}^{n}(\tau_{ij}^{(k)} e_{1,ij}^{(k)}), \tag{7}$$

$$\boldsymbol{\Omega}_i^{(k+1)} = \sum_{j=1}^{n} \tau_{ij}^{(k)}\{e_{1,ij}^{(k)}(\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k)})(\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k)})^T -$$

$$e_{2,ij}^{(k)}\boldsymbol{S}_{ij}^{(k)}\boldsymbol{S}_{ij}^{(k)T} + e_{3,ij}^{(k)}\boldsymbol{D}_i^{(k)}\boldsymbol{D}_i^{(k)T}\}/\sum_{j=1}^{n}\tau_{ij}^{(k)}, \tag{8}$$

$$\boldsymbol{D}_i^{(k+1)} = \sum_{j=1}^{n} \tau_{ij}^{(k)} e_{2,ij}^{(k)}(\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k)})/\sum_{j=1}^{n}\tau_{ij}^{(k)} e_{3,ij}^{(k)}, \tag{9}$$

and

$$\sum_{j=1}^{n} \tau_{ij}^{(k)}[\log(\nu_i^{(k+1)}/2) - \psi(\nu_i^{(k+1)}/2) + 1] +$$

$$\sum_{j=1}^{n} \tau_{ij}^{(k)}(e_{4,ij}^{(k)} - e_{1,ij}^{(k)}) = 0, \tag{10}$$

where in (8), $\boldsymbol{S}_{ij}^{(k)} = \boldsymbol{D}_i^{(k)}(\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k)})^T$, and $\psi(s) = \{\partial\Gamma(s)/\partial s\}/\Gamma(s)$ is the Digamma function in (10).

The E- and M-steps are alternated repeatedly until the likelihood changes by an arbitrarily small amount in the case of convergence of the sequence of likelihood values [6].

III. Singularity Problem in EM algorithm

A singularity problem occurs in a few circumstances with the use of EM algorithm [6]. In the present study involving multivariate data, we may encounter a singularity problem in two occasions when the component-scale matrices are unconstrained. The first one is known as the collapse cluster problem, which is present when the feature data are lying in almost a lower dimensional subspace. For example, in a two dimensional plane, if some data pile up at its boundary line and are separated from other feature data. The variance of this cluster will be singular because the feature data are in fact one dimensional. The second occasion is the empty cluster problem, where a component converges to a cluster containing only a few data points relatively close together. The variance of this cluster will also tend to be singular. An example of two-dimensional feature data is given in Fig. 1, where one of three clusters (say, Cluster A) contains a set of data points lying on the line $y = -4$.

To handle the singularity problem, we add a singularity handling procedure within the framework of the EM algorithm. Before performing the E-step at each EM iteration, the covariance matrices $\boldsymbol{\Omega}_i$ are checked for singularity or being close to singularity (very small determinant). Those covariance matrices that are singular will be re-defined by first determining in which coordinates the covariance matrix is degenerated. The corresponding diagonal elements are changed to a small pre-defined value ε (say, $\varepsilon = 0.0001$), and other elements at the same column and row are changed to zero. The re-definition of a marginal distribution for those coordinates that lead to degenerated covariance matrix will give a higher posterior probability belonging to the particular

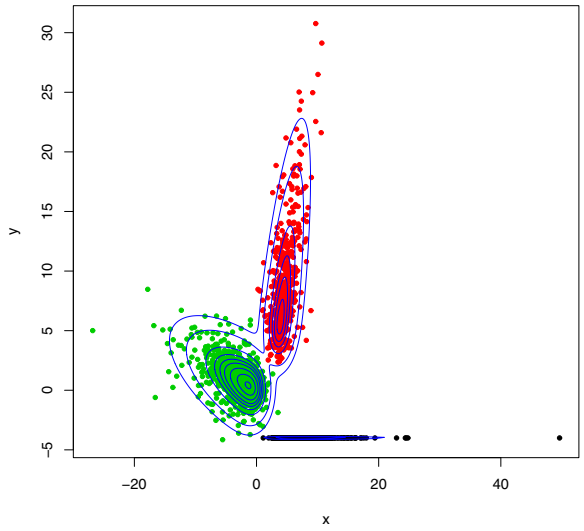

Figure 1. An example data from a three-component skew t mixture with a collapse cluster

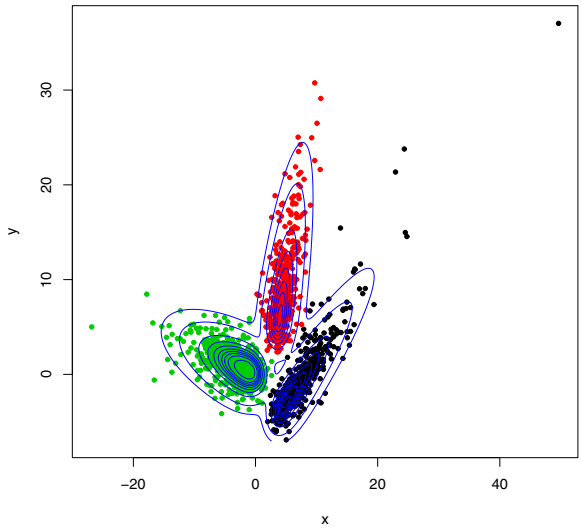

Figure 2. A simulated dataset from a three-component skew t mixture

cluster. Similarly, the corresponding elements of the skew parameter vector $\boldsymbol{D}_i$ will be set to zero. On the M-step at each iteration, the singularity handling procedure will also check whether the expected number of feature data is less than two in each cluster, $\sum_{j=1}^{n} \tau_{ij}^{(k)}$, for $i = 1, \ldots, g$. Clusters containing only a few data points will have their means and covariance matrices re-defined to be zero vector and matrix, respectively. The skew parameter vector $\boldsymbol{D}_i$ will set to be a zero vector as well. The empty cluster problem will therefore be handled by the singularity check on the next E-step as mentioned above.

With the above three-cluster example (Fig. 1), it is the second coordinate that leads to degenerated covariance matrix for Cluster A. The singularity handling procedure will re-define its covariance matrix on the E-step as

$$\begin{pmatrix} \sigma_1^2 & 0 \\ 0 & 0 \end{pmatrix} \rightarrow \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \varepsilon \end{pmatrix}.$$

Thus, the data points with its second coordinate value equal to -4 will have a higher posterior probability belonging to Cluster A than those points with same first coordinate but second coordinate not equal to -4. The second element of $\boldsymbol{D}_1$ will be set to be zero as well.

IV. Estimation of the degrees of freedom

As mentioned in Section II (Equation (10)), the updated estimate of the degrees of freedom ν_i does not exist in closed form. Here we consider three methods for its computation. With the first method (Method 1), an approximation to the term $e_{4,ij}^{(k)}$ on the right-hand-side of (10) is adopted; see the supplementary information of [8]. Method 2 attempts to obtain the exact values at each stage of the iterative process, where the term $e_{4,ij}^{(k)}$ is calculated by truncating an infinite series expansion of it. With Method 3, the default pre-defined values for the degrees of freedom $\nu_i = 4$ $(i = 1, \ldots, g)$ are adopted. A simulation study has been conducted to compare the three methods.

We generate 100 datasets from a three-component skew t mixture model. For each dataset, there are 1000 two-dimensional samples, in which 300 samples come from the first component, 300 from the second, and 400 from the third. One example dataset is plotted in Fig. 2. With this simulation study, true parameter values are used as the initial values. The EM algorithm proceeds until the relative change of log likelihood is less than 0.0001, or until it reaches the maximum number of iterations 100, whichever first occurs. The total computing time, the bias, and mean square errors (MSEs) for each method are summarised in Table I.

From Table I, it can be observed that Method 2 takes the longest computing time (more than twelve times of Method 1 and fifteen times of Method 3). Both Methods 1 and 2 overestimate the degrees of freedom ν_i, and are unbiased in the estimation of $\boldsymbol{\theta}_i$ $(i = 1, 2, 3)$.

V. An Example

A flurescence-activated cell sorting (FACS) dataset is used as an illustration. Flurescence-activated cell sorting machines yield readout on large mixed populations of single cells, with around $n = 5,000-50,000$ cells from blood samples of individuals with multiple sclerosis and other diseases. The flurescence intensities of tagged antibodies are measured

Table I
COMPARISON OF THREE METHODS FOR ESTIMATING THE DEGREES OF FREEDOM

	Method 1 Bias (MSE)	Method 2 Bias (MSE)	Method 3 Bias (MSE)
1st cluster			
μ_1	-0.0023 (0.109)	-0.0026 (0.111)	0.0066 (0.136)
μ_2	-0.0117 (0.132)	-0.0125 (0.132)	-0.0131 (0.133)
σ_{11}	0.0076 (0.148)	0.0072 (0.148)	0.0023 (0.149)
σ_{12}	-0.0062 (0.110)	-0.0060 (0.110)	-0.0080 (0.112)
σ_{22}	-0.0064 (0.150)	-0.0058 (0.150)	-0.0139 (0.135)
D_1	0.0024 (0.123)	0.0032 (0.127)	-0.0113 (0.157)
D_2	0.0093 (0.152)	0.0101 (0.153)	0.0118 (0.151)
ν	0.1785 (0.589)	0.1881 (0.617)	NA
π_1	-0.0010 (0.007)	-0.0010 (0.007)	-0.0005 (0.006)
2nd cluster			
μ_1	-0.0052 (0.119)	-0.0059 (0.121)	-0.0018 (0.126)
μ_2	-0.0095 (0.103)	-0.0093 (0.104)	-0.0125 (0.116)
σ_{11}	0.0077 (0.163)	0.0081 (0.163)	0.0034 (0.157)
σ_{12}	-0.0068 (0.105)	-0.0068 (0.105)	-0.0026 (0.102)
σ_{22}	0.0003 (0.132)	0.0003 (0.132)	-0.0042 (0.124)
D_1	0.0107 (0.150)	0.0114 (0.153)	0.0046 (0.157)
D_2	-0.0118 (0.123)	-0.0117 (0.127)	-0.0066 (0.138)
ν	0.0602 (0.573)	0.0626 (0.577)	NA
π_1	0.0016 (0.007)	0.0016 (0.007)	0.0010 (0.007)
3rd cluster			
μ_1	-0.0283 (0.135)	-0.0289 (0.135)	-0.0281 (0.136)
μ_2	0.0139 (0.095)	0.0144 (0.097)	0.0139 (0.109)
σ_{11}	-0.0112 (0.127)	-0.0104 (0.125)	-0.0133 (0.118)
σ_{12}	0.0171 (0.083)	0.0177 (0.083)	0.0162 (0.084)
σ_{22}	-0.0242 (0.142)	-0.0239 (0.142)	-0.0272 (0.145)
D_1	0.0181 (0.148)	0.0192 (0.148)	0.0168 (0.146)
D_2	-0.0206 (0.116)	-0.0220 (0.122)	-0.0199 (0.133)
ν	0.0917 (0.530)	0.1005 (0.517)	NA
π_1	-0.0006 (0.006)	-0.0006 (0.006)	-0.0006 (0.005)
Computing time	59.95 seconds	773.8 seconds	49.36 seconds

by the scanners and are reported as multi-dimensional points (in general 4 or 8 colour FACS corresponding to different markers). The objective here is to cluster the blood cells on the basis of multivariate FACS data with an attempt to detect important subpopulations of regulatory cells. However, the FACS data are also multimodal, asymmetric, and have many outliers. Thus, the fitting of multivariate normal or t mixture models for the analysis of FACS data often generates distorted contours and may lead to invalid conclusions.

The dataset was captured using a BD Biosciences FACS Calibur system [16]. It consists of $n = 4952$ blood cells samples stained for 4 markers (CD4, CD56, CD8, and CD3). The proposed multivariate skew t mixture model is fitted to the 4-dimensional data with $g = 2$ to $g = 15$ components. Based on the Bayesian information criterion (BIC) for model selection [1], [17], we identify there are eight clusters of blood cells. The pairwise two-dimensional contours of the fitted skew t mixture model are presented in Fig. 3. The contour lines indicate the asymmetric nature of the FACS data, such as the cluster indicated by blue coloured dots in the graph CD56 against CD4, and the cluster indicated by black coloured dots in the graph CD8 against CD56. In addition, it can be seen from Fig. 3 that the model fits well

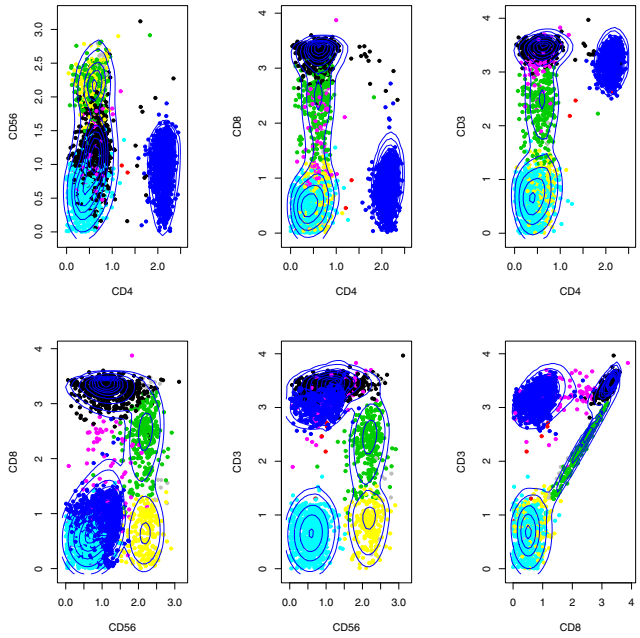

Figure 3. Two-dimensional contours of the fitted skew t mixture model

to the data. The clusters so formed and their estimated model parameters may be used to infer various disease signatures in FACS samples and match corresponding cell populations across samples that enables quantitative downstream analysis, such as classification and prediction of clinically relevant phenotypes [8].

VI. DISCUSSION

We have developed mixtures of multivariate skew t distributions to handle highly asymmetric multivariate data. A singularity handling procedure has been considered to solve singularity problems within the framework of the EM algorithm. We also compare three methods for estimating the degrees of freedom for the component-t distributions. The proposed method has been applied to a real flurorescence-activated cell sorting dataset.

An alternative method has been considered recently to handle asymmetric data [18]. Their method transforms the asymmetric data via a Box-Cox transformation to minimize skewness of the data. Symmetric t distributions are then adopted to model the transformed data. In contrast, our proposed multivariate skew t mixture approach models directly the asymmetric populations and hence gets to understand the distinctive shape and location of each sub-population. These estimated parameters may further be employed to identify distinctive features that are relevant to predict disease outcomes.

Multivariate skew t mixture modelling can be readily applied to other important pattern recognition problems. For

example, it can offer insights into FACS experiment design by detecting redundant (i.e. less informative) and discriminative (i.e. more informative) antibody profiles. Applications of the model can be found in wide areas of scientific fields where (multivariate) data exhibit a mixture of asymmetric patterns with atypical observations; see, for example, [9], [10], [11].

With applications of mixture models, the likelihood equation will have multiple roots corresponding to local maxima. The EM algorithm described in Section II.B should be applied from a wide choice of initial values in any search for all local maxima [6]. The intent is to choose as the maximum likelihood estimate of the parameter vector $\mathbf{\Psi}$ the local maximizer corresponding to the largest of the local maxima located [1]. In practice, consideration has to be given to the problem of relatively large local maxima that occur as a consequence of a fitted component having a very small (but nonzero) generalized variance (the determinant of the covariance matrix). Such a component corresponds to a cluster containing a few data points either relatively close together or almost lying in a lower-dimensional subspace. As described in Section III, a singularity handling procedure is required to identify the spurious local maximizers [1].

ACKNOWLEDGMENT

The work was supported by grants from the Australian Research Council and the University of Queensland, Australia.

REFERENCES

[1] G. J. McLachlan and D. Peel, *Finite Mixture Models*. New York: Wiley, 2000.

[2] S. K. Ng and G. J. McLachlan, "Speeding up the EM algorithm for mixture model-based segmentation of magnetic resonance images," *Pattern Recognition*, vol. 37, pp. 1573–1589, 2004.

[3] S. K. Ng, G. J. McLachlan, K. Wang, L. Ben-Tovim Jones, and S. W. Ng, "A mixture model with random-effects components for clustering correlated gene-expression profiles," *Bioinformatics*, vol. 22, pp. 1745–1752, 2006.

[4] K. Wang, K. K. W. Yau, and A. H. Lee, "A hierarchical Poisson mixture regression model to analyse maternity length of hospital stay," *Stat. Med.*, vol. 21, pp. 3639–3654, 2002.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm (with discussion)," *J. Roy. Stat. Soc. B*, vol. 39, pp. 1–38, 1977.

[6] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, 2nd ed. New Jersey: Wiley, 2008.

[7] G. J. McLachlan and D. Peel, "Robust cluster analysis via mixtures of multivariate t-distributions," *Lect. Notes Computer Science*, vol. 1451, pp. 658–666, 1998.

[8] 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. Mesirov, "Automated high-dimensional flow cytometric data analysis," *Proc. National Acad. Sciences USA*, vol. 106, pp. 8519–8524, 2009.

[9] A. Azzalini and A. Capitanio, "Statistical applications of the multivariate skew-normal distribution," *J. Roy. Stat. Soc. B*, vol. 61, pp. 579–602, 1999.

[10] A. Azzalini and A. Capitanio, "Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution," *J. Roy. Stat. Soc. B*, vol. 65, pp. 367–389, 2003.

[11] A. Azzalini and A. Dalla Valle, "The multivariate skew-normal distribution," *Biometrika*, vol. 83, pp. 715–726, 1996.

[12] A. K. Gupta, "Multivariate skew t-distribution," *Statistics*, vol. 37, pp. 359–363, 2003.

[13] T. I. Lin, J. C. Lee, and W. Hsieh, "Robust mixture modeling using the skew t distribution," *Stat. Comp.*, vol. 17, pp. 81–92, 2007.

[14] T. I. Lin, J. C. Lee, and S. Y. Yen, "Finite mixture modeling using the skew normal distribution," *Stat. Sinica*, vol. 17, pp. 909–927, 2007.

[15] S. K. Sahu, D. K. Dey, and M. D. Branco, "A new class of multivariate skew distributions with applications to Bayesian regression models," *Canadian J. Stat.*, vol. 31, pp. 129–150, 2003.

[16] L. M. Maier, D. E. Anderson, P. L. De Jager, L. S. Wicker, D. A. Hafler, "Allelic variant in CTLA4 alters T cell phosphorylation patterns," *Proc. National Acad. Sciences USA*, vol. 104, pp. 18607–18612, 2007.

[17] G. Schwarz, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6, pp. 461–464, 1978.

[18] K. Lo, R. R. Brinkman, and R. Gottardo, "Automated gating of flow cytometry data via robust model-based clustering," *Cytometry A*, vol. 73, pp. 321–332, 2008.