Stat Comput (2012) 22:33–52
DOI 10.1007/s11222-010-9204-1

Flexible mixture modeling via the multivariate *t* distribution with the Box-Cox transformation: an alternative to the skew-*t* distribution

Kenneth Lo · Raphael Gottardo

Received: 12 August 2009 / Accepted: 14 September 2010 / Published online: 8 October 2010

Abstract Cluster analysis is the automated search for groups of homogeneous observations in a data set. A popular modeling approach for clustering is based on finite normal mixture models, which assume that each cluster is modeled as a multivariate normal distribution. However, the normality assumption that each component is symmetric is often unrealistic. Furthermore, normal mixture models are not robust against outliers; they often require extra components for modeling outliers and/or give a poor representation of the data. To address these issues, we propose a new class of distributions, multivariate *t* distributions with the Box-Cox transformation, for mixture modeling. This class of distributions generalizes the normal distribution with the more heavy-tailed *t* distribution, and introduces skewness via the Box-Cox transformation. As a result, this provides a unified framework to simultaneously handle outlier identification and data transformation, two interrelated issues. We describe an Expectation-Maximization algorithm for parameter estimation along with transformation selection. We demonstrate the proposed methodology with three real data sets and simulation studies. Compared with a wealth of approaches including the skew-*t* mixture model, the proposed *t* mixture model with the Box-Cox transformation performs favorably in terms of accuracy in the assignment of observations, robustness against model misspecification, and selection of the number of components.

Keywords Box-Cox transformation · Clustering · EM algorithm · Outliers · Robustness · Skewness

1 Introduction

In statistics, model-based clustering (McLachlan 1982; Titterington et al. 1985; McLachlan and Basford 1988; Banfield and Raftery 1993; Celeux and Govaert 1995; McLachlan and Peel 2000; Fraley and Raftery 2002; McLachlan et al. 2003; Bouveyron et al. 2007; McNicholas and Murphy 2008; Scrucca 2010; Andrews and McNicholas 2010) is a popular unsupervised approach to look for homogeneous groups of observations. The most commonly used model-based clustering approach is based on finite normal mixture models, which has been shown to give good results in various applied fields, for example, gene expression (Yeung et al. 2001; McLachlan et al. 2002; Pan et al. 2002), image analysis (Wehrens et al. 2004; Fraley et al. 2005; Li et al. 2005), medical diagnosis (Schroeter et al. 1998; Forbes et al. 2006) and astronomy (Kriessler and Beers 1997; Mukherjee et al. 1998).

However, normal mixture models rely on the assumption that each component follows a normal distribution, which is often unrealistic for data with groups asymmetric in shape (Lin et al. 2007b; Lo et al. 2008; Lin 2009a, 2010). A common remedy for the asymmetry issue is to look for transformations of the data that make the normality assumption more realistic. Box and Cox (1964) discussed the power transformation in the context of linear regression, which has

Electronic supplementary material The online version of this article (doi:10.1007/s11222-010-9204-1) contains supplementary material, which is available to authorized users.

K. Lo
Department of Microbiology, University of Washington, Seattle, WA, USA

R. Gottardo (✉)
Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
e-mail: rgottard@fhcrc.org

also been applied to normal mixture models (Schork and Schork 1988; Gutierrez et al. 1995).

Another line of attempts to resolve the asymmetry observed in data is to enhance the flexibility of the normal distribution by introducing skewness. Azzalini (1985) developed a class of univariate skew-normal distributions with the introduction of a shape parameter to account for skewness, which had been put to use in a mixture modeling context by Lin et al. (2007b). A multivariate version of the skew-normal distributions was first proposed by Azzalini and Dalla Valle (1996), with various generalizations or modifications ensuing. One such modification was found in Sahu et al. (2003), who developed a new class of multivariate skew elliptically symmetric distributions with applications to Bayesian regression models, and included the multivariate skew-normal distribution as a special case. As opposed to Azzalini and Dalla Valle's (1996) formulation of the skew-normal distribution, the correlation structure in that of Sahu et al. (2003) is not affected by the introduction of skewness in the sense that independence between elements of a random vector is preserved irrespective of changes in the skewness parameters. The latter formulation was adopted by Lin (2009a), who introduced the multivariate skew-normal mixture model, and described a variant of the Expectation-Maximization (EM) algorithm (Dempster et al. 1977), called the Expectation-Constrained-Maximization (ECM) algorithm (Meng and Rubin 1993), for maximum likelihood estimation. However, the implementation of this methodology is fairly computationally intensive. A simplified version of Sahu et al.'s (2003) formulation has recently been suggested by Pyne et al. (2009), who parameterized skewness in the form of a vector in place of a matrix. As a result of this simplification, the computational complexity of parameter estimation has been reduced considerably.

In addition to non-normality, there is also the problem of outlier identification in mixture modeling. Outliers can have a significant effect on the resulting clustering. For example, they will usually lead to overestimating the number of components in order to provide a good representation of the data (Fraley and Raftery 2002). If a more robust model is used, fewer clusters may suffice. Outliers can be handled in the model-based clustering framework, by either replacing the normal distribution with a more robust one (e.g., t; see Peel and McLachlan 2000; McLachlan and Peel 2000) or adding an extra component to accommodate the outliers (e.g., uniform; see Schroeter et al. 1998).

Transformation selection and outlier identification are two issues which can have heavy mutual influence (Carroll 1982; Atkinson 1988). While a stepwise approach in which transformation is preselected ahead of outlier detection (or vice versa) may be considered, it is unlikely to address the problem well, as the preselected transformation may be influenced by the presence of outliers. One possible means of handling the two issues simultaneously is through the application of skew-t distributions (Azzalini and Capitanio 2003; Sahu et al. 2003) in mixture modeling. Such an attempt was given by Lin et al. (2007a), who proposed a skew-t mixture model based on the formulation of Azzalini and Capitanio (2003), but it is confined to the univariate case. Not until recently has a multivariate version of the skew-t mixture model come to light. Lin (2010) and (Pyne et al. 2009) adopted a similar approach to the case of skew-normal in defining the multivariate skew-t distribution, thereby simplifying Sahu et al.'s (2003) formulation with a vector in place of a skewness matrix.

In view of the aforementioned issues, we propose a unified framework based on mixture models using a new class of skewed distributions, namely, multivariate t distributions with the Box-Cox transformation, to handle transformation selection and outlier identification simultaneously. The t distribution provides a robust mechanism against outliers with its heavier tails relative to the normal distribution (Lange et al. 1989). The Box-Cox transformation is a type of power transformation, which can bring skewed data back to symmetry, a property of both the normal and t distributions. Along with the introduction of the mixture model using this new class of distributions, we also describe a convenient means of parameter estimation via the EM algorithm. Whilst the proposed framework is computationally much simpler than mixture modeling using skew-t distributions, it performs well in various scenarios compared to a wealth of competing approaches, as shown in subsequent sections of this article. A simplified form of our proposed framework has been applied to flow cytometry, which shows a favorable performance in identifying cell populations (Lo et al. 2008). This article presents a comprehensive framework that substantially enriches that previous simplified version, including the selection of component-specific transformations, and the provision of a level of robustness adaptive to the data. In addition, the emphasis is laid upon computational development of the proposed methodology. We have also included a large-scale comparison with competing approaches such as those using the skew-normal or skew-t mixture distributions.

The structure of this article is as follows. In Sect. 2 we first introduce the new class of skewed distributions, multivariate t distributions with the Box-Cox transformation. Then we introduce the mixture model using the proposed distributions, and present details including outlier identification, density estimation and the selection of the number of components. In addition, we describe an EM algorithm to simultaneously handle parameter estimation and transformation selection for our proposed mixture model. In Sect. 3, the performance of the proposed framework is examined on real data sets and compared to a wealth of commonly used approaches. Section 4 presents extensive simulation studies

to further evaluate our proposed framework relative to the other approaches. Finally, in Sect. 5 we summarize and discuss our findings.

2 Methodology

2.1 Preliminaries

2.1.1 The multivariate t distribution

The multivariate t distribution has found its use as a robust modeling tool in various fields of applied statistics like linear and non-linear regression, time series, and pedigree analysis; see Lange et al. (1989) and Kotz and Nadarajah (2004) for examples. The t distribution is applied in place of the normal distribution when the latter fails to offer long enough tails for the error distribution. Formally, a random vector $\mathbf{y}$ of length p is said to follow a p-dimensional multivariate t distribution with mean $\boldsymbol{\mu}$ ($\nu > 1$), covariance matrix $\nu(\nu-2)^{-1}\boldsymbol{\Sigma}$ ($\nu > 2$) and ν degrees of freedom if its density function is given by

$$\varphi_p(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma(\frac{\nu+p}{2})|\boldsymbol{\Sigma}|^{-1/2}}{(\pi\nu)^{p/2}\Gamma(\frac{\nu}{2})\{1+(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})/\nu\}^{\frac{\nu+p}{2}}}. \quad (1)$$

The degrees of freedom ν may be viewed as a robustness tuning parameter, as it controls the fatness of the tails of the distribution. When $\nu \to \infty$, the t distribution approaches a p-dimensional multivariate normal distribution with mean $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$, and density function

$$\phi_p(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\}. \quad (2)$$

An account of the development of the maximum likelihood estimation of the multivariate t distribution can be found in Liu and Rubin (1995), Liu (1997) and Peel and McLachlan (2000). The estimation involves the use of the EM algorithm or its variants including the ECM and ECME (Liu and Rubin 1994) algorithms. The crux of these algorithms constitutes the fact that we can parameterize a t distribution using a normal-gamma compound distribution. The degrees of freedom ν may be jointly estimated along with other unknown parameters, or fixed *a priori* when the sample size is small. In the latter case, setting $\nu = 4$ has been found to provide good protection against outliers and work well in many applications (see, for example, Lange et al. 1989; Stephens 2000).

2.1.2 Box-Cox transformation

The power transformation proposed by Box and Cox (1964) was originally introduced to make asymmetric data fulfill the normality assumption in a regression model. The Box-Cox transformation of an observation y is defined as follows:

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \log y, & \lambda = 0, \end{cases} \quad (3)$$

where λ is referred to as the transformation parameter. Note that this function is defined for positive values of y only. In view of the need to handle negative-valued data in some applications, we adopt a modified version (Bickel and Doksum 1981) of the Box-Cox transformation which is also defined for negative values:

$$y^{(\lambda)} = \frac{\operatorname{sgn}(y)|y|^\lambda - 1}{\lambda}, \quad \lambda > 0. \quad (4)$$

When all data values are positive, this modified Box-Cox transformation reduces to the original version.

2.2 The multivariate t distribution with the Box-Cox transformation

In this article, we propose a new class of distributions, namely, multivariate t distributions with the Box-Cox transformation (tBC), to handle transformation and to accommodate outliers simultaneously. Explicitly, a random vector $\mathbf{y}$ of length p following such a distribution has a density function specified by

$$\varphi_p(\mathbf{y}^{(\lambda)} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) \cdot |J_p(\mathbf{y}; \lambda)|, \quad (5)$$

where $|J_p(\mathbf{y};\lambda)| = |y_1^{\lambda-1}y_2^{\lambda-1}\cdots y_p^{\lambda-1}|$ is the Jacobian induced by the Box-Cox transformation. Equivalently, the random vector $\mathbf{y}$ follows a multivariate t distribution after being Box-Cox transformed. It is difficult to derive the exact mean and variance of the distribution in closed form. However, using first-order Taylor series expansion, approximations for the mean and covariance matrix can be derived. The mean can be approximated by a vector of length p with the j-th element being $\operatorname{sgn}(\lambda\mu_j+1)\,|\lambda\mu_j+1|^{1/\lambda}$, and the variance by $\nu/(\nu-2)D_p(\boldsymbol{\mu};\lambda)\boldsymbol{\Sigma}D_p(\boldsymbol{\mu};\lambda)$, where $D_p(\boldsymbol{\mu};\lambda)$ is a diagonal matrix of order p with the j-th diagonal element being $|\lambda\mu_j+1|^{1/\lambda-1}$. The various shapes that can be represented by the tBC are shown in Fig. 1.

Analogous to the case of the t distribution without transformation, the tBC approaches a multivariate normal distribution with the Box-Cox transformation (NBC) when $\nu \to \infty$. In addition, this class of distributions also includes the untransformed version of the multivariate t or normal

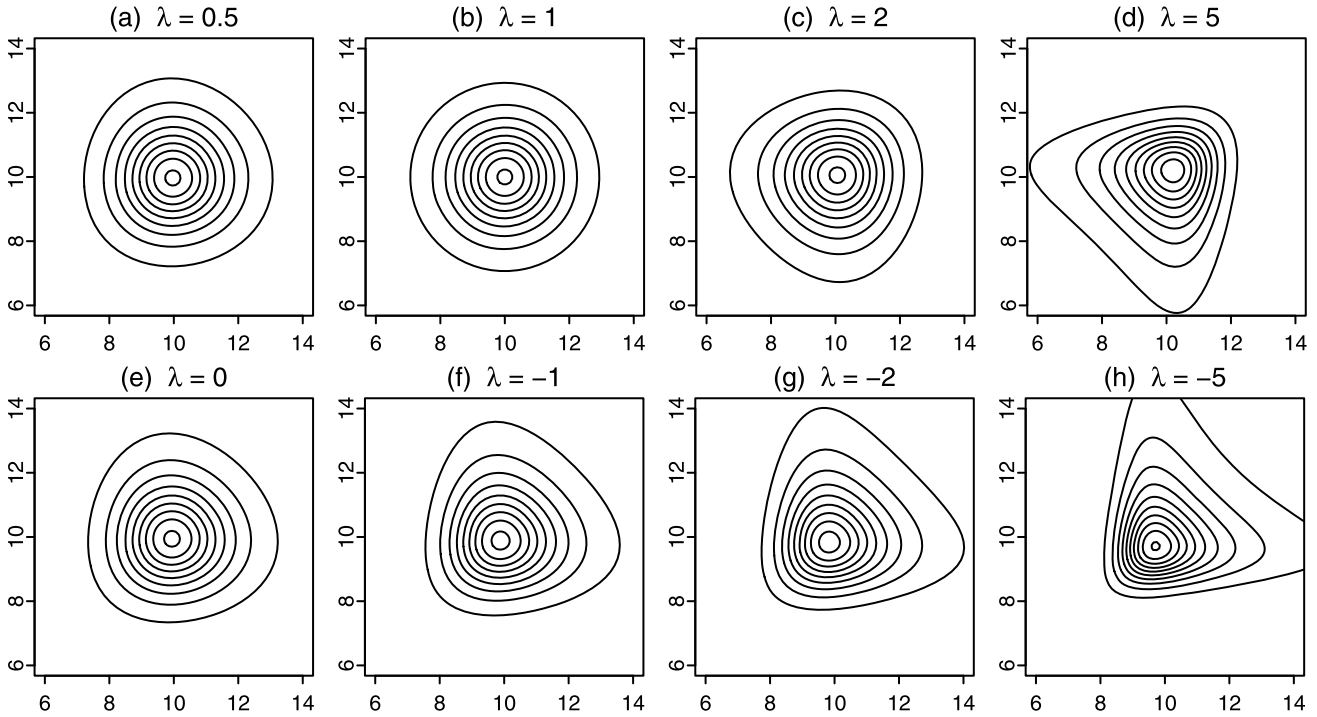

Fig. 1 Contour plots revealing the shape of bivariate t distributions with the Box-Cox transformation for different values of the transformation parameter. Each distribution has a mean of 10 and unit variance along each dimension. The degrees of freedom parameter is fixed at eight. The values of the transformation parameter λ range from -5 (extremely right-skewed) to 5 (extremely left-skewed)

distribution. The untransformed t or normal distribution is recovered by setting λ in (5) to one, although there is a translation of one unit to the left in each direction on the original scale (due to the term $-1/\lambda$ in (4)).

The flexible class of tBC offers robustness against both outliers and asymmetry observed in data. Comparatively, the t distribution alone is deemed robust in the sense that it offers a mechanism to accommodate outliers. As noted by Lange et al. (1989), however, the t distribution is not robust against asymmetric error distributions. When asymmetry is observed, data transformation is desired for the sake of restoring symmetry, and subsequently drawing proper inferences. The introduction of the tBC is therefore in line with the notion of Lange et al. (1989).

2.3 The mixture model of t distributions with the Box-Cox transformation

2.3.1 The model

Making use of the tBC introduced in the last subsection, we now define a G-component mixture model in which each component is described by a tBC. Given data $\mathbf{y}$, with independent p-dimensional observation vectors $\mathbf{y}_i, i = 1, \ldots, n$, the likelihood for the tBC mixture model is given as follows:

$$L(\boldsymbol{\Psi} \mid \mathbf{y}) = \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g \varphi_p(\mathbf{y}_i^{(\lambda_g)} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \nu_g) \cdot |J_p(\mathbf{y}_i; \lambda_g)|,$$

$$\sum_{g=1}^{G} \pi_g = 1. \tag{6}$$

The mixing proportion π_g is the probability that an observation belongs to the g-th component. Estimates of the unknown parameters $\boldsymbol{\Psi} = (\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_G)$ where $\boldsymbol{\Psi}_g = (\pi_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \nu_g, \lambda_g)$ can be obtained conveniently using the EM algorithm described in the next subsection. Analogous to the case of tBC, the mixture distribution approaches that of an NBC mixture model with $\varphi_p(\cdot \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \nu_g)$ being replaced by $\phi_p(\cdot \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ when $\nu_g \to \infty$ for all g. Also, the class of tBC mixture models includes the conventional, untransformed t or normal mixture model, obtained by fixing $\lambda_g = 1$ for all g. Note that a restricted form of (6) has been previously applied in Lo et al. (2008) to identify cell populations in flow cytometry data, on setting a global transformation parameter $\lambda = \lambda_g$ and fixing $\nu_g = 4$ for all g.

2.3.2 Maximum likelihood estimation

In this subsection we illustrate how transformation selection can be handled along with parameter estimation simultane-

ously via an EM algorithm. We first define two types of missing data, same as the ones used in the maximum likelihood estimation for the t mixture model described in Peel and McLachlan (2000). One is the unobserved component membership $\mathbf{z}_i = (z_{i1}, \ldots, z_{iG})$ with

$$z_{ig} = \begin{cases} 1 & \text{if } \mathbf{y}_i \text{ belongs to the } g\text{-th component,} \\ 0 & \text{otherwise,} \end{cases}$$

associated with each observation $\mathbf{y}_i$. Each vector $\mathbf{Z}_i$ follows independently a multinomial distribution with one trial and event properties $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_G)$, denoted as $\mathbf{Z}_i \sim \mathcal{M}_G(1, \boldsymbol{\pi})$. Another type of missing data is the weight u_i, coming from the normal-gamma compound parameterization for the t distribution, such that

$$\mathbf{Y}_i \mid u_i, z_{ig} = 1 \sim \mathrm{N}(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g/u_i) \tag{7}$$

independently for $i = 1, \ldots, n$, and $U_i \sim \mathrm{Ga}(\nu_g/2, \nu_g/2)$. The advantage of writing the model in this way is that, conditional upon the U_i's, the sampling errors are again normal but with different precisions, and estimation becomes a weighted least squares problem. The complete-data log-likelihood becomes

$$\begin{aligned} & l_c(\boldsymbol{\Psi} \mid \mathbf{y}, \mathbf{z}, \mathbf{u}) \\ &= \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig} \Big\{ \log\Big[\pi_g \phi_p(\mathbf{y}_i^{(\lambda_g)} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g/u_i) \\ &\quad \cdot |J_p(\mathbf{y}_i; \lambda_g)|\Big] + \log \mathrm{Ga}(u_i \mid \nu_g/2, \nu_g/2) \Big\} \\ &= \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig} \Big\{ \log \pi_g - \frac{p}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_g| \\ &\quad - \frac{u_i}{2}(\mathbf{y}_i^{(\lambda_g)} - \boldsymbol{\mu}_g)^T \boldsymbol{\Sigma}_g^{-1}(\mathbf{y}_i^{(\lambda_g)} - \boldsymbol{\mu}_g) \\ &\quad + (\lambda_g - 1)\sum_{j=1}^{p}\log|y_{ij}| + \frac{\nu_g}{2}\log\frac{\nu_g}{2} - \log\Gamma\left(\frac{\nu_g}{2}\right) \\ &\quad + \frac{\nu_g}{2}(\log u_i - u_i) + \left(\frac{p}{2} - 1\right)\log u_i \Big\}, \end{aligned} \tag{8}$$

where $\mathrm{Ga}(\cdot \mid \cdot)$ is the density function of u_i. The E-step of the EM algorithm involves the computation of the conditional expectation of the complete-data log-likelihood $E_{\boldsymbol{\Psi}}(l_c \mid \mathbf{y})$. To facilitate this, we need to compute $\tilde{z}_{ig} \equiv E_{\boldsymbol{\Psi}}(Z_{ig} \mid \mathbf{y}_i)$, $\tilde{u}_{ig} \equiv E_{\boldsymbol{\Psi}}(U_i \mid \mathbf{y}_i, z_{ig} = 1)$ and $\tilde{s}_{ig} \equiv E_{\boldsymbol{\Psi}}(\log U_i \mid \mathbf{y}_i, z_{ig} = 1)$:

$$\tilde{z}_{ig} \leftarrow \frac{\pi_g \varphi_p(\mathbf{y}_i^{(\lambda_g)} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \nu_g) \cdot |J_p(\mathbf{y}_i; \lambda_g)|}{\sum_{k=1}^{G} \pi_k \varphi_p(\mathbf{y}_i^{(\lambda_k)} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \nu_k) \cdot |J_p(\mathbf{y}_i; \lambda_k)|}, \tag{9}$$

$$\tilde{u}_{ig} \leftarrow \frac{\nu_g + p}{\nu_g + (\mathbf{y}_i^{(\lambda_g)} - \boldsymbol{\mu}_g)^T \boldsymbol{\Sigma}_g{}^{-1}(\mathbf{y}_i^{(\lambda_g)} - \boldsymbol{\mu}_g)} \tag{10}$$

and

$$\tilde{s}_{ig} \leftarrow \log \tilde{u}_{ig} + \psi\left(\frac{\nu_g + p}{2}\right) - \log\left(\frac{\nu_g + p}{2}\right), \tag{11}$$

where $\psi(\cdot)$ is the digamma function. Note that, if we assume a global transformation parameter λ, (9) used to compute $\tilde{z}_{ig}$ is slightly simplified as

$$\tilde{z}_{ig} \leftarrow \frac{\pi_g \varphi_p(\mathbf{y}_i^{(\lambda)} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \nu_g)}{\sum_{k=1}^{G} \pi_k \varphi_p(\mathbf{y}_i^{(\lambda)} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \nu_k)}. \tag{12}$$

As can be seen in the following, $\tilde{s}_{ig}$ only appears in (18) or (19) for the update of the degrees of freedom ν_g. If we fix ν_g to some predetermined value, then $\tilde{s}_{ig}$ is not needed and so the quantity in (11) does not need to be computed. Upon plugging $\tilde{z}_{ig}$, $\tilde{u}_{ig}$ and $\tilde{s}_{ig}$ into (8) for z_{ig}, u_i and $\log u_i$ respectively, we obtain the conditional expectation of the complete-data log-likelihood.

In the M-step, we update the parameter estimates with values which maximize the conditional expectation of the complete-data log-likelihood. The mixing proportions are updated with the following formula:

$$\hat{w}_g \leftarrow \frac{n_g}{n}, \tag{13}$$

where $n_g \equiv \sum_i \tilde{z}_{ig}$. The estimation of $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ needs to be considered along with the transformation parameter λ_g of the Box-Cox transformation. Closed-form solutions for $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ are available conditional on λ_g as follows,

$$\hat{\boldsymbol{\mu}}_g = \frac{\sum_{i=1}^{n} \tilde{z}_{ig}\tilde{u}_{ig}\mathbf{y}_i^{(\lambda_g)}}{\sum_{i=1}^{n} \tilde{z}_{ig}\tilde{u}_{ig}} = h_1(\lambda_g); \tag{14}$$

$$\begin{aligned} \hat{\boldsymbol{\Sigma}}_g &= \frac{\sum_{i=1}^{n} \tilde{z}_{ig}\tilde{u}_{ig}(\mathbf{y}_i^{(\lambda_g)} - \hat{\boldsymbol{\mu}}_g)(\mathbf{y}_i^{(\lambda_g)} - \hat{\boldsymbol{\mu}}_g)^T}{n_g} \\ &= h_2(\lambda_g). \end{aligned} \tag{15}$$

No closed-form solution is available for λ_g, but on substituting $\hat{\boldsymbol{\mu}}_g = h_1(\lambda_g)$ and $\hat{\boldsymbol{\Sigma}}_g = h_2(\lambda_g)$ into the conditional expectation of the complete-data log-likelihood for $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ respectively, the problem reduces to a one-dimensional search of λ_g. Explicitly, the optimization is recast as a one-dimensional root-finding problem of the equation $\partial E_{\boldsymbol{\Psi}}(l_c \mid \mathbf{y})/\partial\lambda_g = 0$, in which

$$\begin{aligned} & \frac{\partial E_{\boldsymbol{\Psi}}(l_c \mid \mathbf{y})}{\partial \lambda_g} \\ &= \frac{\partial}{\partial \lambda_g}\sum_{i=1}^{n} \tilde{z}_{ig} \Big\{ -\frac{\tilde{u}_{ig}}{2}\Big[(\mathbf{y}_i^{(\lambda_g)} - \boldsymbol{\mu}_g)^T \boldsymbol{\Sigma}_g^{-1}(\mathbf{y}_i^{(\lambda_g)} - \boldsymbol{\mu}_g)\Big] \\ &\quad + (\lambda_g - 1)\sum_{j=1}^{p}\log|y_{ij}| \Big\} \end{aligned}$$

$$= \sum_{i=1}^{n} \left[-\tilde{z}_{ig} \tilde{u}_{ig} (\mathbf{y}_i^{(\lambda_g)} - \boldsymbol{\mu}_g)^T \boldsymbol{\Sigma}_g^{-1} \right] \frac{\partial \mathbf{y}_i^{(\lambda_g)}}{\partial \lambda_g} + \sum_{i=1}^{n} \tilde{z}_{ig} \sum_{j=1}^{p} \log |y_{ij}|, \quad (16)$$

where $\partial \mathbf{y}_i^{(\lambda_g)} / \partial \lambda_g$ is a vector of length p whose j-th element is $\lambda_g^{-2}[\mathrm{sgn}(y_{ij})|y_{ij}|^{\lambda_g} (\lambda_g \log |y_{ij}| - 1) + 1]$, and $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ are replaced with $\hat{\boldsymbol{\mu}}_g = h_1(\lambda_g)$ and $\hat{\boldsymbol{\Sigma}}_g = h_2(\lambda_g)$ respectively. The equation may be solved numerically using, for example, Brent's (1973) algorithm. If we assume a global transformation parameter λ instead, the left hand side of the equation to consider is slightly modified from (16) as

$$\frac{\partial E_{\boldsymbol{\Psi}}(l_c \mid \mathbf{y})}{\partial \lambda} = \sum_{i=1}^{n} \left\{ \sum_{g=1}^{G} \left[-\tilde{z}_{ig} \tilde{u}_{ig} (\mathbf{y}_i^{(\lambda)} - \boldsymbol{\mu}_g)^T \boldsymbol{\Sigma}_g^{-1} \right] \right\} \frac{\partial \mathbf{y}_i^{(\lambda)}}{\partial \lambda} + \sum_{i=1}^{n} \sum_{j=1}^{p} \log |y_{ij}|. \quad (17)$$

Once a numerical estimate of λ_g has been obtained, we substitute it back into (14–15) to update $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ respectively.

To complete the M-step, we need to update the estimate of the degrees of freedom ν_g, unless it is fixed *a priori*. From (8), we see that there are no overlaps between terms involving $(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \lambda_g)$ and those involving ν_g. Hence, the incorporation of the Box-Cox transformation does not complicate the estimation of ν_g. Again, since there is no closed-form solution available for ν_g, we turn it into a one-dimensional root-finding problem by considering the equation $\partial E_{\boldsymbol{\Psi}}(l_c \mid \mathbf{y}) / \partial \nu_g = 0$, in which

$$\frac{\partial E_{\boldsymbol{\Psi}}(l_c \mid \mathbf{y})}{\partial \nu_g} = \frac{\partial}{\partial \nu_g} \sum_{i=1}^{n} \tilde{z}_{ig} \left\{ \frac{\nu_g}{2} \log \frac{\nu_g}{2} - \log \Gamma\left(\frac{\nu_g}{2}\right) + \frac{\nu_g}{2} (\tilde{s}_{ig} - \tilde{u}_{ig}) \right\} \propto n_g \left\{ \log \frac{\nu_g}{2} + 1 - \psi\left(\frac{\nu_g}{2}\right) \right\} + \sum_{i=1}^{n} \tilde{z}_{ig} (\tilde{s}_{ig} - \tilde{u}_{ig}). \quad (18)$$

If we assume a global degrees of freedom $\nu = \nu_g$ for all g, the derivative $\partial E_{\boldsymbol{\Psi}}(l_c \mid \mathbf{y}) / \partial \nu$ is given by

$$\frac{\partial E_{\boldsymbol{\Psi}}(l_c \mid \mathbf{y})}{\partial \nu} \propto n \left\{ \log \frac{\nu}{2} + 1 - \psi\left(\frac{\nu}{2}\right) \right\} + \sum_{i=1}^{n} \sum_{g=1}^{G} \tilde{z}_{ig} (\tilde{s}_{ig} - \tilde{u}_{ig}). \quad (19)$$

Alternatively, to improve the convergence, we may exploit the advantage of the ECME algorithm and switch to update ν by optimizing the constrained actual log-likelihood function:

$$\hat{\nu} \leftarrow \arg\max_{\nu} \left\{ \sum_{i=1}^{n} \log \left(\sum_{g=1}^{G} \pi_g \varphi_p(\mathbf{y}_i^{(\lambda_g)} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \nu) \cdot |J_p(\mathbf{y}_i; \lambda_g)| \right) \right\}. \quad (20)$$

Apart from an intuitive sense that a faster convergence is expected on disregarding the information of the parameter estimates obtained from the previous iteration (which is carried over by the conditional expectation of the complete-data log-likelihood otherwise) as well as considering the actual likelihood instead of its approximation, it also saves a little computational burden by circumventing the computation of $\tilde{s}_{ig}$.

The EM algorithm alternates between the E and M-steps until convergence. The quantity $\tilde{z}_{ig}$ may be interpreted as the posterior probability that observation $\mathbf{y}_i$ belongs to the g-th component. The maximum *a posteriori* configuration results from assigning each observation to the component associated with the largest $\tilde{z}_{ig}$ value. The uncertainty corresponding to each assignment may be conveniently quantified as $1 - \max_g \tilde{z}_{ig}$ (Bensmail et al. 1997).

2.3.3 Outlier identification

Just like the case of $\tilde{z}_{ig}$, the introduction of $\tilde{u}_{ig}$ does not only facilitate the implementation of the EM algorithm, but also aids in the interpretation of the final estimated model. As seen from (14–15), $\tilde{u}_{ig}$ serves as the weight in the weighted least squares estimation of $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$. It holds a negative relationship with the Mahalanobis distance $(\mathbf{y}_i^{(\lambda_g)} - \boldsymbol{\mu}_g)^T \boldsymbol{\Sigma}_g{}^{-1} (\mathbf{y}_i^{(\lambda_g)} - \boldsymbol{\mu}_g)$ between $\mathbf{y}_i$ and $\boldsymbol{\mu}_g$ on the transformed scale, as given by (10). Hence, a small value of $\tilde{u}_{ig}$ would suggest that the corresponding observation is an outlier, and diminish its influence on the estimation of the parameters. In contrast, in the absence of such a mechanism, a normal mixture model is not robust against outliers, as the constraint $\sum_g \tilde{z}_{ig} = 1$ for all i restricts all observations to make equal contributions overall towards parameter estimation.

Exploiting such a mechanism, we may conveniently set up a rule of calling an observation with the associated $\tilde{u}_{ig}$ value smaller than a threshold, say, 0.5, an outlier. Such a threshold may be selected on a theoretical basis by considering the one-to-one correspondence between $\tilde{u}_{ig}$ and the Mahalanobis distance which follows some standard, known distribution. On noting that

$$(\mathbf{y}_i^{(\lambda_g)} - \boldsymbol{\mu}_g)^T \boldsymbol{\Sigma}_g{}^{-1} (\mathbf{y}_i^{(\lambda_g)} - \boldsymbol{\mu}_g) / p \sim F(p, \nu_g), \quad (21)$$

where $\mathbf{y}_i^{(\lambda_g)}$ follows a p-dimensional t distribution with parameters $(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \nu_g)$ and $F(\cdot)$ denotes an F distribution, a threshold c for $\tilde{u}_{ig}$ may be determined by considering the desired threshold quantile level α of the distribution stated in (21):

$$c = \frac{\nu_g + p}{\nu_g + pF_{1-\alpha}(p, \nu_g)}, \tag{22}$$

where $F_{1-\alpha}(\cdot)$ denotes the α quantile of the F distribution such that $\Pr(F \geq F_{1-\alpha}) = 1 - \alpha$. For instance, if $\nu_g = 4$, $p = 5$, and the desired threshold quantile level is $\alpha = 0.9$, then the corresponding threshold for $\tilde{u}_{ig}$ is $c = 0.37$ given the 0.9 quantile $F_{0.1}(5, 4) = 4.051$. Any observation with the associated $\tilde{u}_{ig} < 0.37$ will be deemed an outlier.

From (10), we can also see how the degrees of freedom ν_g contributes to the robustness of the parameter estimation process. A smaller value of ν_g tends to downweight outliers to a greater extent, while a large enough value tends to regress all weights to one, approaching the case of the NBC model.

2.3.4 Density estimation

One advantage of mixture modeling based on the normal distribution is that the marginal distribution for any subset of the dimensions is also normally distributed with the mean and covariance matrix extracted from the conformable dimensions (Johnson and Wichern 2002). This favorable property is also observed in the multivariate t distribution (Liu and Rubin 1995; Kotz and Nadarajah 2004), making the estimation of the marginal density for any dimensions available at a very low computational cost. Consider the partition $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)$ as an example. If $\mathbf{Y}$ comes from a multivariate t distribution with ν degrees of freedom and with mean and covariance matrix conformably partitioned as

$$\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) \quad \text{and} \quad \frac{\nu}{\nu - 2}\boldsymbol{\Sigma} = \frac{\nu}{\nu - 2}\begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

respectively, then its subset $\mathbf{Y}_1$ will follow a t distribution with mean $\boldsymbol{\mu}_1$, covariance matrix $\nu/(\nu-2)\boldsymbol{\Sigma}_{11}$ and the same ν degrees of freedom. This nice property is easily extended to a t mixture model with more than one component, and, in addition, preserved in our proposed tBC mixture model. One can easily derive the marginal density by extracting the conformable partitions from the means, covariance matrices and the Jacobian. The 90th percentile region of the mixture components shown in Supplementary Fig. 1 (Online Resource 1) is produced by these means.

2.3.5 Selecting the number of components

When the number of mixture components is unknown, we apply the Bayesian Information Criterion (BIC) (Schwarz 1978) to guide the selection. The BIC provides a convenient approximation to the integrated likelihood of a model and, in the context of mixture models, is defined as

$$\text{BIC}_G = 2\log \tilde{L}_G - K_G \log n, \tag{23}$$

where $\tilde{L}_G$ is the likelihood value of (6) evaluated at the maximum likelihood estimates of $\boldsymbol{\Psi}$, and K_G is the number of independent parameters for a G-component mixture model. The BIC would then be computed for a range of possible values for G and the one with the largest BIC (or relatively close to it) would be selected. Although the asymptotic approximation of the integrated likelihood leading to the BIC depends on regularity conditions that are not satisfied in the context of mixture models, theoretical and practical justifications can be found in favor of the application of the BIC on model selection for mixture models. Leroux (1992) shows that the BIC will not underestimate the number of components asymptotically. Keribin (2000) shows that the BIC gives a consistent estimate of the number of components. Empirical evidence justifying the use of the BIC in the context of mixture models can be found in a wealth of literature; see, for example, Fraley and Raftery (1998, 2002), and more recently, McNicholas and Murphy (2008), Wang et al. (2009) and Andrews and McNicholas (2010).

Often, the BIC is applied in line with the principle of parsimony, by which we favor a simpler model if it does not incur a downgrade of the modeling performance. Suppose there are two tBC mixture models with G_1 and G_2 components respectively such that $G_1 < G_2$. Under the notion of this principle, we would prefer the simpler model, i.e., the one with G_1 components, unless a very strong evidence of improved performance signified by an increase of >10 (Kass and Raftery 1995; Fraley and Raftery 2002) is observed from BIC_{G_1} over BIC_{G_2}.

3 Application to real data

3.1 Data description

To illustrate our methodology we use the following real data sets.

3.1.1 The bankruptcy data set

This data set was obtained from a study which conducted financial ratio analysis to predict corporate bankruptcy (Altman 1968). The sample consists of 66 manufacturing firms in the United States, of which 33 are bankrupt and the other 33 solvent. The data collected include the ratio of retained earnings (RE) to total assets, and the ratio of earnings before interest and taxes (EBIT) to total assets. They were derived from financial statements released two years prior to bankruptcy, and statements from the solvent firms during the same period.

3.1.2 The crabs data set

Measurements in this data set were collected from a study of rock crabs of genus *Leptograpsus* (Campbell and Mahon 1974). The sample is composed of 50 crabs for each combination of species (blue and orange color forms) and sex (male and female), resulting in a total of 200 observations. There are five morphological measurements, namely, the frontal lobe size, the width of the rear region of the carapace, the length of the carapace along the midline, the maximum width of the carapace, and the depth of the body, for each crab.

3.1.3 The wine data set

To identify the discriminating factors from a set of chemical and physical characteristics to classify wines by type and origin, a study with a sample of 178 red wines grown in Piedmont, Italy was conducted (Forina et al. 1986). Three cultivars were represented in the sample: Barolo (59), Grignolino (71) and Barbera (48). For each wine 28 measurements were made. Here, we consider a 13-variable data subset available in the `gclus` R package (Hurley 2004); the 13 continuous measurements include alcohol, malic acid, ash, alkalinity of ash, magnesium, total phenols, flavanoids, nonflavanoid phenols, proanthocyanins, color intensity, hue, OD280/OD315 of diluted wines and proline.

3.2 Implementation

We compare the performance of six mixture modeling approaches using different mixture distributions, namely, t with the Box-Cox transformation (tBC), t, normal with the Box-Cox transformation (NBC), normal, skew-t, and skew-normal. Since all observations in the data sets come with known labels, we can assess and compare the models based on misclassification rates. For the bankruptcy and crabs data sets, we also take the number of components selected by the BIC as an assessment criterion. In the analysis of the wine data set, we examine an interesting problem that how dimension reduction refines the classification performance.

We first fit the data sets using the six aforementioned models in turn, on fixing the number of mixture components at the known values, i.e., two for the bankruptcy data set, four for the crabs data set, and three for the wine data set. The same initialization strategy is applied to the EM algorithm for all the models. Each time, 10 random partitions are generated. Following each partition formed, $\tilde{z}_{ig}$ will be assigned one or zero accordingly, and the model parameters are initialized using the formulae in the M-step. A few EM runs ensue, terminated at a premature tolerance level of 10^{-3} for the relative change in the likelihood values between two successive iterations. Out of the 10 random partitions, the one delivering the highest likelihood value after a few EM runs will be taken as the initial configuration for the eventual EM algorithm. Convergence of the EM algorithm is detected at a tolerance level of 10^{-6}. At convergence, misclassification rates, i.e., the proportions of observations assigned to the incorrect group, are computed. Each misclassification rate is determined as the minimum considering all permutations of the labels of the components.

To facilitate the comparison of their capability to select the correct number of components, when we apply the aforementioned models, we fit the data and compute the BIC once for each choice of the number of mixture components $G = 1, 2, \ldots, M$, where $M = 6$ for the bankruptcy data set and $M = 8$ for the crabs data set. These values are chosen for M because they are well above the true number of groups (two for the bankruptcy data set and four for the crabs data set) such that little change in the result is expected when we further increase M; numerical problems may arise when M is too large, moreover.

All results presented here were obtained with our software implementation of clustering for flow cytometry, flowClust (Lo et al. 2009), except for the skew-normal and skew-t mixture modeling that were obtained with the FLAME software (http://www.broadinstitute.org/cancer/software/genepattern/modules/FLAME/) developed by Pyne et al. (2009).

3.3 Results

3.3.1 Classification

Table 1 shows the misclassification rates for the different models. As can be seen, for the bankruptcy data set, the tBC and NBC mixture models deliver misclassification rates (15.2% and 16.7% respectively) lower than the other methods by a wide margin. By taking a graphical inspection of the results, we find that the poor classification performance of the other four methods is due to the inability to resolve the shape of the two groups of observations properly

Table 1 Misclassification rates for different models applied to the bankruptcy and crabs data sets

Model	Bankruptcy	Crabs
tBC	**0.152 (10)**	**0.070 (14)**
t	0.273 (18)	0.075 (15)
NBC	0.167 (11)	0.345 (69)
Normal	0.318 (21)	0.290 (58)
Skew-t	0.303 (20)	0.085 (17)
Skew-normal	0.394 (26)	0.175 (35)

The best results are shown in bold. The numbers of misclassified cases are given within parentheses

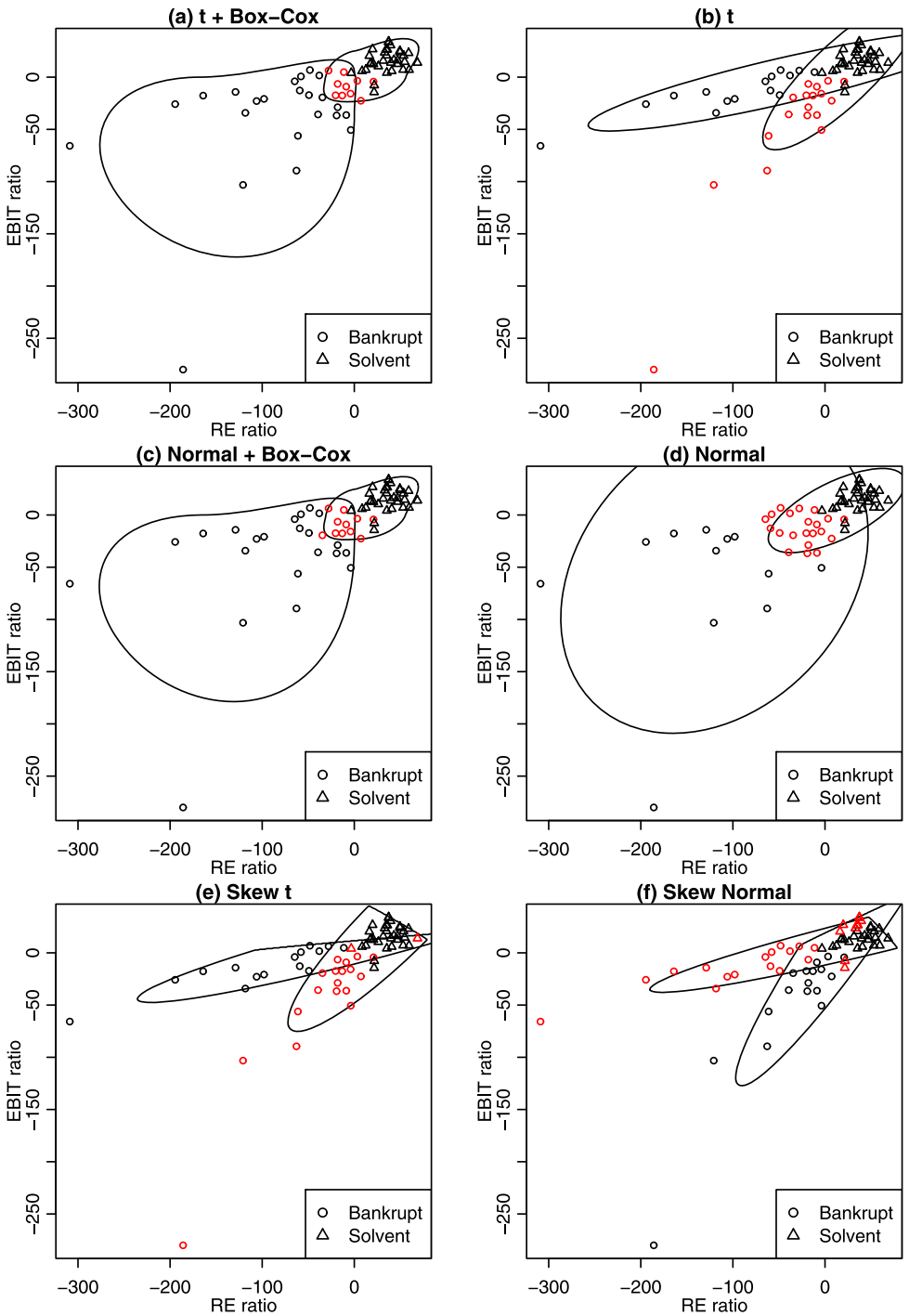

Fig. 2 Scatterplots revealing the assignment of observations for different models applied to the bankruptcy data set. The *black solid lines* represent the 90th percentile region of the components in the mixture models. Misclassified observations are drawn in *red*

(Fig. 2(b, d–f)). The challenge likely arises from the scattered group of bankrupt firms, with its most concentrated region located at the upper right corner and in close proximity to the dense group of solvent firms. The sensitivity of normal mixture models to outliers is clearly demonstrated in this example: the obvious outlier at the bottom of the scatterplot leads to an excessively sparse component representing the bankrupt group. Consequently, most observations in the bankrupt group have been absorbed by the compact component representing the solvent group. The shapes of the components in the t, skew-t and skew-normal mixture models are not all the same, but it appears that for all of them the scattered group of bankrupt firms are split into two components with one absorbing a concentration extending to the left and the other to the bottom. In contrast, both the tBC and NBC mixture models provide a nice representation of both

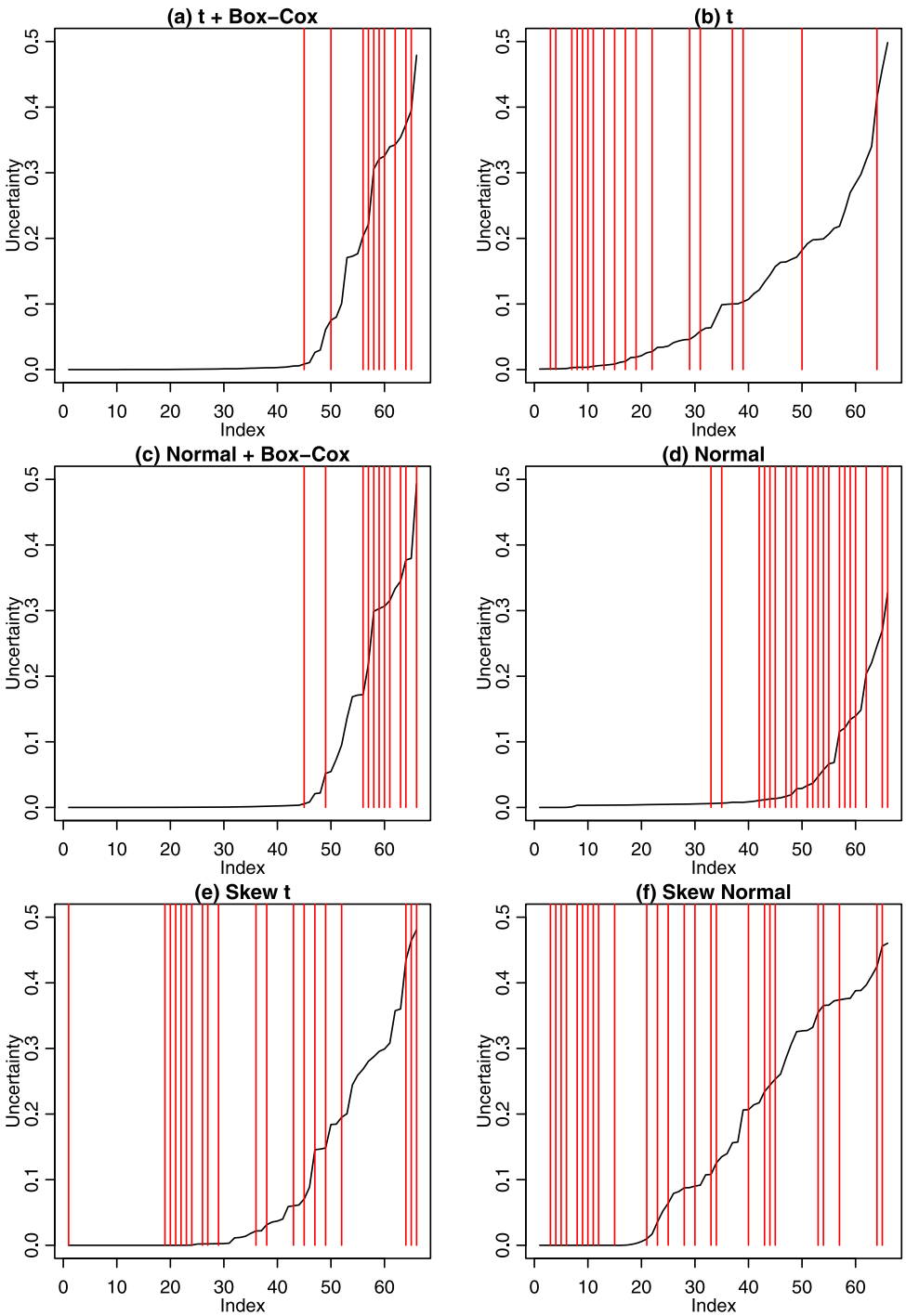

Fig. 3 Plots revealing the location of misclassified observations relative to the ordered uncertainties of all observations for different models applied to the bankruptcy data set. Locations of the misclassified observations are marked with *red vertical lines*

groups of observations (Fig. 2(a, c)). The group of bankrupt firms is resolved quite successfully upon a proper transformation ($\hat{\lambda} \approx 0.5$ for both models) of the observations.

As another means of performance assessment, we look into the location of the misclassified observations in a plot of the ordered uncertainties (Fig. 3). On observing that the misclassified observations have spread over the entire range of the uncertainties, it suggests that the t, skew-t and skew-normal mixture models simply provide an incorrect representation of the two groups (Fig. 3(b, e, f)). The quality of the fit using the tBC and NBC mixture models respectively is confirmed by the corresponding uncertainty plots (Fig. 3(a, c)). The observations associated with high uncertainties are also the ones most likely to be misclassified.

The results on the crabs data set once again show that the tBC mixture model delivers the best performance in

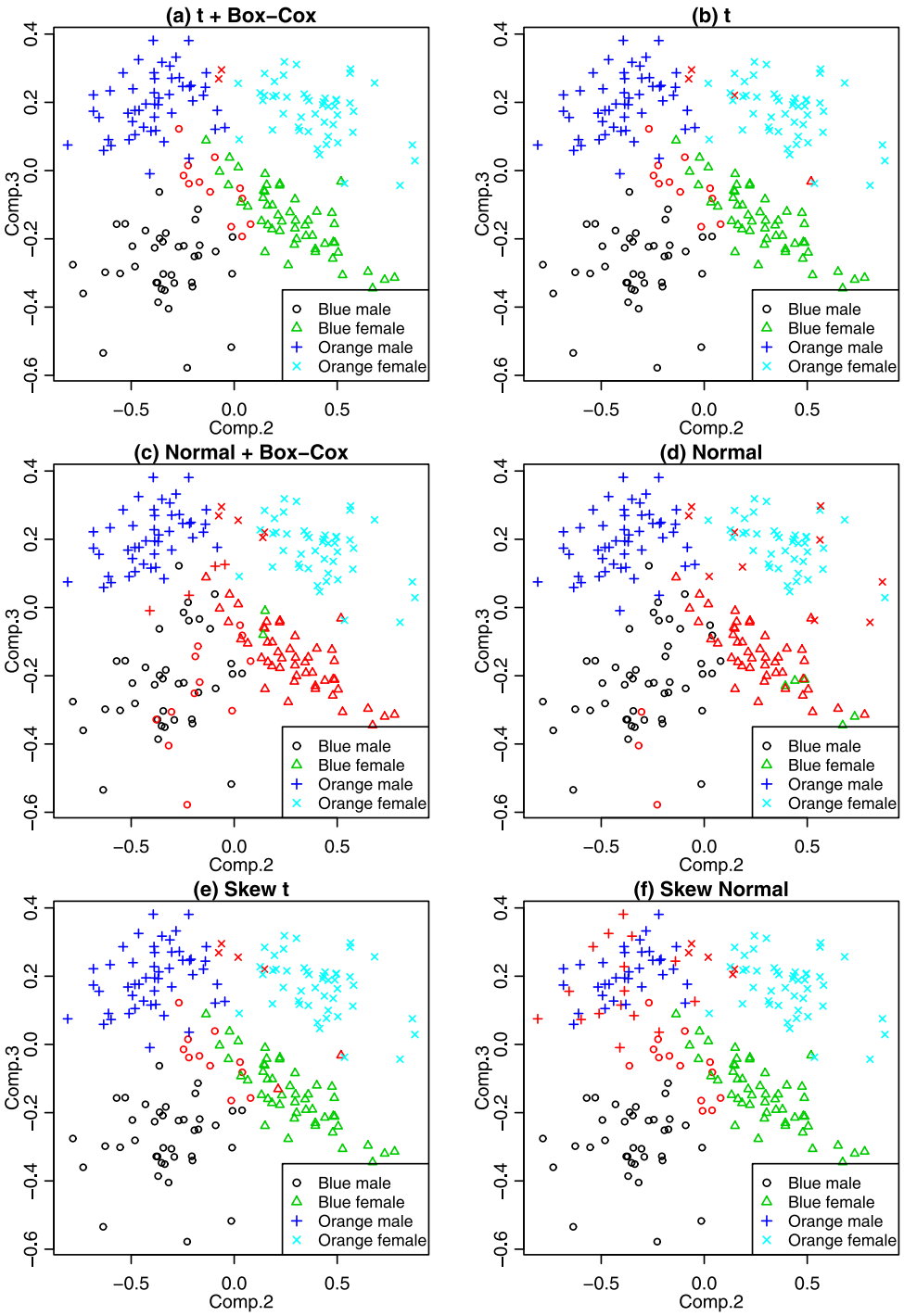

Fig. 4 Plots revealing the assignment of observations for different models applied to the crabs data set, displayed via the second and third principal components. Misclassified observations are drawn in *red*, overriding the original colors used to reveal their true group memberships

terms of misclassification rate (7%). It is followed closely by the t (7.5%) and skew-t (8.5%) mixture models. The crabs data set has also been recently analyzed by Raftery and Dean (2006), Bouveyron et al. (2007) and Scrucca (2010), who employed dimension reduction techniques or subspace clustering to improve Gaussian mixture modeling on high-dimensional data by seeking a more informative subspace. Their reported misclassification rates range between 5–7.5%. Achieving competitive performance against these approaches, the tBC model certainly appears favorable facilitated by allowing for more flexible shapes of the components on the original data space.

Supplementary Fig. 1 (Online Resource 1) shows a scatterplot of the crabs data set projected onto the first two dimensions, namely, the frontal lobe size and the width of the rear region of the carapace. However, unlike the case for the

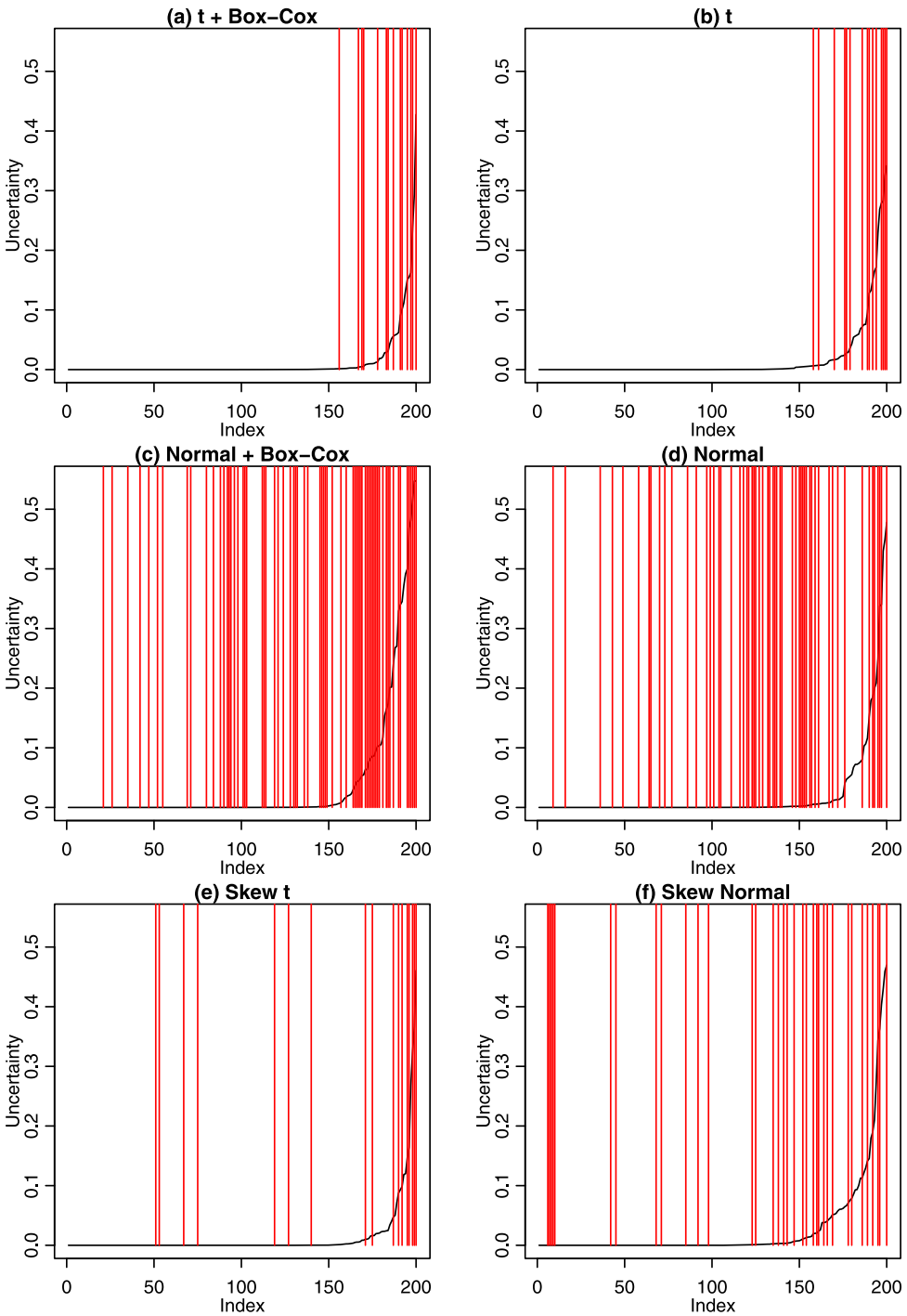

Fig. 5 Plots revealing the location of misclassified observations relative to the ordered uncertainties of all observations for different models applied to the crabs data set. Locations of the misclassified observations are marked with *red vertical lines*

bankruptcy data set with only two dimensions, a visually clear discrimination of the four groups in the crabs data set cannot be achieved by projecting the observations onto any two out of the five dimensions. Therefore, we opt for displaying the crabs data set on its second versus third principal components which provides a good visually discriminating effect. Figure 4(a, b) suggest that those few misclassified observations in the tBC and t mixture models are all likely in the overlapping region of neighboring groups, justifying that these models provide a good representation of all the four groups in the data set. This is further confirmed by a check on the uncertainty plots, in which the misclassified observations are also among the ones with the highest uncertainties (Fig. 5(a, b)). Meanwhile, from Fig. 4(c, d, f) we find that, for the poorly performed NBC, normal and skew-normal mixture models, misclassified cases are concentrated

on one or two of the groups. Supplementary Fig. 1(c, d, f) (Online Resource 1) reveal that these models incorrectly split the assignment of the observations from those groups into other components. As expected, these three poorly performing normal-based models have misclassified observations spreading over the entire range in the uncertainty plots (Fig. 5(c, d, f)).

In the following, we examine whether a pleasant classification performance persists when we apply the tBC mixture modeling methodology to the wine data set with a higher dimension ($p = 13$). From Table 2, we see that the tBC mixture model delivers a misclassification rate (10.7%) drastically lower than the other models (28.1%–38.2%). The high misclassification rates observed likely arise from the fact that the number of free parameters to estimate increases in the order of p^2, while the sample size is not very large ($n = 178$) relative to the data dimension. We therefore attempt to reduce the data dimension via principal component analysis and proceed with the major principal components that account for the majority of the total variance in the data. A turning point at four is observed from the scree plot (Supplementary Fig. 2 (Online Resource 2)), suggesting that the first four principal components are the major contributors to the total variance; the cumulative proportion of the total variance explained by these four principal components is about 80%. We thereby extract the first four principal components, repeat the model-fitting procedure as detailed in Sect. 3.2, and determine the misclassification rates. As reported in Table 2, a significant improvement in classification has been observed for all the six approaches. With the exception of the skew-normal mixture model, each of them succeeds in arriving at a much lower misclassification rate of 5.1%–7.3%. The tBC mixture model still manages to deliver marginally the best classification performance among the six approaches. Its count of nine misclassified observations lags behind the result of Andrews and McNicholas (2010) who correctly classified all but one of the observations, but levels the finding reported in Scrucca (2010); both cited references integrated dimension reduction techniques into Gaussian or t mixture modeling.

Figure 6 displays the wine data set projected onto its first and second principal components. The five models represented in Fig. 6(a–e) accordingly by and large recover the group structure of the data, while the skew-normal mixture model poorly identifies the separation between Barolo and Grignolino wines. As in the analysis of the crabs data set, once again, all the misclassified observations are found in the rightmost region of the plot of ordered uncertainties for the tBC and t mixture models (Fig. 7(a–b)). On the contrary, from Fig. 7(f) we see that five of the misclassified observations in the skew-normal mixture model fall in the least-uncertainty region (left end of the plot). This empirically justifies its failure to recover the shape of the groups of the wine data set.

Table 2 Misclassification rates for different models applied to the wine data set and its first four principal components

Model	Wine	Prin. Comp.
tBC	**0.107 (19)**	**0.051 (9)**
t	0.320 (57)	0.062 (11)
NBC	0.382 (68)	0.062 (11)
Normal	0.303 (54)	0.062 (11)
Skew-t	0.281 (50)	0.073 (13)
Skew-normal	0.354 (63)	0.124 (22)

The best results are shown in bold. The numbers of misclassified cases are given within parentheses

3.3.2 Selecting the number of components

From the BIC curves shown in Fig. 8, we observe that one single peak is observed for each modeling choice over the range of the number of components attempted. The number of components at which a peak is observed is deemed optimal by the BIC for the respective model. The BIC has selected the correct number of components (two) for all the mixture models except normal when applied to the bankruptcy data set (Table 3). As to the crabs data set in which the separation of the groups is less clear-cut, it poses a challenge of selecting the right number of components to most models. Only the tBC and t mixture models have recovered the correct number of components (four) guided by the BIC. This result further confirms our observation stated in the last subsection that the four-component mixture model using the tBC or t mixture model provides the best representation of the data out of all candidates.

4 Simulation studies

We have conducted a series of simulations to further evaluate the relative performance of our proposed framework to the other approaches presented in Sect. 3.2. The different approaches are evaluated for their sensitivity against model misspecification, using the following two criteria: the accuracy in the assignment of observations, and the accuracy in selecting the number of components.

4.1 Implementation

To facilitate the comparison, we generate data from the following mixture models: tBC, skew-t, t and normal. To assess the accuracy in the assignment of observations, two settings of parameter values have been adopted: one taken from the estimates obtained when applying each of the aforementioned models to the bankruptcy data set, and the other one from the crabs data set, with the number of components set

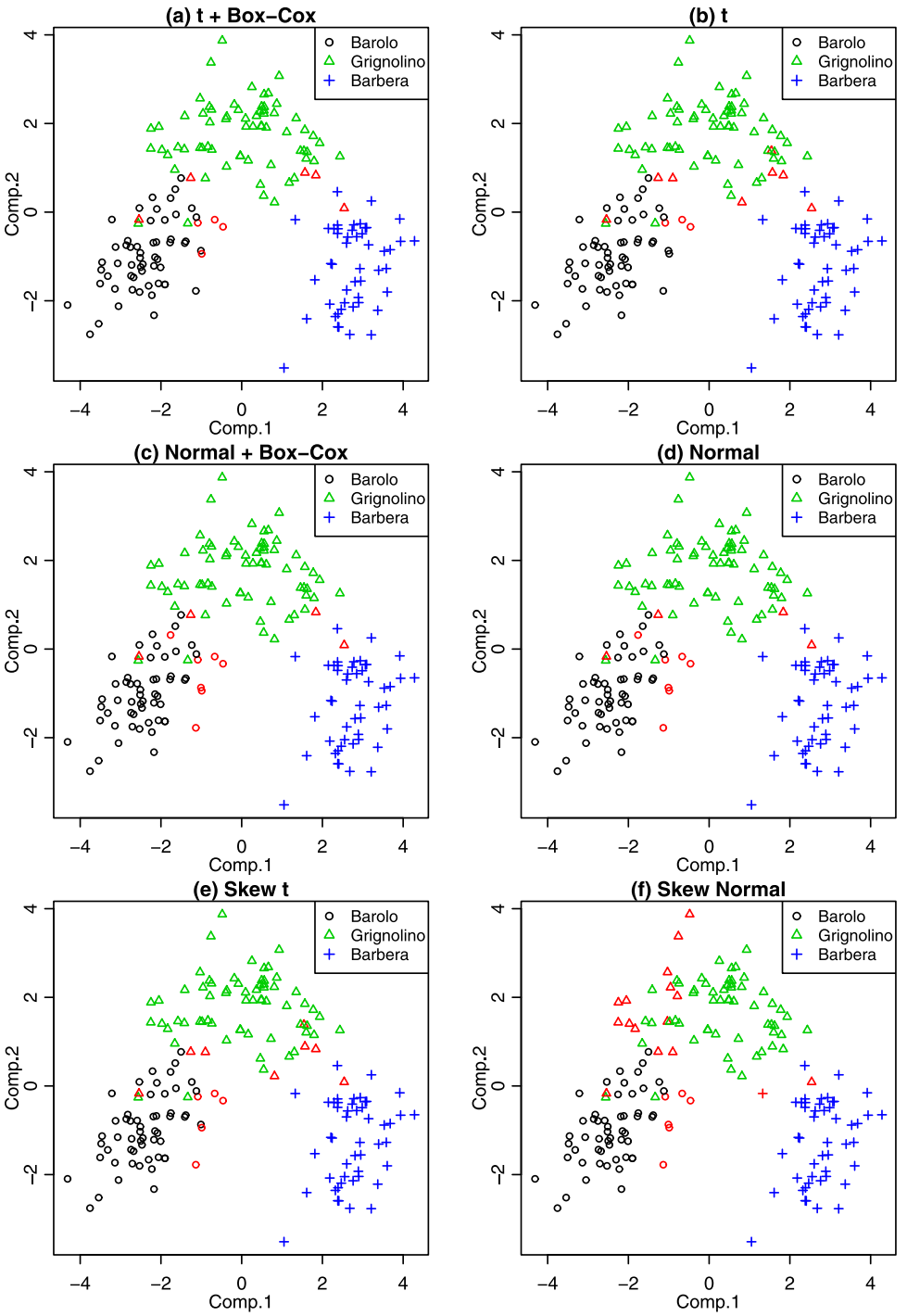

Fig. 6 Plots revealing the assignment of observations for different models applied to the wine data set, displayed via the first and second principal components. Misclassified observations are drawn in *red*, overriding the original colors used to reveal their true group memberships

as the respective known values. As a result, each data set generated from the bankruptcy setting consists of two components and two dimensions, while that from the crabs setting has four components and five dimensions. For data sets generated under the bankruptcy setting we fix the number of observations at 200, while it is set as 500 for the crabs setting. 100 data sets are generated from each of the aforementioned models under each setting. To study the accuracy in selecting the number of components, we focus at the crabs setting. Pertaining to this criterion, the crabs setting offers a better platform to discriminate the relative performance of the different approaches for its larger number of groups and higher dimensions. 1000 observations are generated from the crabs setting instead to avoid numerical problems that may arise from small components formed when the number of components is significantly larger than the true number.

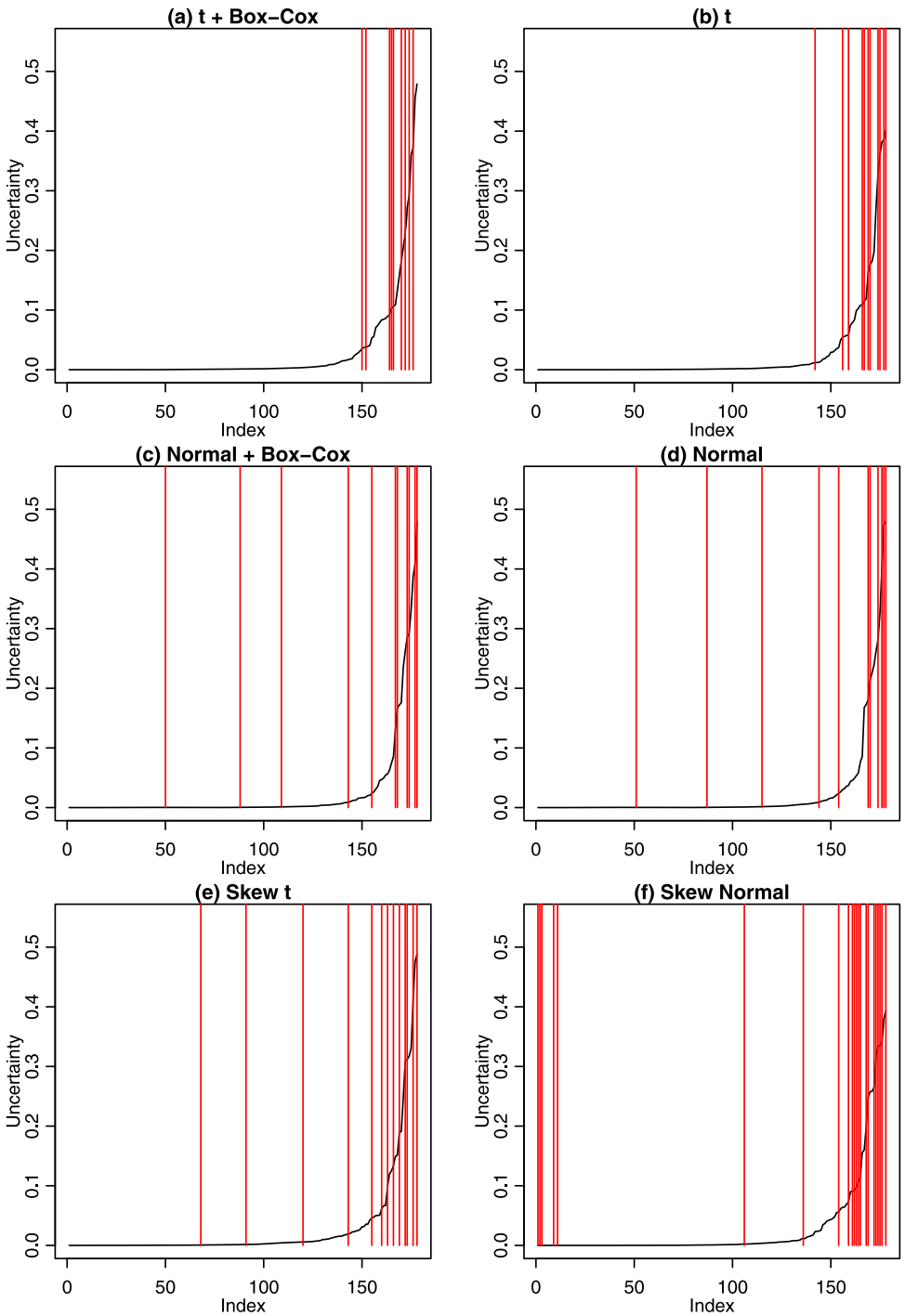

Fig. 7 Plots revealing the location of misclassified observations relative to the ordered uncertainties of all observations for different models applied to the wine data set. Locations of the misclassified observations are marked with *red vertical lines*

We apply the six approaches presented in Sect. 3.2 in turn to each generated data set. The same implementation details elucidated in Sect. 3.2 apply here, with the exceptions stated below. In order to complete the simulation studies within a reasonable timeframe, whilst guaranteeing a satisfactory level of convergence of the EM algorithm, the premature tolerance level and the convergence tolerance level are changed to 10^{-2} and 10^{-5} respectively.

In the study of the accuracy in the assignment of observations, model fitting is done by presuming that the number of components is known, i.e., two for the bankruptcy setting and four for the crabs setting. Similar to the way we determined the misclassification rates in our real data analysis, we consider all permutations of the labels of the components and take the lowest one out of all misclassification rates computed. The performance of the different models is

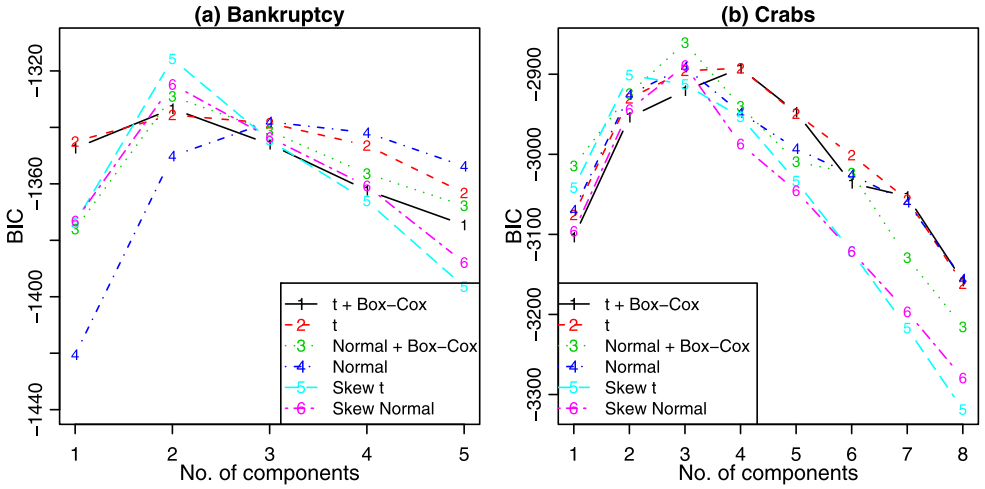

Fig. 8 Plots of BIC against the number of components for the different models applied to the bankruptcy and crabs data sets

Table 3 The number of components selected by the BIC for different models applied to the bankruptcy and crabs data sets

Model	Bankruptcy	Crabs
*t*BC	**2**	**4**
t	**2**	**4**
NBC	**2**	3
Normal	3	3
Skew-*t*	**2**	2
Skew-normal	**2**	3

The best results are shown in bold

compared via the average misclassification rates. To study the accuracy in selecting the number of components, each time when we apply a model to a data set generated under the crabs setting, we set the number of components from one up to eight in turn. The number of components is then selected to be the one which delivers the highest BIC.

On a sidetrack, as suggested by a reviewer of this article, we make a parallel run of the entire simulation study of the accuracy in the assignment of observations by replacing random partitioning with 10 partitions returned by K-means clustering (MacQueen 1967) in initialization. All implementation settings are exactly the same as what have been stated above, otherwise. The purpose of conducting such a parallel run is to see if K-means, a popular choice for initialization in clustering analysis, improves classification by providing apparently more refined partitions than random partitioning.

4.2 Results

4.2.1 Classification

As shown clearly in Table 4 that summarizes the result using random partitioning in initialization, our proposed *t*BC mixture model is the only model that remains the best or close to the best in all the comparisons made. It delivers the lowest misclassification rates under both settings (7.5% and 1.1% respectively) when data are generated from the *t*BC mixture model. The flexibility of the *t*BC mixture model is exhibited when we look into its performance in the scenario of model misspecification. It remains close to the respective true model in those cases, and even delivers the lowest misclassification rate when the true model is *t* under the bankruptcy setting (10.9%), or normal under the crabs setting (2.7%). Contrariwise, when data are generated from the *t*BC mixture model, with a lack of mechanisms to handle asymmetric components, both the *t* and normal mixture models do not perform well. It is worth noting that even the skew-*t* mixture model, which is intended for data departing from symmetry, also performs poorly; the associated misclassification rate is as high as 14.2% under the bankruptcy setting, while that for *t*BC is only 7.5%. When data are generated from the skew-*t* mixture model, taking advantage of the correct specification the skew-*t* mixture model performs well. The *t*BC mixture model also shows a competent performance, however. Meanwhile, the skew-*t* mixture model performs satisfactorily when the true mixture model is *t* or normal. The normal mixture model cannot match the others at all when data are generated from any other model included in this study, showing its vulnerability to outliers and asymmetric components. In addition, it is interesting to notice that the normal mixture model gives a rather high misclassification rate (3.8%) relative to the levels attained by *t*BC, *t* and skew-*t* (2.7%–2.9%) when it itself is the true model for data generation under the crabs setting.

The result of a parallel run of this part of study using K-means initialization is summarized in Supplementary Table 1 (Online Resource 3). Compared with Table 4, whilst a slight increase overall in the misclassification rates pertaining to the bankruptcy setting when using K-means is observed, the increase as to the crabs setting is significant and consistent across all the six approaches in comparison.

Table 4 Average misclassification rates for different models applied to data sets generated under the bankruptcy or crabs setting

		Model used to fit data					
		*t*BC	*t*	NBC	Normal	Skew-*t*	Skew-normal
Model used to generate data under the bankruptcy setting	*t*BC	**0.075**	0.124	**0.075**	0.126	0.142	0.110
	Skew-*t*	0.100	0.094	0.225	0.189	**0.087**	0.191
	t	**0.109**	**0.109**	0.126	0.134	0.114	0.144
	Normal	0.032	0.032	0.033	**0.030**	0.032	0.031
Model used to generate data under the crabs setting	*t*BC	**0.011**	0.014	0.057	0.074	0.015	0.016
	Skew-*t*	0.024	0.023	0.046	0.060	**0.020**	0.021
	t	0.024	**0.021**	0.048	0.070	0.023	0.023
	Normal	**0.027**	0.029	0.042	0.038	0.028	0.028

The best results are shown in bold

Table 5 90% coverage intervals of the number of components, and numbers of times the correct number of components is selected by the BIC for different models applied to data sets generated under the crabs setting

		Model used to fit data											
		*t*BC		*t*		NBC		Normal		Skew-*t*		Skew-normal	
		Int.	#corr.	Int.	#corr.	Int.	#corr.	Int.	#corr.	Int.	#corr.	Int.	#corr.
Model used to generate data	*t*BC	**(4, 4)**	96	**(4, 4)**	**100**	**(4, 4)**	96	**(4, 4)**	99	(3, 5)	72	(3, 5)	78
	Skew-*t*	**(4, 4)**	99	**(4, 4)**	96	(4, 5)	88	(4, 5)	90	**(4, 4)**	**100**	**(4, 4)**	**100**
	t	**(4, 4)**	97	**(4, 4)**	97	(4, 5)	94	(4, 5)	85	**(4, 4)**	**100**	**(4, 4)**	**100**
	Normal	**(4, 4)**	**100**	(4, 5)	95	**(4, 4)**	96	(4, 5)	95	**(4, 4)**	**100**	**(4, 4)**	**100**

The best results are shown in bold. Int. = Interval; #corr = number of times the correct number of components is selected

We, however, do not deem this finding counter-intuitive. As can be seen, there is considerable skewness underlying the groups in data sets generated under the bankruptcy or crabs setting. K-means is equivalent to the classification EM algorithm for a normal mixture model assuming a common covariance matrix spherical in shape (Celeux and Govaert 1992; Celeux and Govaert 1995). These assumptions are obviously violated by the asymmetric groups found in the generated data sets. Such a violation of the assumptions incurs K-means to deliver misleading initial partitions, which ultimately converge to sub-optimal local maxima. The result presented here is a mere example revealing how an apparently more informative initialization scheme based upon an incorrect assumption can compromise the result. As such, initialization other than random partitioning that brings in additional assumptions should be used with caveat.

4.2.2 Selecting the number of components

Table 5 summarizes the result of this part of study, giving 90% coverage intervals of the number of components, together with numbers of times the correct number of components is selected for each model out of the 100 repetitions. The *t*BC mixture model selects the correct number of components (four) in the majority of repetitions, even in case of model misspecification. It is the only model that remains to contain only the true number of components in all the 90% coverage intervals. On the other hand, both the skew-*t* and skew-normal mixture models fail to distinguish the four groups properly in about 25% of the data sets generated from the *t*BC mixture model. Besides, both the NBC and normal mixture models, when applied to data sets generated from the *t* or skew-*t* mixture model, tend to require an additional component to accommodate the data in an excess of outliers.

5 Discussion

In this article, we have introduced a new class of distributions, the *t* distributions with the Box-Cox transformation, for mixture modeling. The proposed methodology is in line with Lange et al. (1989) notion that transformation selection and outlier identification are two issues of mutual influence and therefore should be handled simultaneously. In our real data applications and simulation studies, we have shown the flexibility of this methodology in accommodating asymmetric components in the presence of outliers, and in

coping with model misspecification. The vulnerability of the normal-based models to outliers is exposed in the analysis of the crabs data set, in which the presence of outliers prevents a clear distinction of the four groups. A lack of mechanisms to downsize the influence of remote observations undermines the ability of these approaches to properly locate the cores of the four groups in the data set. On the other hand, the analysis of the bankruptcy data set provides a very good example of demonstrating the importance of incorporating data transformation in clustering. In the absence of a means to accommodate components departing from symmetry, the t mixture model fails to provide a reasonable representation of the data, while the number of groups is known in advance. Our simulation studies have confirmed these findings.

As mentioned in the Introduction, although mixture modeling using our proposed tBC distributions and that using the skew-t distributions follow two lines of development with more or less the same aim, our approach has the appeal of being computationally much simpler to implement. As noted in Lin (2010), difficulties have been encountered in evaluating the conditional expectation of the complete-data log-likelihood in the E-step of the EM algorithm for the skew-t mixture model. The objective function cannot be derived in closed form due to the presence of analytically intractable quantities. Numerical techniques for optimization as well as integration need to be employed extensively to update a vast amount of quantities in both the E and M-steps of the algorithm, undermining the computational stability therein. Besides, the parameterization that accounts for skewness in our proposed model originates from the family of power transformations, which is intuitively interpretable. It is less trivial to interpret the skewness vector parameterized in the skew-t distribution, however. In addition, as presented in Sect. 2.3.3, the way to identify outliers using our approach is straightforward and supported with a theoretical justification: exploiting the relationship between $\tilde{u}_{ig}$ and the quantile of an F distribution through (22), it is almost costless to proceed with outlier identification once the EM algorithm is completed. On the contrary, when the skew-t mixture model is used, we cannot determine such a threshold by recasting it as a known quantity obtained from a standard distribution. Consequently, it demands extra computational effort to identify outliers, especially when the dimension of the data is high. Finally, and perhaps most importantly, as demonstrated from our real data applications and simulation studies, the simplicity of the computational implementation of our proposed methodology is not achieved at the expense of the quality of performance. The results have shown that our proposed approach performs as well as that based on the skew-t mixture model, or even slightly better.

In this article, we present a flexible form of multivariate mixture modeling that simultaneously incorporates outlier identification and data transformation. Nevertheless, we are aware that the modeling performance may not be very satisfactory when we apply the proposed methodology to data sets with a high dimension, say, $p \geq 10$. As noted by Fraley and Raftery (2002) and Bouveyron et al. (2007), mixture models in the Gaussian (or related) family suffer from an excessive number of parameters to estimate when the data dimension is high: the number of free parameters increases in the order of p^2. Incidentally, a high-dimensional data set likely contains variables redundant in providing clustering information. To resolve the aforementioned issues, we need to reduce the number of free parameters. Constrained parameterization of the covariance matrices via eigenvalue decomposition first proposed by Banfield and Raftery (1993) would be a potential direction to investigate at a subsequent stage. Another, but not mutually exclusive, approach is to incorporate dimension reduction techniques. In this article, the analysis of the wine data set has revealed the potential benefit of dimension reduction with a simple application of principal component analysis. More vigorous approaches (see, for example, McNicholas and Murphy 2008; Scrucca 2010; Andrews and McNicholas 2010) that integrate dimension reduction into mixture modeling would be desirable such that our proposed methodology can handle feature selection along with data transformation and outlier identification concurrently.

An open-source software package that facilitates flow cytometry analysis with the methodology proposed in this article has been developed and is available at Bioconductor (Gentleman et al. 2004). It is released as an R package called flowClust (Lo et al. 2009) and addresses the vast demand for software development from the flow cytometry community. flowClust is dedicated to the automated identification of cell populations, and is well integrated into other flow cytometry packages. Meanwhile, we recognize the potential of applying the proposed methodology in other fields, and the importance of developing a general-purpose tool like MCLUST (Fraley and Raftery 2002, 2006), the popular software that performs clustering analysis based on normal mixture models. We are going to work on such a general-purpose, standalone software that will serve as a contribution to the general public.

Acknowledgements The authors thank the Associate Editor and the two reviewers for comments that improved an earlier draft of the article. This research was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada, NIH grant EB008400 and a MITACS internship.

References

Altman, E.I.: Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. J. Finance **23**(4), 589–609 (1968)

Andrews, J.L., McNicholas, P.D.: Extending mixtures of multivariate t-factor analyzers. Stat. Comput. (2010, in press). doi: 10.1007/s11222-010-9175-2

Atkinson, A.C.: Transformations unmasked. Technometrics **30**, 311–318 (1988)

Azzalini, A.: A class of distributions which includes the normal ones. Scand. J. Statist. **12**, 171–178 (1985)

Azzalini, A., Capitanio, A.: Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. J. R. Stat. Soc. Ser. B **65**(2), 367–389 (2003)

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

Banfield, J.D., Raftery, A.E.: Model-based Gaussian and non-Gaussian clustering. Biometrics **49**, 803–821 (1993)

Bensmail, H., Celeux, G., Raftery, A.E., Robert, C.P.: Inference in model-based cluster analysis. Stat. Comput. **7**, 1–10 (1997)

Bickel, P.J., Doksum, K.A.: An analysis of transformations revisited. J. Am. Stat. Assoc. **76**(374), 296–311 (1981)

Bouveyron, C., Girard, S., Schmid, C.: High-dimensional data clustering. Comput. Stat. Data Anal. **52**, 502–519 (2007)

Box, G.E.P., Cox, D.R.: An analysis of transformations. J. R. Stat. Soc. Ser. B **26**, 211–252 (1964)

Brent, R.: Algorithms for Minimization without Derivatives. Prentice-Hall, Englewood Cliffs (1973)

Campbell, N.A., Mahon, R.J.: A multivariate study of variation in two species of rock crab of the genus Leptograpsus. Aust. J. Zool. **22**(3), 417–425 (1974)

Carroll, R.J.: Prediction and power transformations when the choice of power is restricted to a finite set. J. Am. Stat. Assoc. **77**(380), 908–915 (1982)

Celeux, G., Govaert, G.: A classification EM algorithm for clustering and two stochastic versions. Comput. Stat. Data Anal. **14**(3), 315–332 (1992)

Celeux, G., Govaert, G.: Gaussian parsimonious clustering models. Pattern Recogn. **28**(5), 781–793 (1995)

Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B **39**, 1–38 (1977)

Forbes, F., Peyrard, N., Fraley, C., Georgian-Smith, D., Goldhaber, D.M., Raftery, A.E.: Model-based region-of-interest selection in dynamic breast MRI. J. Comput. Assist. Tomogr. **30**, 675–687 (2006)

Forina, M., Armanino, C., Castino, M., Ubigli, M.: Multivariate data analysis as a discriminating method of the origin of wines. Vitis **25**(3), 189–201 (1986)

Fraley, C., Raftery, A.E.: How many clusters? Which clustering method? Answers via model-based cluster analysis. Comput. J. **41**(8), 578–588 (1998)

Fraley, C., Raftery, A.E.: Model-based clustering, discriminant analysis, and density estimation. J. Am. Stat. Assoc. **97**(458), 611–631 (2002)

Fraley, C., Raftery, A., Wehrens, R.: Incremental model-based clustering for large datasets with small clusters. J. Comput. Graph. Stat. **14**(3), 529–546 (2005)

Fraley, C., Raftery, A.E.: MCLUST version 3 for R: Normal mixture modeling and model-based clustering. Tech. Rep. 504, University of Washington, Department of Statistics (2006, revised 2009)

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y.H., Zhang, J.: Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. **5**(10), R80 (2004)

Gutierrez, R.G., Carroll, R.J., Wang, N., Lee, G.H., Taylor, B.H.: Analysis of tomato root initiation using a normal mixture distribution. Biometrics **51**, 1461–1468 (1995)

Hurley, C.: Clustering visualizations of multivariate data. J. Comput. Graph. Stat. **13**(4), 788–806 (2004)

Johnson, R.A., Wichern, D.W.: Applied Multivariate Statistical Analysis. Prentice Hall, Upper Saddle River (2002)

Kass, R.E., Raftery, A.E.: Bayes factors. J. Am. Stat. Assoc. **90**(430), 773–795 (1995)

Keribin, C.: Consistent estimation of the order of mixture models. Sankhyā Ser. A **62**(1), 49–66 (2000)

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

Kriessler, J.R., Beers, T.C.: Substructure in galaxy clusters: a two-dimensional approach. Astron. J. **113**, 80–100 (1997)

Lange, K.L., Little, R.J.A., Taylor, J.M.G.: Robust statistical modeling using the t-distribution. J. Am. Stat. Assoc. **84**, 881–896 (1989)

Leroux, M.: Consistent estimation of a mixing distribution. Ann. Stat. **20**, 1350–1360 (1992)

Li, Q., Fraley, C., Bumgarner, R.E., Yeung, K.Y., Raftery, A.E.: Donuts, scratches and blanks: Robust model-based segmentation of microarray images. Bioinformatics **21**(12), 2875–2882 (2005)

Lin, T.I.: Maximum likelihood estimation for multivariate skew normal mixture models. J. Multivar. Anal. **100**(2), 257–265 (2009a)

Lin, T.I.: Robust mixture modeling using multivariate skew t distributions. Stat. Comput. **20**(3), 343–356 (2010)

Lin, T.I., Lee, J.C., Hsieh, W.J.: Robust mixture modeling using the skew t distribution. Stat. Comput. **17**, 81–92 (2007a)

Lin, T.I., Lee, J.C., Yen, S.Y.: Finite mixture modelling using the skew normal distribution. Stat. Sin. **17**, 909–927 (2007b)

Liu, C.: ML estimation of the multivariate t distribution and the EM algorithm. J. Multivar. Anal. **63**, 296–312 (1997)

Liu, C., Rubin, D.: The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. Biometrika **81**(4), 633–648 (1994)

Liu, C., Rubin, D.: ML estimation of the t distribution using EM and its extensions, ECM and ECME. Stat. Sin. **5**, 19–39 (1995)

Lo, K., Brinkman, R.R., Gottardo, R.: Automated gating of flow cytometry data via robust model-based clustering. Cytometry A **73A**(4), 321–332 (2008)

Lo, K., Hahne, F., Brinkman, R.R., Gottardo, R.: flowClust: a Bioconductor package for automated gating of flow cytometry data. BMC Bioinformatics **10**, 145 (2009)

MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: LeCam, L., Neyman, J. (eds.) Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. vol. 1, pp. 281–297. University of California Press, Berkeley (1967)

McLachlan, G.J.: The classification and mixture maximum likelihood approaches to cluster analysis. In: Krishnaiah, P.R., Kanal, L. (eds.) Handbook of Statistics. vol. 2, pp. 199–208. North-Holland, Amsterdam (1982)

McLachlan, G.J., Basford, K.E.: Mixture Models: Inference and Applications to Clustering. Dekker, New York (1988)

McLachlan, G., Peel, D.: Finite Mixture Models. Wiley-Interscience, New York (2000)

McLachlan, G.J., Bean, R.W., Peel, D.: A mixture model-based approach to the clustering of microarray expression data. Bioinformatics **18**(3), 413–422 (2002)

McLachlan, G., Peel, D., Bean, R.W.: Modelling high-dimensional data by mixtures of factor analyzers. Comput. Stat. Data Anal. **41**, 379–388 (2003)

McNicholas, P.D., Murphy, T.B.: Parsimonious Gaussian mixture models. Stat. Comput. **18**, 285–296 (2008)

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

Mukherjee, S., Feigelson, E.D., Babu, G.J., Murtagh, F., Fraley, C., Raftery, A.E.: Three types of gamma ray bursts. Astrophys. J. **508**, 314–327 (1998)

Pan, W., Lin, J., Le, C.T.: Model-based cluster analysis of microarray gene-expression data. Genome Biol. **3**(2), R9 (2002)

Peel, D., McLachlan, G.J.: Robust mixture modelling using the t distribution. Stat. Comput. **10**(4), 339–348 (2000)

Pyne, S., Hu, X., Wang, K., Rossin, E., Lin, T.I., Maier, L.M., Baecher-Allan, C., McLachlan, G.J., Tamayo, P., Hafler, D.A., De Jager, P.L., Mesirov, J.P.: Automated high-dimensional flow cytometric data analysis. Proc. Natl. Acad. Sci. USA **106**(21), 8519–8524 (2009)

Raftery, A.E., Dean, N.: Variable selection for model-based clustering. J. Am. Stat. Assoc. **101**(473), 168–178 (2006)

Sahu, S.K., Dey, D.K., Branco, M.D.: A new class of multivariate skew distributions with applications to Bayesian regression. Can. J. Stat. **31**(2), 129–150 (2003)

Schork, N.J., Schork, M.A.: Skewness and mixtures of normal distributions. Commun. Stat. Theory Methods **17**, 3951–3969 (1988)

Schroeter, P., Vesin, J.M., Langenberger, T., Meuli, R.: Robust parameter estimation of intensity distributions for brain magnetic resonance images. IEEE Trans. Med. Imag. **17**(2), 172–186 (1998)

Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)

Scrucca, L.: Dimension reduction for model-based clustering. Stat. Comput. **20**(4), 471–484 (2010)

Stephens, M.: Bayesian analysis of mixture models with an unknown number of components—an alternative to reversible jump methods. Ann. Stat. **28**, 40–74 (2000)

Titterington, D.M., Smith, A.F.M., Makov, U.E.: Statistical Analysis of Finite Mixture Distributions. Wiley, Chichester (1985)

Wang, K., Ng, S.K., McLachlan, G.J.: Multivariate skew t mixture models: applications to fluorescence-activated cell sorting data. In: Conference Proceedings of Digital Image Computing: Techniques and Applications, pp. 526–531. IEEE Computer Society, Los Alamitos (2009)

Wehrens, R., Buydens, L.M.C., Fraley, C., Raftery, A.E.: Model-based clustering for image segmentation and large datasets via sampling. J. Classif. **21**, 231–253 (2004)

Yeung, K.Y., Fraley, C., Murua, A., Raftery, A.E., Ruzzo, W.L.: Model-based clustering and data transformations for gene expression data. Bioinformatics **17**(10), 977–987 (2001)