Adv Data Anal Classif (2013) 7:321–338
DOI 10.1007/s11634-013-0137-3

REGULAR ARTICLE

Dimension reduction for model-based clustering via mixtures of multivariate t-distributions

Katherine Morris · Paul D. McNicholas · Luca Scrucca

Received: 28 November 2012 / Revised: 5 May 2013 / Accepted: 27 May 2013 /
Published online: 14 June 2013

Abstract We introduce a dimension reduction method for model-based clustering obtained from a finite mixture of t-distributions. This approach is based on existing work on reducing dimensionality in the case of finite Gaussian mixtures. The method relies on identifying a reduced subspace of the data by considering the extent to which group means and group covariances vary. This subspace contains linear combinations of the original data, which are ordered by importance via the associated eigenvalues. Observations can be projected onto the subspace and the resulting set of variables captures most of the clustering structure available in the data. The approach is illustrated using simulated and real data, where it outperforms its Gaussian analogue.

Keywords Dimension reduction · Mixture models · Model-based clustering

Mathematics Subject Classification 62H30

This work was supported by a Queen Elizabeth II Scholarship in Science and Technology (Morris), as well as an Early Researcher Award from the government of Ontario and a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (McNicholas).

K. Morris · P. D. McNicholas (✉)
Department of Mathematics and Statistics, University of Guelph, Ontario, Canada
e-mail: pmcnicho@uoguelph.ca

K. Morris
e-mail: kmorri09@uoguelph.ca

L. Scrucca
Dipartimento di Economia, Finanza e Statistica, Università degli Studi di Perugia, Perugia, Italy
e-mail: luca@stat.unipg.it

1 Introduction

In this paper, we introduce a dimension reduction method for model-based clustering via t-mixtures, which is analogous to the approach of Scrucca (2010) for Gaussian mixtures. Dimension reduction methods summarize the information available in a set of variables through a reduced subset of features derived from the original variables. At the same time, using multivariate t-distributions to model data can be advantageous because it provides robustness to outliers (cf. Peel and McLachlan 2000). Our aim is to estimate a subspace which captures most of the clustering structure contained in the data. Following the work of Li (1991, 2000), the dimension reduction subspace is found by looking at the variation in both group means and group covariances. This subspace contains linear combinations of the original data, which are ordered by importance via the associated eigenvalues. Observations can be projected onto the subspace and the resulting set of variables captures most of the clustering structure available in the data.

The remainder of the paper is outlined as follows. Background material is presented in Sect. 2. In Sect. 3, we outline our dimension reduction clustering method and highlight an algorithm for selecting a subset of the variables while retaining most of the clustering information contained within the data. We apply the algorithm to simulated and real data sets, including comparison of the performance of our method with its Gaussian analogue and five other dimension reduction techniques (Sect. 4). The paper concludes with discussion and suggestions for future work (Sect. 5). All the computational work in this paper was carried out within R (R Development Core Team 2012).

2 Background

Clustering algorithms based on probability models are a popular choice for exploring structures in modern data sets, which continue to grow in size and complexity. The model-based approach assumes that the data are generated by a finite mixture of probability distributions. A p-dimensional random vector $\boldsymbol{X}$ is said to arise from a parametric finite mixture distribution if $f(\boldsymbol{x}|\boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_g(\boldsymbol{x}|\boldsymbol{\theta}_g)$, where G is the number of components, π_g are mixing proportions, so that $\sum_{g=1}^{G} \pi_g = 1$ and $\pi_g > 0$, and $\boldsymbol{\vartheta} = (\pi_1, \ldots, \pi_G, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G)$ is the parameter vector. The $f_g(\boldsymbol{x}|\boldsymbol{\theta}_g)$ are called component densities and $f(\boldsymbol{x}|\boldsymbol{\vartheta})$ is formally referred to as a G-component parametric finite mixture distribution.

For Gaussian mixtures, Banfield and Raftery (1993) developed a model-based framework for clustering by using the following eigenvalue decomposition of the covariance matrix

$$\boldsymbol{\Sigma}_g = \lambda_g \boldsymbol{D}_g \boldsymbol{A}_g \boldsymbol{D}_g^{\top}, \tag{1}$$

where $\boldsymbol{D}_g$ is the orthogonal matrix of eigenvectors of $\boldsymbol{\Sigma}_g$, $\boldsymbol{A}_g$ is a diagonal matrix with elements proportional to the eigenvalues of $\boldsymbol{\Sigma}_g$, and λ_g is a scalar. Here, $\boldsymbol{D}_g$ determines the orientation of the principal components of $\boldsymbol{\Sigma}_g$, $\boldsymbol{A}_g$ determines the shape of the density contours, and λ_g specifies the volume of the corresponding ellipsoid.

Table 1 Nomenclature for a subset of models in the MCLUST family which appear in this paper: 'E' indicates that a constraint is imposed, 'V' indicates that a constraint is not imposed (Fraley and Raftery 1999)

Model	Volume λ_g	Shape $\boldsymbol{A}_g$	Orientation $\boldsymbol{D}_g$	Covariance $\boldsymbol{\Sigma}_g$	Free covariance parameters
EEE	Equal	Equal	Equal	$\lambda\boldsymbol{DAD}^\top$	$p(p+1)/2$
VEV	Variable	Equal	Variable	$\lambda_g\boldsymbol{D}_g\boldsymbol{A}\boldsymbol{D}_g^\top$	$Gp(p+1)/2-(G-1)(p-1)$
VVV	Variable	Variable	Variable	$\lambda_g\boldsymbol{D}_g\boldsymbol{A}_g\boldsymbol{D}_g^\top$	$Gp(p+1)/2$

Table 2 Nomenclature for a subset of models in the tEIGEN family which appear in this paper: 'C' indicates that a constraint is imposed, 'U' indicates that a constraint is not imposed (Andrews and McNicholas 2012a)

Model	Volume λ_g	Shape $\boldsymbol{D}_g$	Orientation $\boldsymbol{A}_g$	Degrees of freedom ν_g	Free covariance parameters
CCCC	C	C	C	C	$[p(p+1)/2]+1$
CUCC	C	U	C	C	$G[p(p+1)/2]-(G-1)p+1$
UUCU	U	U	C	U	$G[p(p+1)/2]-(G-1)(p-1)+G$
UUUC	U	U	U	C	$G[p(p+1)/2]+1$
UUUU	U	U	U	U	$G[p(p+1)/2]+G$

By imposing constraints on the elements of (1), a family of Gaussian parsimonious clustering models (GPCM) is obtained and discussed by Celeux and Govaert (1995). A description of a subset of these models appears in Table 1.

Recent work on model-based clustering using the multivariate t-distribution has been contributed by Peel and McLachlan (2000), McLachlan et al (2007), Greselin and Ingrassia (2010a,b), Andrews et al. (2011), Baek and McLachlan (2011), Andrews and McNicholas (2011a,b), Andrews and McNicholas (2012a), Steane et al. (2012), McNicholas and Subedi (2012), and McNicholas (2013). In particular, Andrews and McNicholas (2012a) used the decomposition (1) of the multivariate t-distribution scale matrix to build a family of 20 mixture models which are called the tEIGEN family (a selection is shown in Table 2). The tEIGEN family uses the ten MCLUST covariance structures as well as two of the other GPCM structures and also includes constraints on the degrees of freedom (cf. Andrews and McNicholas 2012a).

Parameter estimation for the tEIGEN family is carried out via the expectation-conditional maximization algorithm (ECM; Meng and Rubin 1993), an iterative procedure for finding maximum likelihood estimates when data are incomplete or treated as being incomplete. Extensive details on the ECM algorithm are given by McLachlan and Krishnan (2008). The ECM algorithm replaces the M-step in the expectation-maximization (EM) algorithm (Dempster et al. 1977) with a number of conditional maximization steps that can be more computationally efficient.

Model selection for the tEIGEN family is carried out using the Bayesian information criterion (BIC; Schwarz 1978):

$$\text{BIC} = 2l(\boldsymbol{x}, \hat{\boldsymbol{\vartheta}}) - r\log n\,,$$

where $l(\boldsymbol{x}, \hat{\boldsymbol{\vartheta}})$ is the maximized log-likelihood, $\hat{\boldsymbol{\vartheta}}$ is the maximum likelihood estimate of $\boldsymbol{\vartheta}$, r is the estimated number of free parameters, and n is the number of observations. While alternatives to the BIC exist, the authors feel safe contending that it remains the most popular mixture model selection criterion within the literature. Note that the tEIGEN family is implemented within the `teigen` package (Andrews and McNicholas 2012b) for R; we use `teigen` in our analyses (Sect. 4).

Recently, Scrucca (2010) proposed a new method of dimension reduction for model-based clustering in the Gaussian framework, called GMMDR. Given a G-component Gaussian mixture model (GMM) of the form

$$f(\boldsymbol{x}|\boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_{\mathrm{N}}(\boldsymbol{x}|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g),$$

the procedure finds the smallest subspace that captures the clustering information contained within the data. The core of the method is to identify those directions where the cluster means $\boldsymbol{\mu}_g$ and the cluster covariances $\boldsymbol{\Sigma}_g$ vary as much as possible, provided that each direction is $\boldsymbol{\Sigma}$-orthogonal to the others. The variations in cluster means and cluster covariances are captured by the matrices $\boldsymbol{M}_{\mathrm{I}}$ and $\boldsymbol{M}_{\mathrm{II}}$, respectively, as given below. Identifying these directions is achieved via the generalized eigen-decomposition of the kernel matrix $\boldsymbol{M}$, defined by Scrucca (2010) as

$$\boldsymbol{M}\boldsymbol{v}_i = l_i \boldsymbol{\Sigma}\boldsymbol{v}_i, \tag{2}$$

where $l_1 \geq l_2 \geq \cdots \geq l_d > 0$ and $\boldsymbol{v}_i^{\top}\boldsymbol{\Sigma}\boldsymbol{v}_j = \begin{cases} 1 & \text{if } i = j, \text{ and} \\ 0 & \text{otherwise.} \end{cases}$

Here,

$$\begin{aligned} \boldsymbol{M} &= \boldsymbol{M}_{\mathrm{I}}\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{\mathrm{I}} + \boldsymbol{M}_{\mathrm{II}}, \\ \boldsymbol{M}_{\mathrm{I}} &= \sum_{g=1}^{G} \pi_g(\boldsymbol{\mu}_g - \boldsymbol{\mu})(\boldsymbol{\mu}_g - \boldsymbol{\mu})^{\top}, \\ \boldsymbol{M}_{\mathrm{II}} &= \sum_{g=1}^{G} \pi_g(\boldsymbol{\Sigma}_g - \bar{\boldsymbol{\Sigma}})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}_g - \bar{\boldsymbol{\Sigma}})^{\top}. \end{aligned}$$

Also, $\boldsymbol{\mu} = \sum_{g=1}^{G} \pi_g \boldsymbol{\mu}_g$ is the global mean, $\boldsymbol{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^{\top}$ is the covariance matrix and $\bar{\boldsymbol{\Sigma}} = \sum_{g=1}^{G} \pi_g \boldsymbol{\Sigma}_g$ is the pooled within-cluster covariance matrix.

3 Mehodology

The dimension reduction approach of Scrucca (2010) is extended herein through development of a t-analogue. The density of a multivariate t-mixture model (tMM) is given by

$$f(\boldsymbol{x}|\boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g \, f_t(\boldsymbol{x}|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \nu_g)\,, \tag{3}$$

where π_g are the mixing proportions and

$$f_t(\boldsymbol{x}|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \nu_g) = \frac{\Gamma(\frac{\nu_g+p}{2})|\boldsymbol{\Sigma}_g|^{-\frac{1}{2}}}{(\pi\nu_g)^{\frac{p}{2}}\Gamma(\frac{\nu_g}{2})(1+\frac{\delta(\boldsymbol{x},\boldsymbol{\mu}_g|\boldsymbol{\Sigma}_g)}{\nu_g})^{\frac{\nu_g+p}{2}}}\,, \tag{4}$$

is the density of a multivariate t-distribution with mean $\boldsymbol{\mu}_g$, scale matrix $\boldsymbol{\Sigma}_g$, and ν_g degrees of freedom, and $\delta(\boldsymbol{x}, \boldsymbol{\mu}_g|\boldsymbol{\Sigma}_g) = (\boldsymbol{x} - \boldsymbol{\mu}_g)^\top \boldsymbol{\Sigma}_g^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_g)$ is the squared Mahalanobis distance between $\boldsymbol{x}$ and $\boldsymbol{\mu}_g$.

Note that, although $\boldsymbol{\Sigma}_g$ is a covariance matrix, it is not the covariance matrix of the random variable $\boldsymbol{X}$ with the density in (4). The covariance matrix of $\boldsymbol{X}$ is $\tilde{\boldsymbol{\Sigma}}_g = \nu_g/(\nu_g - 2)\boldsymbol{\Sigma}_g$, for $\nu_g > 2$. Thus, we obtain a modified version of the kernel matrix $\boldsymbol{M}_t$ (cf. 2), where $\boldsymbol{M}_{\text{II}} = \sum_{g=1}^{G} \pi_g(\tilde{\boldsymbol{\Sigma}}_g - \bar{\boldsymbol{\Sigma}})\boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{\Sigma}_g} - \bar{\boldsymbol{\Sigma}})^\top$.

Given a tMM (3), we wish to find a subspace, $\mathscr{S}(\boldsymbol{\beta})$, where the cluster means and cluster covariances vary the most. As outlined in Sect. 2, this is achieved through the eigen-decomposition of the modified kernel matrix $\boldsymbol{M}_t$.

Definition 3.1 The tMMDR directions are the eigenvectors $[\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d] \equiv \boldsymbol{\beta}$ which form the basis of the dimension reduction subspace $\mathscr{S}(\boldsymbol{\beta})$.

Suppose $\mathscr{S}(\boldsymbol{\beta})$ is the subspace spanned by the tMMDR directions obtained from the eigen-decompostion of $\boldsymbol{M}_t$, and that $\boldsymbol{\mu}_g$ and $\tilde{\boldsymbol{\Sigma}_g}$ are the mean and covariance matrix, respectively, for the gth component. Then the projections of the parameters onto $\mathscr{S}(\boldsymbol{\beta})$ are given by $\boldsymbol{\beta}^\top \boldsymbol{\mu}_g$ and $\boldsymbol{\beta}^\top \tilde{\boldsymbol{\Sigma}}_g \boldsymbol{\beta}$, respectively. For an $n \times p$ sample data matrix $\boldsymbol{X}$, the sample version $\hat{\boldsymbol{M}}_t$ of the kernel $\boldsymbol{M}_t$ is obtained using the corresponding estimates from the fit of a t-mixture model via the ECM algorithm. Then the tMMDR directions are calculated from the generalized eigen-decomposition of $\hat{\boldsymbol{M}}_t$ with respect to $\hat{\boldsymbol{\Sigma}}$. The tMMDR directions are ordered based on eigenvalues; this means that directions associated with approximately zero eigenvalues can be discarded in practice because clusters will overlap substantially along these directions. Also, their contribution to the overall location of the sample points in an eigenvector expansion is approximately zero and so they provide little positional information.

Definition 3.2 The tMMDR variables, $\boldsymbol{Z}$, are the projections of the $n \times p$ data matrix $\boldsymbol{X}$ onto the subspace $\mathscr{S}(\boldsymbol{\beta})$ and can be computed as $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{\beta}$.

As in the case of GMMDR, the estimation of the tMMDR variables can be viewed as a form of feature extraction where the components are reduced through a set of linear combinations of the original variables. This set of features may contain estimated tMMDR variables that provide no clustering information but require parameter estimation. Thus, the next step in the process of model-based clustering is to detect and remove these unnecessary tMMDR variables.

Scrucca (2010) used the subset selection method of Raftery and Dean (2006) to prune the subset of GMMDR features. We will also use this approach to select the most appropriate tMMDR features. We compare two subsets of features, s and $s' = \{s \backslash i\} \subset s$, using the BIC difference

$$\begin{aligned}\text{BIC}_{\text{diff}}(Z_{i\in s}) &= \text{BIC}_{\text{clust}}(Z_s) - \text{BIC}_{\text{not clust}}(Z_s)\\ &= \text{BIC}_{\text{clust}}(Z_s) - [\text{BIC}_{\text{clust}}(Z_{s'}) + \text{BIC}_{\text{reg}}(Z_i|Z_{s'})],\end{aligned} \tag{5}$$

where $\text{BIC}_{\text{clust}}(Z_s)$ is the BIC value for the best clustering model fitted using features in s, $\text{BIC}_{\text{clust}}(Z_{s'})$ is the BIC value for the best clustering model fitted using features in s', and $\text{BIC}_{\text{reg}}(Z_i|Z_{s'})$ is the BIC value for the regression of the ith feature on the remaining features in s'.

Now, the space of all possible subsets contains $2^d - 1$ elements and an exhaustive search is not feasible. To bypass this issue, we employ the greedy search algorithm of Scrucca (2010) to find a local optimum in the model space, which is based on the forward-backward search algorithm of Raftery and Dean (2006). The greedy search from Scrucca (2010) is a forward-only procedure; a backward step is not necessary because the tMMDR variables are $\boldsymbol{\Sigma}$-orthogonal. Because a backward step is not needed, computing time is decreased.

1. Select the first feature to be the one which maximizes the BIC difference in (5) between the best clustering model and the model which assumes no clustering, i.e., a single component.
2. Select the next feature amongst those not previously included, to be the one which maximizes the BIC difference in (5).
3. Iterate the previous step until all the BIC differences for the inclusion of a variable become negative.

At each step, the search over the model space is performed with respect to the model parametrization and the number of clusters. A summary of our new method of dimension reduction for model-based clustering via t-mixtures, tMMDR, appears below.

Algorithm: tMMDR estimation and feature selection

1. Fit a tMM to the data using the tEIGEN family.
2. Estimate the tMMDR directions: identify directions where the cluster means and cluster variances vary the most, provided each direction is $\boldsymbol{\Sigma}$-orthogonal to the others. This is done through the eigen-decomposition of the kernel matrix $\boldsymbol{M}_t$.
3. Select the tMMDR variables: compute the set of features by projecting the data onto the estimated subspace and use the greedy search algorithm to discard the ones which provide no clustering information.
4. Fit a tMM on the selected tMMDR variables and return to step 2.
5. Repeat steps 2–4 until none of the remaining features can be dropped.

4 Applications

4.1 Simulated data

We ran the tMMDR algorithm on the data simulation schemes outlined in Scrucca (2010) and compared its performance to that of the GMMDR procedure. Data were generated from the Gaussian distribution and the following models (corresponding to family members) were considered: three overlapping clusters with common covariance, three overlapping clusters with common shape, and three overlapping clusters with unconstrained covariance. For each model we ran three scenarios, namely:

I. No noise variables: generated three variables from a multivariate normal distribution.
II. Noise variables: started with scenario one and added seven noise variables generated from independent standard normal variables.
III. Redundant and noise variables: started with scenario one and added three variables correlated with each clustering variable (with correlation coefficients equal to 0.9, 0.7, 0.5, respectively) as well as four independent standard normal variables.

For the full details on developing the synthetic data see Scrucca (2010). To ascertain the performance of the clustering methods under varying data dimensions, each scenario was run for three data sets consisting of 100, 300, and 1,000 data points, respectively, generated according to the schemes described earlier. Every run comprised 1,000 simulations. We evaluated the clustering results by computing the adjusted Rand index (ARI; Hubert and Arabie 1985) for each scenario: higher values of ARI correspond to better performance, with the value 1 reflecting perfect class agreement.

We chose to generate only 10 variables in the latter two simulation scenarios so that we could provide a direct comparison between our results for tMMDR and the results for GMMDR from Scrucca (2010); results are given in Table 3. When no noise or redundant variables are present, the performance of the two methods is quite similar. However, for the scenarios which include noise and redundant variables, tMMDR exhibits ARI values which are higher than those for GMMDR, particularly for small sample sizes. This occurs consistently for all models as well as for varying data dimensions.

Next, we simulated data with higher dimensions using the R package `clusterGeneration` (Qiu and Joe 2006). With this procedure, we chose to generate five clusters with equal numbers of observations by setting the degree of separation between them to 0.4 (with 1 representing the most separation), and choosing arbitrary positive definite covariance matrices. We generated four data sets, each comprising 300 observations, with dimensions 10, 30, 50, and 100, respectively.

Figure 1 depicts the scatterplot of five random variables from the data which indicate the underlying structure. The clustering results (Table 4) show that the number of features selected and the computing time increases with the dimensionality of the data. These simulations were performed on a computer with 128 GB of RAM utilizing one core of an Intel® Xeon® E7-8837 CPU running at 2.67 GHz. Note that the computing times for our approach will decrease with any future improvements in the efficiency of the `teigen` package.

Table 3 Average ARIs, with standard errors in brackets, for all three scenarios, based on 1,000 simulations

	I: No noise (3 variables)			II: Noise variables (10 variables)			III: Noise and redundant variables (10 variables)		
	$n = 100$	$n = 300$	$n = 1{,}000$	$n = 100$	$n = 300$	$n = 1{,}000$	$n = 100$	$n = 300$	$n = 1{,}000$
Model 1									
GMMDR (EEE)	0.9716	0.9742	0.9753	0.8612	0.9674	0.9742	0.8832	0.9699	0.9742
	(0.0347)	(0.0172)	(0.0088)	(0.1426)	(0.0234)	(0.009)	(0.1613)	(0.0197)	(0.0088)
*t*MMDR (CCCC)	0.9705	0.9722	0.9747	0.9210	0.9709	0.9737	0.9987	0.9994	0.9995
	(0.0314)	(0.0163)	(0.0085)	(0.1763)	(0.0172)	(0.0087)	(0.0116)	(0.0028)	(0.0012)
Model 2									
GMMDR (VEV)	0.9709	0.9802	0.9819	0.9201	0.9747	0.9806	0.9231	0.9727	0.9806
	(0.0351)	(0.0141)	(0.0073)	(0.0799)	(0.0165)	(0.0082)	(0.0811)	(0.0169)	(0.0077)
*t*MMDR (UUCU)	0.9768	0.9796	0.9815	0.9609	0.9781	0.9818	0.9898	0.9995	0.9997
	(0.0283)	(0.0142)	(0.0074)	(0.1571)	(0.0101)	(0.0079)	(0.0197)	(0.0014)	(0.0008)
Model 3									
GMMDR (VVV)	0.9952	0.9981	0.9983	0.8643	0.9674	0.9751	0.8799	0.9712	0.9740
	(0.0146)	(0.0042)	(0.0024)	(0.1384)	(0.021)	(0.0081)	(0.1609)	(0.0185)	(0.0099)
*t*MMDR (UUUU)	0.9972	0.9982	0.9981	0.9865	0.9973	0.9981	0.9612	0.9991	0.9994
	(0.0092)	(0.0042)	(0.0023)	(0.0801)	(0.0052)	(0.0025)	(0.1241)	(0.0168)	(0.0073)

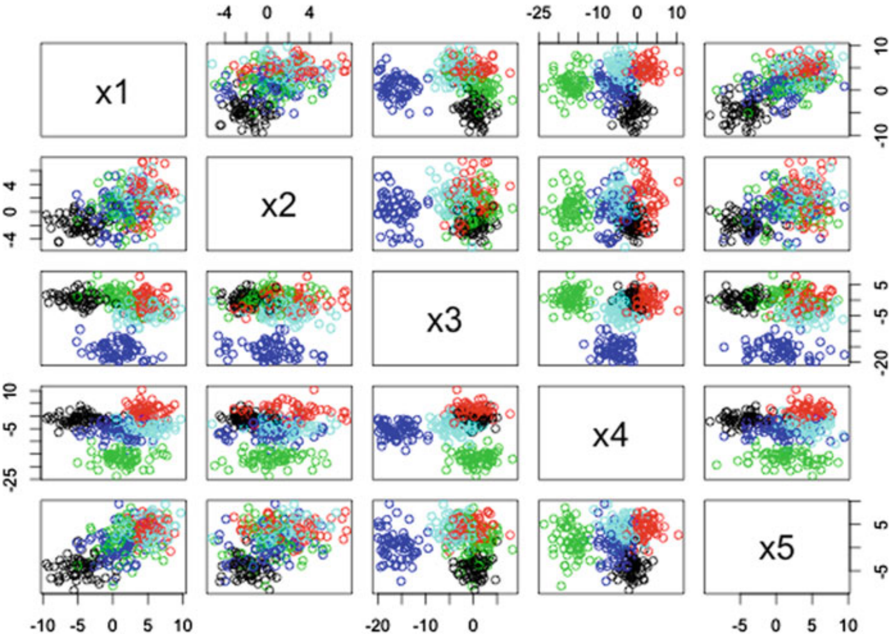

Fig. 1 Scatterplot of five variables from the higher dimensional simulation indicating the five clusters present in the data (colour figure online)

Table 4 Computing time for the *t*MMDR procedure applied to simulated data, based on five runs

Dim.	Obs.	Comp.	Features	ARI	Avg. time (seconds)
10	300	5	4	1	807
30	300	5	5	1	3,174
50	300	5	10	1	4,540
100	300	5	12	0.97	16,676

4.2 Real data

For the analyses in this section, we ran the GMMDR and *t*MMDR algorithms on the scaled version of each data set. We used the MCLUST hierarchical agglomerative procedure for initialization (cf. Fraley and Raftery 1999). In order to gauge the performance of our algorithm, we compared our results with five other dimension reduction procedures, outlined briefly below.

1. Robust PCA algorithm (Hubert et al. 2005) paired with *t*-mixtures via the *t*EIGEN family: principal components analysis resistant to outliers, with robust loadings computed by using projection-pursuit techniques and the minimum covariance determinant method. We used the R package `rrcov` (Todorov and Filzmoser 2009) for the ROBPCA computations as well as the `teigen` package.

2. Mixtures of t-factor analyzers (MMtFA; Andrews and McNicholas 2011a,b) extend the mixtures of multivariate t-factor analyzers model to include constraints on the degrees of freedom, the factor loadings, and the error variance matrices. These models are essentially a t-analogue of the Gaussian family developed by McNicholas and Murphy (2008, 2010). For the MMtFA results, the algorithms were initialized using a hierarchical agglomerative clustering and they were run for a range of 1–5 components and 1–4 factors. The BIC was used to select the number of components and the number of factors. The numbers of components and factors do not need to be specified *a priori* but when we write about the number of features, note that we are referring to the number of latent factors q, where each component requires q factors.
3. FisherEM algorithm (Bouveyron and Brunet 2012): a subspace clustering method based on Gaussian mixtures where the EM-like algorithm estimates both the discriminative subspace and the parameters of the model. This procedure requires the number of clusters to be specified. We used the R package `FisherEM` (Bouveyron and Brunet 2012) in our analyses.
4. Clustvarsel algorithm (Raftery and Dean 2006) paired with Gaussian mixtures via the MCLUST family: a greedy procedure to find the (locally) optimal subset of variables in a dataset. We employed the R package `clustvarsel` (Dean and Raftery 2009) in our analyses.
5. SelvarClust algorithm (Maugis et al 2009): a greedy algorithm for variable selection in model-based clustering via Gaussian mixtures which modifies the method of Raftery and Dean (2006) by allowing data where individuals are described by quantitative block variables. We used the software available at http://www.math.univ-toulouse.fr/~maugis/SelvarClustHomepage.html for our analyses.

4.2.1 Wine data

Forina et al (1986) recorded several chemical and physical properties for three types of Italian wines: Barolo, Grignolino, and Barbera. For our analysis, we used the data set comprising 13 variables and 178 observations which is available from the `gclus` package (Hurley 2004) in R. The resulting classifications (Table 5) and algorithm comparisons (Table 6) show that the MMtFA with 2 factors produces the best results (ARI $=$ 0.96). Also, the tMMDR and `FisherEM` procedures (both with ARI $=$ 0.93) outperform the GMMDR method (ARI $=$ 0.85), while using less features in the process. Figure 2 illustrates a scatterplot of the estimated tMMDR directions corresponding to the three clusters found by the procedure. The separation between the varieties of wine is clear in the plots of direction 1 against directions 2 and 3.

Table 5 A classification table for the best tMMDR model fitted to the wine data

	1	2	3
Barolo	59	2	0
Grignolino	0	67	0
Barbera	0	2	48

Table 6 Summary of results for the best models fitted to the wine data

Method	ARI	Components	Features
tMMDR	0.93	3	4
GMMDR	0.85	3	5
`FisherEM`	0.93	3	2
selvarclust	0.59	5	7
`clustvarsel` with Gaussian mixtures	0.78	3	5
ROBPCA with t-mixtures	0.86	3	5
MMtFA	0.96	3	2

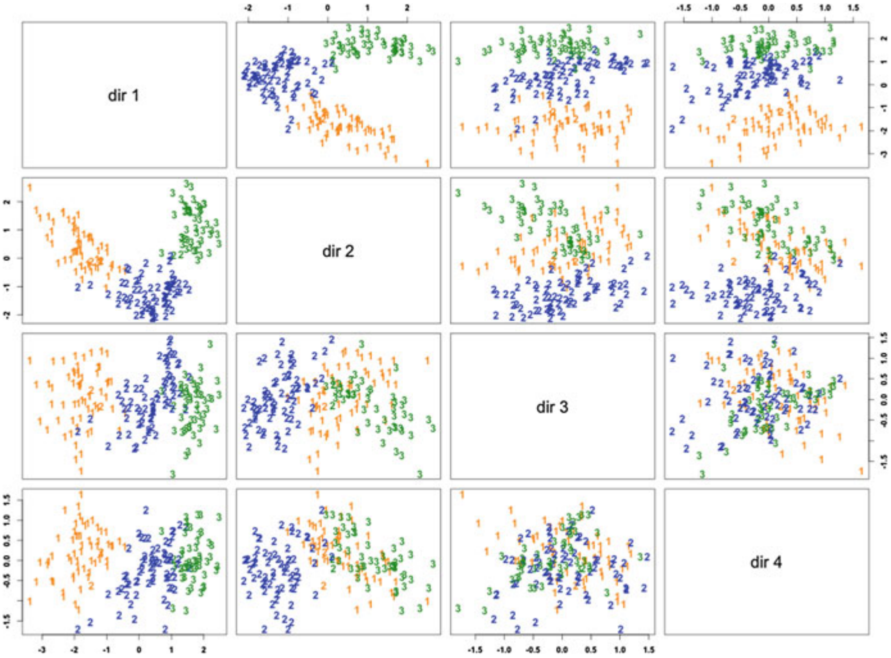

Fig. 2 Plots of estimated tMMDR directions for the wine data. The labels of the observations indicate their true cluster classification and the *colour* gives their estimated tMMDR cluster allocation (colour figure online)

4.2.2 Crabs data

Campbell and Mahon (1974) recorded five measurements for specimens of *Leptograpsus* crabs found in Australia. Crabs were classified according to their colour (blue or orange) and their gender. This data set, which consists of 200 observations, is available through the R package `MASS` (Venables and Ripley 2002). The known classification of the crabs data by colour and gender has four groups: 50 blue/orange males and 50 blue/orange females. The tMMDR method (ARI $= 0.86$) produces a better clustering than the GMMDR method (ARI $= 0.82$) on the crabs data but it requires one more feature than GMMDR. The resulting classification is presented in Table 7, while the

Table 7 A classification table for the best tMMDR model fitted to the crabs data

	1	2	3	4
Female				
Blue	50	8	0	0
Orange	0	42	0	0
Male				
Blue	0	0	47	0
Orange	0	0	3	50

Table 8 Summary of results for the best models fitted to the crabs data

Method	ARI	Components	Features
tMMDR	0.86	4	4
GMMDR	0.82	4	3
`FisherEM`	0.03	4	3
selvarclust	0.84	4	4
`clustvarsel`	0.76	5	4
ROBPCA with t-mixtures	0.82	4	5
MMtFA	0.61	3	3

procedure comparison appearing in Table 8 shows that tMMDR outperforms all other algorithms.

Figure 3 illustrates a scatterplot of the estimated tMMDR directions corresponding to the four clusters found by the procedure. The separation between the crabs species is most clear in the plots of direction 1 against direction 3. As the number of clusters increases, it becomes more difficult to visualize their separation as evidenced in the plots of direction 2 against directions 3 and 4. Looking at Fig. 3, it is interesting to compute the ARI for each set of directions and compare the values to those obtained by GMMDR. Using tMMDR, direction 1 has ARI $= 0.1394$, directions 1 and 2 have ARI $= 0.4507$, directions 1–3 have ARI $= 0.5556$, and directions 1–4 have ARI $= 0.8617$. By comparison, using GMMDR, direction 1 has ARI $= 0.5342$, directions 1 and 2 have ARI $= 0.7738$, and directions 1–3 have ARI $= 0.8195$. We observe that, for up to three directions, the ARI for tMMDR is less than those directions computed for GMMDR, but when the 4th direction is used for tMMDR, the ARI gets larger. Thus, adding this 'extra' direction improves accuracy.

4.2.3 Diabetes data

Reaven and Miller (1979) examined the relationship between five measures of blood plasma glucose and insulin in order to classify people as normal, overt diabetic, or chemical diabetic. This data set consists of observations from 145 adult patients at the Stanford Clinical Research Centre and is available through the R package `locfit` (Loader 2012). There are 76 normal patients, 36 chemical diabetics, and 33 overt diabetics; their tMMDR classifications are given in Table 9. Table 10 indicates that the selvarclust algorithm (ARI = 0.81) outperforms both the tMMDR (ARI = 0.70)

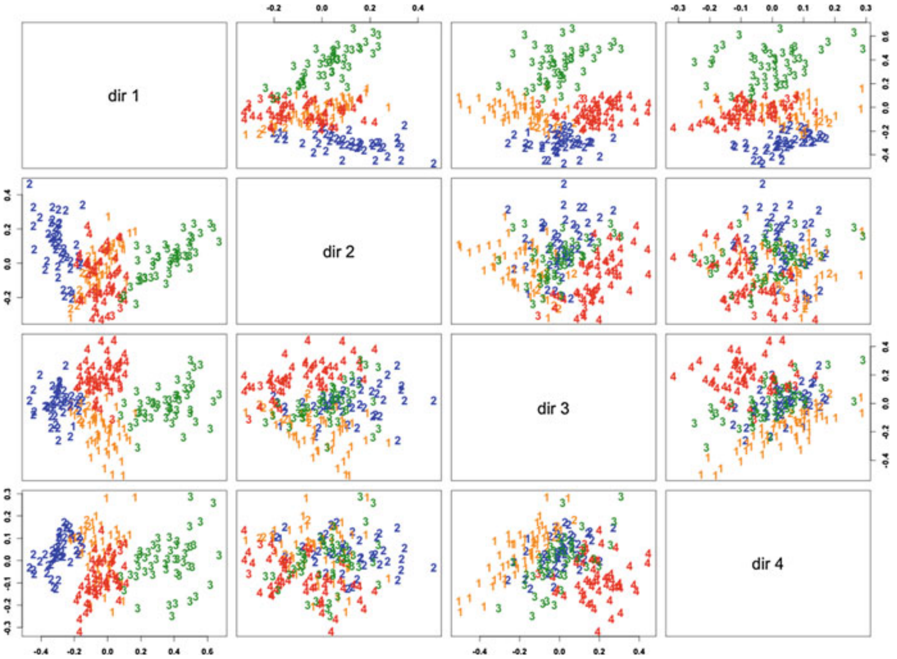

Fig. 3 Plots of estimated tMMDR directions for the crabs data. The labels of the observations indicate their true cluster classification and the *colour* gives their estimated tMMDR cluster allocation (colour figure online)

Table 9 A classification table for the best tMMDR model fitted to the diabetes data

	1	2	3
Overt	26	0	0
Chemical	7	27	2
Normal	0	9	74

and GMMDR methods (ARI = 0.65), while the rest of the procedures do not do particularly well on these data.

Figure 4 illustrates the overlap between the predicted groups. In particular, the chemical group is sometimes classified into the normal or overt group, but normal is never wrongly classified as overt or *vice-versa*. The average uncertainty associated with each identified cluster is as follows: 0.0106 for overt, 0.0524 for chemical, and 0.0158 for normal. The modest ARI is likely due to the nature of chemical diabetes.

4.2.4 SRBCT data

Khan et al. (2001) used microarray experiments of small round blue cell tumours (SRBCT) to classify childhood cancer. Their data contained 2308 genes and 83 tissue samples which were used to measure gene expressions in four types of tumours, namely Ewing sarcoma (EWS), Burkitt lymphoma (BL), neuroblastoma (NB), and rhabdomyosarcoma (RMS). The SRBCTs of childhood are given this collective name

Table 10 Summary of results for the best models fitted to the diabetes data

Method	ARI	Components	Features
tMMDR	0.70	3	4
GMMDR	0.65	3	3
`FisherEM`	0.54	3	2
selvarclust	0.81	3	4
`clustvarsel`	0.66	3	4
ROBPCA with t-mixtures	0.59	4	5
MMtFA	0.67	3	1

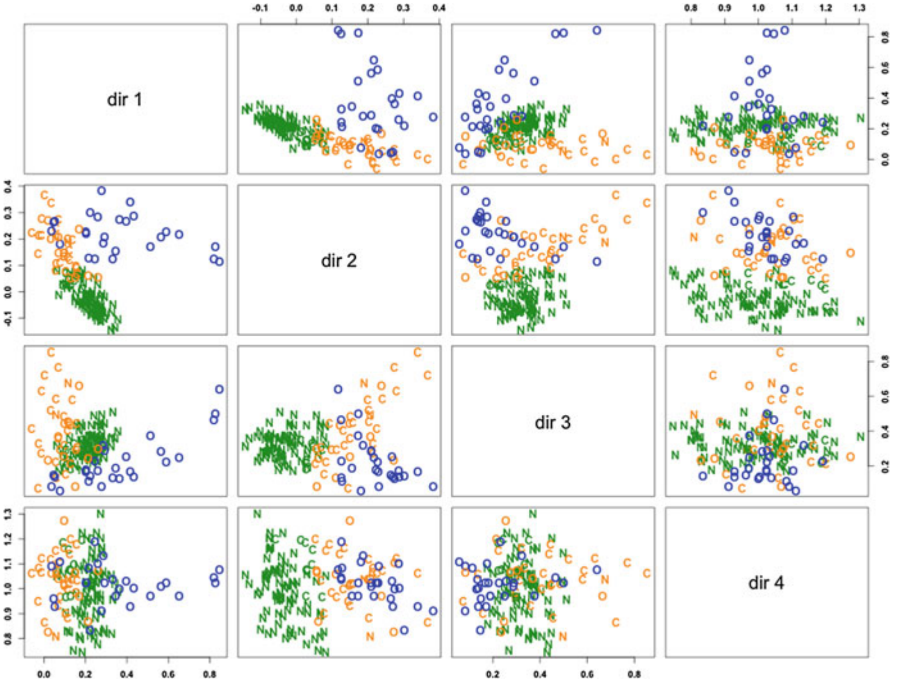

Fig. 4 Plots of estimated tMMDR directions for the diabetes data. The *letters* of the observations indicate their true cluster classification and the *colour* gives their estimated tMMDR cluster allocation (colour figure online)

because of their similar appearance on routine histology. These data are available via the R package `plsgenomics` (Boulesteix et al. 2011).

One of the main challenges in clustering these data is the large number of gene expression levels compared with the small number of cancer samples. As is the case with microarrays in general, there may be many non-informative genes which could hinder the clustering of the cancer clusters. Thus, gene filtering is very important for dimensionality reduction and further analysis. Khan et al. (2001) identified 96 genes which were useful in the classification of tissue samples, while Tibshirani et al (2002) isolated 43 such genes. We used gene-filtering based on t-tests and identified 36

Table 11 A classification table for the best tMMDR model fitted to the SRBCT data

	1	2	3	4
Ewing sarcoma	28	0	0	0
Burkitt lymphoma	0	11	0	0
Neuroblastoma	0	0	18	0
Rhabdomyosarcoma	1	0	0	25

Table 12 Summary of results for the best models fitted to the SRBCT data

Method	ARI	Components	Features
tMMDR	0.96	4	7
GMMDR	0.83	6	5
`FisherEM`	0.96	4	3
selvarclust	0.72	6	17
`clustvarsel`	0.80	6	36
ROBPCA with t-mixtures	0.80	5	7
MMtFA	0.39	2	3

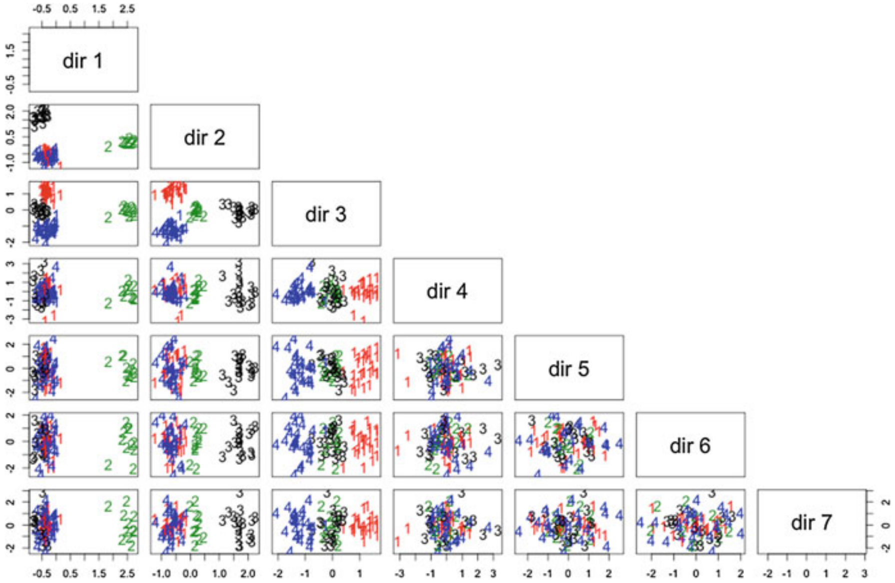

Fig. 5 Plots of the estimated tMMDR directions for the SRBCT data. The *labels* of the observations indicate their true cluster classification and the *colour* gives their estimated tMMDR cluster allocation (colour figure online)

differentially expressed genes which were then analyzed via our dimension reduction procedure.

The resulting classification (Table 11) and algorithm comparisons (Table 12) show that the tMMDR and `FisherEM` procedures (both with ARI = 0.96) outperform the

Table 13 Summary results for the *t*MMDR and GMMDR approaches on all five real data sets

Data	Dim.	Method	Model	Comp.	Deg. of freedom	Feat.	ARI
`wine`	13	GMMDR	VEV	3	–	5	0.85
		*t*MMDR	CUCC	3	58	4	0.9309
`crabs`	5	GMMDR	EEV	4	–	3	0.8195
		*t*MMDR	CUCU	4	{157, 7, 60, 97}	4	0.8617
`diabetes`	5	GMMDR	VEV	3	–	3	0.6536
		*t*MMDR	UUUC	3	66	4	0.702
`SRBCT`	36	GMMDR	VEI	6	–	5	0.8325
		*t*MMDR	CICU	4	{63,15,66,64}	7	0.96

GMMDR method (ARI = 0.8), which does not identify the correct number of clusters. Figure 5 depicts a scatterplot of the estimated *t*MMDR directions corresponding to the four clusters found by the procedure. The separation between the tumour classes is clear in the plots of directions 2 and 3 against the others, as they clearly reveal the underlying structure.

4.2.5 Summary

The results of our real data analyses are summarized in Table 13. Clearly, the *t*MMDR approach has outperformed the GMMDR method on these data. The *t*MMDR approach also performed very well when benchmarked against other well established dimension reduction procedures.

5 Conclusion

This paper introduced an effective dimension reduction technique for model-based clustering within the multivariate *t*-distribution framework. Our method, known as *t*MMDR, focused on identifying the smallest subspace of the data that captured the inherent cluster structure. The *t*MMDR approach was illustrated using simulated and real data, where it performed favourably compared to its Gaussian analogue (GMMDR; Scrucca 2010), as well as five other dimension reduction methods (ROBPCA with *t*-mixtures, mixtures of *t*-factor analyzers, clustvarsel with Gaussian mixtures, selvarclust, and FisherEM).

Future work will focus on dimension reduction using distributions that account for skewness (e.g., Karlis and Santourian 2009; Lin 2010; Vrbik and McNicholas 2012; Franczak et al. 2012; Lee and McLachlan 2013). However, it is not clear that the resulting methods would necessarily outperform the *t*MMDR method introduced herein.

Acknowledgments The authors thank Dr. Jeffrey Andrews for running the MM*t*FA models and providing the results reported herein. The authors are grateful to a guest editor and two anonymous reviewers for their very helpful comments and suggestions.

References

Andrews JL, McNicholas PD (2011a) Extending mixtures of multivariate t-factor analyzers. Stat Comput 21(3):361–373

Andrews JL, McNicholas PD (2011b) Mixtures of modified t-factor analyzers for model-based clustering, classification, and discriminant analysis. J Stat Plan Inference 141(4):1479–1486

Andrews JL, McNicholas PD (2012a) Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions: the tEIGEN family. Stat Comput 22(5):1021–1029

Andrews JL, McNicholas PD (2012b) teigen: model-based clustering and classification with the multivariate t-distribution. R package version 1.0

Andrews JL, McNicholas PD, Subedi S (2011) Model-based classification via mixtures of multivariate t-distributions. Comput Stat Data Anal 55(1):520–529

Baek J, McLachlan GJ (2011) Mixtures of common t-factor analyzers for clustering high-dimensional microarray data. Bioinformatics 27:1269–1276

Banfield JD, Raftery AE (1993) Model-based Gaussian and non-Gaussian clustering. Biometrics 49(3): 803–821

Boulesteix AL, Lambert-Lacroix S, Peyre J, Strimmer K (2011) plsgenomics: PLS analyses for genomics. R package version 1.2-6

Bouveyron C, Brunet C (2012) Simultaneous model-based clustering and visualization in the Fisher discriminative subspace. Stat Comput 22(1):301–324

Campbell NA, Mahon RJ (1974) A multivariate study of variation in two species of rock crab of genus leptograpsus. Aust J Zoo l 22:417–425

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

Dean N, Raftery AE (2009) clustvarsel: Variable selection for model-based clustering. R package version 1.3

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Royal Stat Soc 39(1):1–38

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

Fraley C, Raftery AE (1999) MCLUST: software for model-based cluster analysis. J Classif 16:297–306

Franczak B, Browne RP, McNicholas PD (2012) Mixtures of shifted asymmetric Laplace distributions. Arxiv, preprint arXiv:1207.1727v3

Greselin F, Ingrassia S (2010a) Constrained monotone EM algorithms for mixtures of multivariate t-distributions. Stat Comput 20(1):9–22

Greselin F, Ingrassia S (2010b) Weakly homoscedastic constraints for mixtures of t-distributions. In: Fink A, Lausen B, Seidel W, Ultsch A (eds) Advances in Data Analysis, Data Handling and Business Intelligence. Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin

Hubert L, Arabie P (1985) Comparing partitions. J Classifi 2:193–218

Hubert M, Rousseeuw PJ, Vanden Branden K (2005) ROBPCA: a new approach to robust principal components analysis. Technometrics 47:64–79

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

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

Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 7:673–679

Lee SX, McLachlan GJ (2013) On mixtures of skew normal and skew t-distributions. Arxiv, preprint arXiv:1211.3602v3

Li KC (1991) Sliced inverse regression for dimension reduction (with discussion). J Am Stat Assoc 86: 316–342

Li KC (2000) High dimensional data analysis via the SIR/PHD approach, unpublished manuscript. http://www.stat.ucla.edu/~kcli/sir-PHD.pdf

Lin TI (2010) Robust mixture modeling using multivariate skew t-distributions. Stat Comput 20:343–356

Loader C (2012) locfit: Local Regression, Likelihood and Density Estimation. R package version 1.5-8

Maugis C, Celeux G, Martin-Magniette ML (2009) Variable selection for clustering with Gaussian mixture models. Biometrics 65:701–709

McLachlan GJ, Krishnan T (2008) The EM algorithm and extensions, 2nd edn. Wiley, New York

McLachlan GJ, Bean RW, Jones LT (2007) Extension of the mixture of factor analyzers model to incorporate the multivariate t-distribution. Comput Stat Data Anal 51(11):5327–5338

McNicholas PD (2013) Model-based clustering and classification via mixtures of multivariate t-distributions. In: Giudici P, Ingrassia S, Vichi M (eds) Statistical models for data analysis, studies in classification, data analysis, and knowledge organization. Springer International Publishing, Switzerland

McNicholas PD, Murphy TB (2008) Parsimonious Gaussian mixture models. Stat Comput 18:285–296

McNicholas PD, Murphy TB (2010) Model-based clustering of microarray expression data via latent Gaussian mixture models. Bioinformatics 26(21):2705–2712

McNicholas PD, Subedi S (2012) Clustering gene expression time course data using mixtures of multivariate t-distributions. J Stat Plan Inference 142(5):1114–1127

Meng XL, Rubin DB (1993) Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika 80:267–278

Peel D, McLachlan GJ (2000) Robust mixture modelling using the t-distribution. Stat Comput 10:339–348

Qiu WL, Joe H (2006) Generation of random clusters with specified degree of separation. J Classifi 23(2):315–334

R Development Core Team (2012) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org

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

Reaven GM, Miller RG (1979) An attempt to define the nature of chemical diabetes using a multidimensional analysis. Diabetologia 16:17–24

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

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

Steane MA, McNicholas PD, Yada R (2012) Model-based classification via mixtures of multivariate t-factor analyzers. Commun Stat Simul Comput 41(4):510–523

Tibshirani R, Hastie T, Narasimhan B, Chu G (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Nat Acad Sci USA 99(10):6567–6572

Todorov V, Filzmoser P (2009) An object-oriented framework for robust multivariate analysis. J Stat Softw 32(3):1–47. http://www.jstatsoft.org/v32/i03/

Venables WN, Ripley BD (2002) Modern applied statistics with S, 4th edn. Springer, New York. http://www.stats.ox.ac.uk/pub/MASS4

Vrbik I, McNicholas PD (2012) Analytic calculations for the EM algorithm for multivariate skew-mixture models. Stat Prob Lett 82(6):1169–1174