Subspace Clustering with the Multivariate-t Distribution

SUBSPACE CLUSTERING
WITH THE MULTIVARIATE-T DISTRIBUTION

BY

ANGELINA PESEVSKI, B.Sc.

A THESIS

SUBMITTED TO THE DEPARTMENT OF MATHEMATICS & STATISTICS

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

Master of Science (2017) McMaster University

(Mathematics & Statistics) Hamilton, Ontario, Canada

TITLE:	Subspace Clustering with the Multivariate-t Distribution
AUTHOR:	Angelina Pesevski B.Sc., (Mathematics and Statistics) McMaster University, Hamilton, Canada
SUPERVISOR:	Dr. Paul D. McNicholas Dr. Brian C. Franczak
NUMBER OF PAGES:	ix, 39

To my loving and supporting partner, Kiril, as well as the individuals who gave me this opportunity, my parents, Liljana and Bobi.

Abstract

Clustering procedures suitable for the analysis of very high-dimensional data are needed for many modern data sets. One model-based clustering approach called high-dimensional data clustering (HDDC) uses a family of Gaussian mixture models to model the sub-populations of the observed data, i.e., to perform cluster analysis. The HDDC approach is based on the idea that high-dimensional data usually exists in lower-dimensional subspaces; as such, the dimension of each subspace, called the intrinsic dimension, can be estimated for each sub-population of the observed data. As a result, each of these Gaussian mixture models can be fitted using only a fraction of the total number of model parameters. This family of models has gained attention due to its superior classification performance compared to other families of mixture models; however, it still suffers from the usual limitations of Gaussian mixture model-based approaches. Herein, a robust analogue of the HDDC approach is proposed. This approach, which extends the HDDC procedure to include the mulitvariate-t distribution, encompasses 28 models that rectify one of the major shortcomings of the HDDC procedure. Our tHDDC procedure is fitted to both simulated and real data sets and is compared to the HDDC procedure using an image reconstruction problem that arose from satellite imagery of Mars' surface.

Acknowledgements

First and foremost, I'd like to thank my supervisor Professor Paul D. McNicholas. Thank you for giving me this opportunity and teaching me to push myself to new limits. I will forever be grateful for your encouraging words that got me to where I am today.

I would also like to thank Dr. Brian C. Franczak for all the guidance and support throughout this project. I cannot even begin to express my gratitude for all of the hours you have dedicated to me, for the daily coding sessions and the occasional pep talk. You have been a great mentor and friend, and I am glad that I will have the pleasure to continue working under your supervision.

To my parents, Lile and Bobi, you have dedicated your lives to your children's academic and personal success, and for that I will be forever thankful. Even if I become half the parent and friend you both are, I know that I will have succeeded in my life. Thank you for giving me a chance at this life and for always believing in me.

To my sister, thank you for all the encouragement and positivity you radiate on a daily basis. I have had an easier life because you always got to go first and I learned from your mistakes, especially when it comes to education.

Last but definitely not least, I would like to thank my partner, Kire. You have always given me that burst of energy when I was frustrated with my code or when I

was in a writing rut. You have given me confidence and made me understand that I can accomplish anything I set my mind to. The following years of studies are not going to be easy, but I am welcoming them with open arms, all because of you.

Contents

List of Figures

Chapter 1

Introduction

Cluster analysis refers to the practice of using statistical approaches to detect subgroups within a given data set. These subgroups can represent a physical attribute not described by the given explanatory variables, e.g., gender, income tax bracket or blood type, which can reveal important relationships among the observed data and may be a crucial component in the effective analysis of a given data set. Due to their construction, finite mixture models are very useful when modelling data that contain a finite collection of sub-populations because each component of the model can be used to represent one of these sub-populations. Reviews of the application of finite mixture models for clustering are given by Fraley and Raftery (2002); Bouveyron and Brunet-Saumard (2014) and McNicholas (2016b), and extensive details can be found in the monographs by McLachlan and Peel (2000) and McNicholas (2016a).

The current influx of information from, e.g., social media sources, has resulted in many modern data sets having the characteristics of big data (Puts *et al.*, 2015). As such, there is an increasing need for statistical methods that can handle these large data sets. In Chapter 2, we provide a review of popular finite mixture-models for

handling big data sets, specifically those with large dimension, and highlight a high-dimensional data clustering (HDDC) approach recently introduced by Bouveyron *et al.* (2007). This approach has gained attention due to its parsimonious parameter estimation scheme. It has been shown to work well for very high-dimensional real data of more than 200 variables. However, it still suffers from the usual limitations of the Gaussian mixture model based approaches, e.g., it is not robust to outliers. The focus of this thesis is the expansion of this HDDC approach to include a multivariate distribution that is robust to outliers.

A formal outline of this thesis follows: in Chapter 2, we provide background information on finite mixture-models and summarize related work, in Chapter 3, we outline the derivation of a multivariate-t high-dimensional data clustering (tHDDC) approach and describe a parameter estimation scheme to fit the resulting models, in Chapter 4 we assess the classification performance of this novel family of models using a simulation study and three real data sets, and in Chapter 5 we conclude with a discussion and suggestions for future work.

Chapter 2

Background

2.1 Normal-Variance Mixtures

Define a real random variable U following a probability distribution H, where $H \in [0, \infty)$ and a p-dimensional random variable $\mathbf{X}$. We can say that $\mathbf{X}$ follows a normal-variance mixture if $\mathbf{X} \mid U = u$ follows a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $u\boldsymbol{\Sigma}$, where $\boldsymbol{\mu}$ is a location parameter and $\boldsymbol{\Sigma}$ is a $p \times p$ covariance matrix (Barndorff-Nielsen *et al.*, 1982). Formally, a random vector $\mathbf{X}$ arising from a normal-variance mixture has density

$$f(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int_0^\infty \phi_p \left(\mathbf{x} \mid u \, ; \boldsymbol{\mu}, u\boldsymbol{\Sigma}\right) h(u) du, \tag{2.1}$$

where $\phi_p \left(\mathbf{X} \mid u \, ; \boldsymbol{\mu}, u\boldsymbol{\Sigma}\right)$ is the density function of a p-variate Gaussian distribution with location parameter $\boldsymbol{\mu}$ and covariance matrix $u\boldsymbol{\Sigma}$. There are several distributions that arise as mixtures of Normal distributions, one being the multivariate-t distribution (Kotz and Nadarajah, 2004).

2.2 Finite Mixture-Models

The density of a parametric finite mixture distribution is

$$f(\mathbf{x} \,|\, \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g \, p_g(\mathbf{x} \,|\, \boldsymbol{\theta}_g), \tag{2.2}$$

where $\pi_g > 0$, such that $\sum_{g=1}^{G} \pi_g = 1$, are called mixing proportions, $p_g(\mathbf{x} \,|\, \boldsymbol{\theta}_g)$ are the component densities and $\boldsymbol{\vartheta} = (\pi_1, \ldots, \pi_G, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G)$ is a vector containing the model parameters. Herein, we follow convention and refer to the application of finite mixture models for clustering as model-based clustering.

As described in McNicholas (2016a), there are many different definitions of a cluster. A common definition discussed by Wolfe (1963) is one where a cluster is described as a group of observations that are more similar to each other than to observations outside that particular group. A more rigorous definition, given by Wolfe (1963), requires the use of finite mixture models and defines a cluster as being one of the components of a mixture of distributions.

2.3 Gaussian Mixture-Model

Each component of the density of the general finite mixture model, given in (2.2), can be specified to be any univariate or multivariate probability distribution. Until the last decade or so, the majority of work on model-based clustering using multivariate component densities focused on the Gaussian mixture-model. A random variable $\mathbf{X} \in \mathbb{R}^p$ arising from a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance

matrix $\boldsymbol{\Sigma}$ has density

$$\phi_p\left(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}\right)' \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right\}. \tag{2.3}$$

It follows that the density of a mixture of multivariate Gaussian distributions is

$$f\left(\mathbf{x} \mid \pi_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right) = \sum_{g=1}^{G} \pi_g \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_g|^{1/2}} \exp\left\{-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_g\right)' \boldsymbol{\Sigma}_g^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_g\right)\right\}, \tag{2.4}$$

where π_g is defined for (2.2), and $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ are the component location parameter and component covariance matrix, respectively. In total, the general Gaussian mixture model has $(G-1)+Gp+Gp(p+1)/2$ free parameters.

2.4 Mixtures of Multivariate-t Distributions

One of the first notable departures from Gaussianity was provided by Peel and McLachlan (2000), who utilized mixtures of multivariate-t distributions for clustering. As the name suggests, mixtures of multivariate-t distributions assume that each sub-population of the observed data follow the multivariate-t distribution. Formally, the density of a mixture of multivariate-t distribution is formulated by writing the component density in (2.2) as

$$p_g\left(\mathbf{x} \mid \boldsymbol{\theta}_g\right) = f_t(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \nu_g) = \frac{\Gamma\left[\left(\nu_g+p\right)/2\right]|\boldsymbol{\Sigma}_g|^{-1/2}(\pi\nu_g)^{-p/2}}{\Gamma\left[\nu_g/2\right]\left[1+\delta\left(\mathbf{x}, \boldsymbol{\mu}_g \mid \boldsymbol{\Sigma}_g\right)/\nu_g\right]^{(\nu_g+p)/2}}, \tag{2.5}$$

where $\Gamma(\cdot)$ is the Gamma function, p is the number of dimensions in the observed data set, $\boldsymbol{\mu}_g$ is the component location parameter, $\boldsymbol{\Sigma}_g$ is the component covariance

matrix, ν_g parameterizes the degrees of freedom in each component, and

$$\delta(\mathbf{x}, \boldsymbol{\mu}_g \mid \boldsymbol{\Sigma}_g) = \left(\mathbf{x} - \boldsymbol{\mu}_g\right)' \boldsymbol{\Sigma}_g^{-1} \left(\mathbf{x} - \boldsymbol{\mu}_g\right) \tag{2.6}$$

is the squared Mahalanobis distance between $\mathbf{x}$ and $\boldsymbol{\mu}_g$ for $g = 1, \ldots, G$. Despite rectifying a well known shortcoming of the Gaussian mixture model by formulating a model that is robust to outliers, mixtures of multivariate-t distributions have only gained popularity in the last few years (see McLachlan *et al.*, 2007; Andrews and McNicholas, 2011a,b; Baek and McLachlan, 2011; Steane *et al.*, 2012; Andrews and McNicholas, 2012; Lin *et al.*, 2014, for examples).

2.5 Gaussian Parsimonious Clustering Models

A family of mixture models emerges when we introduce constraints on the component densities. Some families of Gaussian mixture models are well established and widely used, e.g., the Gaussian Parsimonious Clustering Models (GPCMs; Celeux and Govaert, 1995) which arise from constraints being imposed on the eigen-decomposed covariance structure in a Gaussian mixture model. This eigen-decomposition is

$$\boldsymbol{\Sigma}_g = \lambda_g \mathbf{D}_g \mathbf{A}_g \mathbf{D}_g^{'}, \tag{2.7}$$

where $\boldsymbol{\Sigma}_g$ is the component covariance matrix, λ_g is a constant, $\mathbf{D}_g$ is a matrix of eigenvectors, and $\mathbf{A}_g$ is a diagonal matrix with $|\mathbf{A}_g| = 1$ and entries proportional to the eigenvalues of $\boldsymbol{\Sigma}_g$. Applying a combination of the constraints: $\lambda_g = \lambda, \mathbf{A}_g = \mathbf{A}, \mathbf{D}_g = \mathbf{D}, \mathbf{D}_g = \mathbf{I}$, $\mathbf{A}_g = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix of the appropriate

dimension, across the components of the Gaussian mixture model results in a family of fourteen models and allows for various shapes and sizes of clusters (see Table 2.1). In more than half of these fourteen models there are $\mathcal{O}(p^2)$ free parameters to be estimated; hence, with higher dimensions, it can be very computationally inefficient to use these models. The GPCMs are supported by the R packages `mclust` (Fraley and Raftery, 2002) and `mixture` (Browne *et al.*, 2015).

Table 2.1: Nomenclature, component volume, shape and orientation, covariance structure, and number of free covariance parameters for each member of the GPCM family.

Model	Volume	Shape	Orientation	$\boldsymbol{\Sigma}_g$	Free cov. parameters
EII	Equal	Spherical	–	$\lambda\mathbf{I}$	1
VII	Variable	Spherical	–	$\lambda_g\mathbf{I}$	G
EEI	Equal	Equal	Axis-Aligned	$\lambda\mathbf{A}$	p
VEI	Variable	Equal	Axis-Aligned	$\lambda_g\mathbf{A}$	$p+G-1$
EVI	Equal	Variable	Axis-Aligned	$\lambda\mathbf{A}_g$	$pG-G+1$
VVI	Variable	Variable	Axis-Aligned	$\lambda_g\mathbf{A}_g$	pG
EEE	Equal	Equal	Equal	$\lambda\mathbf{DAD}'$	$p(p+1)/2$
EEV	Equal	Equal	Variable	$\lambda\mathbf{D}_g\mathbf{AD}_g'$	$Gp(p+1)/2-(G-1)p$
VEV	Variable	Equal	Variable	$\lambda_g\mathbf{D}_g\mathbf{AD}_g'$	$Gp(p+1)/2-(G-1)(p-1)$
VVV	Variable	Variable	Variable	$\lambda_g\mathbf{D}_g\mathbf{A}_g\mathbf{D}_g'$	$Gp(p+1)/2$
EVE	Equal	Variable	Equal	$\lambda\mathbf{DA}_g\mathbf{D}'$	$p(p+1)/2+(G-1)(p-1)$
VVE	Variable	Variable	Equal	$\lambda_g\mathbf{DA}_g\mathbf{D}'$	$p(p+1)/2+(G-1)p$
VEE	Variable	Equal	Equal	$\lambda_g\mathbf{DAD}'$	$p(p+1)/2+(G-1)$
EVV	Equal	Variable	Variable	$\lambda\mathbf{D}_g\mathbf{A}_g\mathbf{D}_g'$	$Gp(p+1)/2-(G-1)$

2.6 tEIGEN Models

The multivariate-t analog of the GPCM family for the mixtures of multivariate-t distributions is the tEIGEN family (Andrews and McNicholas, 2012). These models use the same eigen-decomposition as the GPCM family and therefore the same constraints mentioned above can be applied, in addition to $\nu_g = \nu$. It is important to note that in this case, it is not the covariance matrix that is decomposed, but rather

the scale matrix. Furthermore, as explained in (Vrbik and Mcnicholas, 2014), the geometric interpretation of the component shapes in non-Gaussian mixture-models is not the same as with the Gaussian cases.

By combining the aforementioned constraints, a total of 28 different models are derived. In R, all 28 models are supported by the `teigen` package (Andrews and McNicholas, 2014; Andrews *et al.*, 2017).

2.7 Parsimonious Gaussian Mixture Models

Another popular family of mixture models are the parsimonious Gaussian mixture models (PGMMs; McNicholas and Murphy, 2008). These models are an extension of the mixture of factor analyzers (Ghahramani and Hinton, 1997) whose component covariance matrices are written as

$$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g, \tag{2.8}$$

where $\boldsymbol{\Lambda}_g$ is a $p \times q$ loading matrix with $q < p$, and $\boldsymbol{\Psi}_g$ is a diagonal $p \times p$ matrix with positive entries for $g = 1, \ldots, G$. By imposing constraints on $\boldsymbol{\Lambda}_g$ and $\boldsymbol{\Psi}_g$ across the components, McNicholas and Murphy (2008) introduced eight parsimonious models in which the number of free parameters is $\mathcal{O}(p)$ so that the number of covariance parameters grows linearly with dimension. For this reason, these models are more appropriate than the GPCMs for high-dimensional data. These models can be implemented via the `pgmm` package for R (McNicholas *et al.*, 2015). Note: McNicholas

and Murphy (2010) extended the PGMMs to include four new models by setting

$$\boldsymbol{\Psi}_g = \omega_g \boldsymbol{\Delta}_g, \tag{2.9}$$

where $\omega_g \in \mathbb{R}^+$ and $\boldsymbol{\Delta}_g = \text{diag}\{\delta_1, \delta_2, \ldots, \delta_p\}$ is a noise matrix, such that $|\boldsymbol{\Delta}_g| = 1$ (see Table 2.2).

Table 2.2: Nomenclature, component covariance matrix structure, and number of free covariance parameters for each parsimonious Gaussian mixture models.

PGMM Nomenclature					
$\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$	$\boldsymbol{\Delta}_g = \boldsymbol{\Delta}$	$\omega_g = \omega$	$\boldsymbol{\Delta}_g = \mathbf{I}_p$	Component Covariance Matrix	Number of Free Scale Parameters
C	C	C	C	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \omega\mathbf{I}_p$	$[pq - q(q-1)/2] + 1$
C	C	U	C	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \omega_g\mathbf{I}_p$	$[pq - q(q-1)/2] + G$
U	C	C	C	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \omega\mathbf{I}_p$	$G[pq - q(q-1)/2] + 1$
U	C	U	C	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \omega_g\mathbf{I}_p$	$G[pq - q(q-1)/2] + G$
C	C	C	U	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \omega\boldsymbol{\Delta}$	$[pq - q(q-1)/2] + p$
C	C	U	U	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \omega_g\boldsymbol{\Delta}$	$[pq - q(q-1)/2] + [G + (p-1)]$
U	C	C	U	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \omega\boldsymbol{\Delta}$	$G[pq - q(q-1)/2] + p$
U	C	U	U	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \omega\boldsymbol{\Delta}_g$	$G[pq - q(q-1)/2] + [G + (p-1)]$
C	U	C	U	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \omega\boldsymbol{\Delta}_g$	$[pq - q(q-1)/2] + [1 + G(p-1)]$
C	U	U	U	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \omega_g\boldsymbol{\Delta}_g$	$[pq - q(q-1)/2] + Gp$
U	U	C	U	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \omega\boldsymbol{\Delta}_g$	$G[pq - q(q-1)/2] + [1 + G(p-1)]$
U	U	U	U	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \omega_g\boldsymbol{\Delta}_g$	$G[pq - q(q-1)/2] + Gp$

2.8 Mixture of Multivariate-t Factor Analyzers

The multivariate-t analogue of the PGMMs, known as the mixture of multivariate t-factor analyzers (MMtFA) were introduced by Andrews and McNicholas (2011a,b). In the MMtFAs, the component scale stucture is also parameterized as $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g$. By applying the constraints: $\boldsymbol{\Psi}_g = \psi_g\mathbf{I}$, $\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$, and $\nu_g = \nu$, Andrews and McNicholas (2011a) created a family of six models, whose covariance parameters grow linearly with p, and Andrews and McNicholas (2011b) extended this to a family of 24 models. It is worth noting that the probabilistic principal t-component analyzer model, MPtCA, is a special case of the MMtFA model, where $\boldsymbol{\Psi}_g = \psi_g\mathbf{I}$. This family

of 24 models is supported by the `mmtfa` package (Andrews *et al.*, 2015).

2.9 High-Dimensional Data Clustering

Bouveyron *et al.* (2007) proposed a high-dimensional data clustering (HDDC) technique that is also based on an eigen-decomposition of the covariance structure of the Gaussian mixture model. This technique projects the observed data into a lower-dimensional subspace spanned by a subset of the eigenvectors of $\boldsymbol{\Sigma}_g$. Formally, given a data set $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of n data points in $\mathbb{R}^p$ with G sub-populations, this method assumes that high-dimensional data mostly rests in lower-dimensional subspaces. This assumption can drastically reduce the number of covariance parameters that require estimation and result in an efficient parameter estimation scheme.

As with the GPCMs, Bouveyron *et al.* (2007) lets $\mathbf{D}_g$ be the orthogonal matrix of eigenvectors of $\boldsymbol{\Sigma}_g$, but instead considers a block-diagonal matrix, $\boldsymbol{\Delta}_g$, which contains the eigenvalues of $\boldsymbol{\Sigma}_g$. Formally, $\boldsymbol{\Delta}_g$ has the following form:

$$\boldsymbol{\Delta}_g = \left(\begin{array}{ccc|ccc} a_{1g} & & 0 & & & \\ & \ddots & & & \mathbf{0} & \\ 0 & & a_{d_g g} & & & \\ \hline & & & b_g & & 0 \\ & \mathbf{0} & & & \ddots & \\ & & & 0 & & b_g \end{array}\right), \tag{2.10}$$

where the upper left block is of size $d_g \times d_g$, $d_g \in \{1, p-1\}$ is the intrinsic dimension in each component, or cluster, and the lower right block is of size $(p - d_g) \times (p - d_g)$,

with $a_{jg} > b_g$, for $j = 1, \ldots, d_g$ and $g = 1, \ldots, G$. Bouveyron *et al.* (2007) proposed two methods for estimating the intrinsic dimension in each component of this eigen-decomposed GMM. The first approach utilizes the scree-test of Cattell (1966) and the second approaches utilizes the probabilistic Bayesian information Criterion (BIC; Schwarz, 1978), which is given by

$$\text{BIC} = 2l(\hat{\boldsymbol{\vartheta}}) - \rho \log n, \tag{2.11}$$

where ρ is the number of free parameters in the model, n is the number of observations, and $l(\hat{\boldsymbol{\vartheta}})$ is the maximized log-likelihood value. The eigenvectors associated with the eigenvalues a_{jg}, for $j = 1, \ldots, d_g$, span a subspace $\mathbb{E}_g \in \mathbb{R}^{d_g}$ for each cluster, such that $\boldsymbol{\mu}_g \in \mathbb{E}_g$. The affine subspace $\mathbb{E}_g^{\perp}$ is defined such that $\mathbb{E}_g \otimes \mathbb{E}_g^{\perp} = \mathbb{R}^p$ and $\boldsymbol{\mu}_g \in \mathbb{E}_g^{\perp}$. Each observation $\mathbf{x}_i$ is then projected onto the subspace $\mathbb{E}_g$, which is called the specific subspace of the gth group, because most of the data are assumed to live on or near this subspace. This decomposition leads to 28 possible models by constraining the parameters $a_{jg}, b_g, \mathbf{D}_g, d_g$ across the G components and d_g dimensions. Of these 28 models, 14 have been implemented in the `R` package `HDclassif` (Berge *et al.*, 2012). The fully unconstrained model will be referred to as UUUU and the letters C, G and D will denote completely constrained, constrained across groups and constrained across dimensions, respectively.

2.10 Model Selection and Performance Assesment

When fitting a family of mixture-models, it is necessary to choose the 'best' model. A very popular choice is the BIC, defined in Section 2.9. It is a widely used method and

has been strongly established in theoretical and applied work (Leroux *et al.*, 1992; Kass and Raftery, 1995; Kass and Wasserman, 1995; Keribin, 2000). It is important to note that the BIC can be used to approximate the Bayes factors (Kass and Raftery, 1995; Dasgupta and Raftery, 1998). In particular, Dasgupta and Raftery (1998) show that

$$2 \log p\left(\mathbf{x} \mid \boldsymbol{\vartheta}_g\right) \approx \mathrm{BIC}_G$$

where $p\left(\mathbf{x} \mid \boldsymbol{\vartheta}_g\right)$ is the integrated likelihood of the data using a G-component mixture model and BIC_G is the BIC value for the model with G groups. The difference in BIC values for two models is an approximation to the Bayes factor, assuming that the prior distributions of the two models are equal.

For performance assessment, a usual choice is the adjusted Rand index (ARI; Hubert and Arabie, 1985). The ARI is used to assess class agreement between the true class labels and the predicted labels rendered by the clustering techniques. It was originally introduced to correct the Rand index (Rand, 1971), which is given by

$$\mathrm{RI} = \frac{\text{number of pairwise agreements}}{\text{number of pairs}}, \tag{2.12}$$

where pairwise agreements refer to pairs of observations being correctly classified as coming from the same group and correctly classified as coming from different groups. The expected value of the Rand index is greater than 0 for a random classification, making it hard to interpret. Hence, the ARI has expected value equal to 0, with a perfect classification being represented by a score of 1. A negative ARI means that the classification performance is worse than would be expected under random classification.

Chapter 3

Methodology

We now lay out some groundwork for a novel HDDC approach that utilizes mixtures of multivariate-t distributions. This approach, which is the t-analouge of the HDDC models described in Section 2.9, rectifies a well-known shortcoming of the Gaussian mixture model-based approach to cluster analysis by incorporating a distribution that is robust to outliers. Like with all the clustering procedures introduced so far, our goal is to cluster a given data set $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ in $\mathbb{R}^p$ into G homogeneous groups.

Herein, we use a mixture of multivariate-t distributions, whose density is given by

$$f\left(\mathbf{x} \mid \pi_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \nu_g\right) = \sum_{g=1}^{G} \pi_g \frac{\Gamma\left[\left(\nu_g + p\right)/2\right] |\boldsymbol{\Sigma}_g|^{-1/2} (\pi\nu_g)^{-p/2}}{\Gamma\left[\nu_g/2\right] \left[1 + \delta\left(\mathbf{x}, \boldsymbol{\mu}_g \mid \boldsymbol{\Sigma}_g\right)/\nu_g\right]^{(\nu_g+p)/2}}, \tag{3.1}$$

where all model parameters are defined for equations (2.2), (2.4) and (2.5).

The general multivariate-t mixture model requires the estimation of the full covariance structure, so the number of parameters to estimate is $\mathcal{O}(p^2)$. As Bouveyron *et al.* (2007) describe, via the *empty space* phenomenon (Scott and Thompson, 1983), we can assume that most of the data live around lower-dimensional subspaces. By

performing clustering in these lower-dimensional subspaces, the number of parameters to be estimated is reduced significantly.

3.1 Formulating the tHDDC Approach

Analogous to the HDDC approach, we specify $\mathbf{D}_g$ to be the orthogonal matrix of eigenvectors and

$$\mathbf{\Delta}_g = \mathbf{D}_g' \mathbf{\Sigma}_g \mathbf{D}_g, \tag{3.2}$$

where $\mathbf{\Delta}_g$ is a class specific matrix of the form given in (2.10). Following Bouveyron *et al.* (2007) we define

$$P_g(\mathbf{x}) = \tilde{\mathbf{D}}_g \tilde{\mathbf{D}}_g'(\mathbf{x} - \boldsymbol{\mu}_g) + \boldsymbol{\mu}_g \quad \text{and} \quad P_g^{\perp}(\mathbf{x}) = \bar{\mathbf{D}}_g \bar{\mathbf{D}}_g'(\mathbf{x} - \boldsymbol{\mu}_g) + \boldsymbol{\mu}_g \tag{3.3}$$

as the projection of $\mathbf{x}$ on $\mathbb{E}_g$ and the projection of $\mathbf{x}$ on $\mathbb{E}_g^{\perp}$, respectively, where $\tilde{\mathbf{D}}_g$ consists of the first d_g columns of $\mathbf{D}_g$, concatenated with $p - d_g$ zero columns, and $\bar{\mathbf{D}}_g = \mathbf{D}_g - \tilde{\mathbf{D}}_g$.

Formally, the covariance structure of each tHDDC mixture model has parameters $a_{jg}, b_g, \mathbf{D}_g, d_g, \nu_g$ for $j = 1, \ldots, d_g$ and $g = 1, \ldots, G$. It follows that applying group-wide constraints to the covariance parameters and allowing the degrees of freedom parameter, ν_g, to either vary or be constrained across G can lead to a total of 56 possible models. (The tHDDC analogues of the HDDC models available in `HDclassif` are listed in Table 3.1).

3.2 Parameter Estimation

In model-based clustering, the expectation-maximization (EM) algorithm (Dempster *et al.*, 1977; McLachlan and Krishnan, 2008) is the usual choice when estimating the parameters of a finite mixture model. As the name suggests, the EM algorithm is an iterative procedure that alternates between two steps: an E-step and a M-step, until convergence is reached. On the E-step, the expected value of the complete-data log-likelihood is updated given the current parameter estimates. In the M-step, the same complete-data log-likelihood is maximized in terms of the model parameters. We use a variation of the EM algorithm called the expectation conditional-maximization (ECM) algorithm (Meng and Rubin, 1993), which replaces each M-step with multiple CM-steps. For each tHDDC model the complete-data is made up of the observed $\mathbf{x}_i$, the latent u_{ig}, and the missing z_{ig} for $i = 1, \ldots, n$ and $g = 1, \ldots, G$. The z_{ig} are introduced to represent component membership. Formally, we write that

$$z_{ig} = \begin{cases} 1 & \text{if observation } \mathbf{x}_i \text{ belongs to component } g \\ 0 & \text{otherwise.} \end{cases} \tag{3.4}$$

The u_{ig} is a realization of a gamma distributed random variable, U_{ig}, that arises because we exploit the fact that the multivariate-t distribution is a normal-variance mixture as defined in Section 2.1. Formally, the random variable $U_{ig} \mid z_{ig} = 1$ follows a $\mathcal{G}(\nu/2, \nu/2)$ distribution (Peel and McLachlan, 2000). Therefore, it will have density

$$f(u; \nu/2, \nu/2) = \frac{(\nu/2)^{\nu/2}\, u^{\nu/2-1}}{\Gamma(\nu/2)} \exp\left[(-\nu/2)\, u\right]. \tag{3.5}$$

It follows that $U \mid \mathbf{X} \sim \mathcal{G}\left((\nu + p)/2, \frac{\nu + \delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma})}{2}\right)$.

3.2.1 The E-step

For the general finite mixture model, the component indicator variables are usually replaced by their expected values,

$$\mathbb{E}[Z_{ig} \mid \mathbf{x}_i] = \frac{\pi_g p_g(\mathbf{x} \mid \boldsymbol{\vartheta}_g)}{\sum_{h=1}^{G} \pi_h p_h(\mathbf{x} \mid \boldsymbol{\vartheta}_h)} =: \hat{z}_{ig},$$

for $i = 1, \ldots, n$ and $g = 1, \ldots, G$. Unfortunately, this usually requires the computation of both the determinant and inverse of a $p \times p$ covariance matrix. To avoid these potentially cumbersome calculations, we follow Bouveyron *et al.* (2007) and derive a cost function that utilizes the projection functions: $P_g(x)$ and $P_g^{\perp}(x)$, defined in Section 3.1. The derivation of the cost function is as follows: first, note that we can write

$$\begin{aligned} -2\log f_t(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) &= -2\log\Gamma\left[(\nu+p)/2\right] + 2\log\Gamma\left[\nu/2\right] + p\left(\log\nu + \log\pi\right) \\ &+ (\nu+p)\log\left[1 + \frac{1}{\nu}\left(||\boldsymbol{\mu} - P\left(\mathbf{x}\right)||_{\mathbf{A}}^2 + \frac{1}{b}||\mathbf{x} - P'\left(\mathbf{x}\right)||^2\right)\right] + \sum_{j=1}^{d}\log a_j + (p-d)\log b, \end{aligned}$$

where $||\mathbf{x}||_{\mathbf{A}}^2 = \mathbf{x}\mathbf{A}\mathbf{x}'$ with $\mathbf{A} = \tilde{\mathbf{D}}_g\boldsymbol{\Delta}_g\tilde{\mathbf{D}}_g'$, and all other values are as previously defined. So, on the E-step of the proposed ECM algorithm we replace each z_{ig} with

$$\hat{z}_{ig} = \frac{1}{\sum_{h=1}^{G}\exp\left[\frac{1}{2}\left(K_g\left(\mathbf{x}_i\right) - K_h\left(\mathbf{x}_i\right)\right)\right]},$$

where we refer to

$$K_g(\mathbf{x}_i) = -2\log\Gamma\left[(\nu_g + p)/2\right] + 2\log\Gamma\left[\nu_g/2\right] + (p - d_g)\log b_g + p\left(\log\nu_g + \log\pi\right)$$
$$+ (\nu_g + p)\log\left[1 + \frac{1}{\nu_g}\left(||\boldsymbol{\mu}_g - P_g(\mathbf{x})||^2_{\mathbf{A}_g} + \frac{1}{b_g}||\mathbf{x} - P'_g(\mathbf{x})||^2\right)\right] + \sum_{j=1}^{d_g}\log a_{jg} - 2\log\pi_g,$$

for $i = 1, \ldots, n$ and $g = 1, \ldots, G$, as the cost function. Each u_{ig} is then replaced by their expected values

$$\mathbb{E}[U_{ig} \mid \mathbf{x}_i, z_{ig} = 1] = \frac{\nu_g + p}{\nu_g + ||\boldsymbol{\mu}_g - P(\mathbf{x}_i)||^2_{\mathbf{A}} + \frac{1}{b_g}||\mathbf{x} - P'(\mathbf{x}_i)||^2} =: \hat{u}_{ig},$$

for $i = 1, \ldots, n$ and $g = 1, \ldots, G$ (cf. Peel and McLachlan, 2000; Andrews and McNicholas, 2012). In this update we exploit a part of the cost function derivation (given in Appendix A), which states that

$$\delta(\mathbf{x}, \boldsymbol{\mu}_g \mid \boldsymbol{\Sigma}_g) = ||\boldsymbol{\mu}_g - P(\mathbf{x}_i)||^2_{\mathbf{A}} + \frac{1}{b_g}||\mathbf{x} - P'(\mathbf{x}_i)||^2.$$

3.2.2 The CM-steps

On the first CM-step we update the mixing proportions and component location parameter using

$$\hat{\pi}_g = \frac{n_g}{n} \quad \text{and} \quad \hat{\boldsymbol{\mu}}_g = \frac{\sum_{i=1}^n \hat{z}_{ig}\hat{u}_{ig}\mathbf{x}_i}{\sum_{i=1}^n \hat{z}_{ig}\hat{u}_{ig}},$$

respectively, where $n_g = \sum_{i=1}^n \hat{z}_{ig}$. The degrees of freedom parameter, ν_g, is updated using the closed form approximation given in Andrews *et al.* (2017). Formally, we let

$$\hat{\nu}_g \approx \frac{-\exp(k) + 2\exp\left(\varphi\left(\frac{\hat{\nu}_g^{\text{old}}}{2}\right) + \frac{1-\hat{\nu}_g^{\text{old}}}{2}\right)\exp(k)}{1 - \exp(k)}$$

with

$$k = -1 - \frac{1}{n_g}\sum_{g=1}^{G}\sum_{i=1}^{n}\hat{z}_{ig}\left(\log\hat{u}_{ig} - \hat{u}_{ig}\right) - \varphi\left(\frac{\hat{\nu}_g^{\text{old}} + p}{2}\right) + \log\left(\frac{\hat{\nu}_g^{\text{old}} + p}{2}\right)$$

where $\hat{\nu}_g^{\text{old}}$ is the estimate of ν_g from the previous iteration of this ECM algorithm, and $\varphi(\cdot)$ is the digamma function.

For each tHDDC model, the updates on the second CM step are analogous to those given in Bouveyron *et al.* (2007). For illustrative purposes, we outline how to update each covariance parameter for the UUUUU model, i.e., the model where $a_{jg}, b_g, \mathbf{D}_g, d_g, \nu_g$ are free to vary across all $g = 1, \ldots, G$ and $j = 1, \ldots, d_g$.

First, we calculate the intrinsic dimension, d_g. For each value of $j \in \{1, p-1\}$ we compute

$$l\left(\hat{\boldsymbol{\vartheta}}\right) = -\frac{n}{2}\left(d_j \log a_{jg} + (p - d_j)\log b_g - 2\log\pi_g + \log\nu_g \right.$$
$$\left. + \log\pi - 2\log\Gamma\left[\left(\nu_g + p\right)/2\right] + 2\log\Gamma\left(\nu_g/2\right)\right) \quad (3.6)$$

and set d_g equal to the value of d_j that maximizes the BIC values associated with the log-likelihood values found using (3.6). Then we let $\mathbf{D}_g$ be the eigenvectors of $\hat{\boldsymbol{\Sigma}}_g$, where

$$\hat{\boldsymbol{\Sigma}}_g = \frac{1}{n_g}\sum_{i=1}^{n}\hat{z}_{ig}\hat{u}_{ig}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)'. \quad (3.7)$$

Each a_{jg}, for $j = 1, \ldots, d_g$ and $g = 1, \ldots, G$ is then replaced with the first d_g eigenvalues of $\hat{\boldsymbol{\Sigma}}_g$ and we estimate b_g using

$$\hat{b}_g = \frac{1}{(p - d_g)}\left(\text{tr}\left(\hat{\boldsymbol{\Sigma}}_g\right) - \sum_{j=1}^{d_g} a_{jg}\right). \quad (3.8)$$

3.3 Convergence

For the proposed ECM algorithm, we initialize each model using either k-means clustering or random starting values and use the Aitkin's acceleration (Aitken, 1926) procedure to determine if the algorithm has converged. That is, we consider this ECM algorithm to have converged when $l_\infty^{(k+1)} - l^{(k)} < \epsilon$, where $\epsilon = 10^{-2}$ (see Lindsay, 1995; McNicholas *et al.*, 2010). In this criterion, $l^{(k)}$ is the log-likelihood value at iteration (k) and $l_\infty^{(k+1)}$ is the asymptotic estimate of the log-likelihood at iteration $(k+1)$. Formally,

$$l_\infty^{(k+1)} = l^{(k)} + \frac{1}{1 - a^{(k)}} \left(l^{(k+1)} - l^{(k)} \right), \tag{3.9}$$

where

$$a^{(k)} = \frac{l^{(k+1)} - l^{(k)}}{l^{(k)} - l^{(k-1)}}.$$

Table 3.1: Nomenclature, covariance decomposition and number of free covariance parameters for the tHDDC models. For constaints on a_{jg}, U represents unconstrained, D represents constrained accross dimension, G represents constrained across groups and C represents constrained accross both dimension and group. For all other components, U and C are unconstrained and constrained across groups, respectively. For the number of free parameters, $\rho = Gp+G+1$ is the number of parameters required to estimate the mean and proportions. The number of parameters required to estimate $\tilde{D}_g$, $\tilde{D}$ and are $\bar{\tau} = d_g[p-(d_g+1)/2]$ and $\tau = d[p-(d+1)/2]$, and $s = \sum_{g=1}^{G} d_g$.

Model	$a_{jg} = a_g/a_j$	$b_g = b$	$\mathbf{D}_g = \mathbf{D}$	$d_g = d$	$\nu_g = \nu$	Number of Cov. Parameters
UUUUU	U	U	U	U	U	$\rho + \bar{\tau} + 3G + s$
UCUUU	U	C	U	U	U	$\rho + \bar{\tau} + 2G + s + 1$
DUUUU	D	U	U	U	U	$\rho + \bar{\tau} + 4G$
CUUUU	C	U	U	U	U	$\rho + \bar{\tau} + 3G + 1$
DCUUU	D	C	U	U	U	$\rho + \bar{\tau} + 3G + 1$
CCUUU	C	C	U	U	U	$\rho + \bar{\tau} + 2G + 2$
UUUCU	U	U	U	C	U	$\rho + G(\tau + d + 2) + 1$
UCUCU	U	C	U	C	U	$\rho + G(\tau + d + 1) + 2$
DUUCU	D	U	U	C	U	$\rho + G(\tau + 2 + 1) + 1$
CUUCU	C	U	U	C	U	$\rho + G(\tau + 2) + 2$
DCUCU	D	C	U	C	U	$\rho + G(\tau + 2) + 2$
CCUCU	C	C	U	C	U	$\rho + G(\tau + 1) + 3$
GCCCU	G	C	C	C	U	$\rho + \tau + d + G + 2$
CCCCU	C	C	C	C	U	$\rho + \tau + G + 3$
UUUUC	U	U	U	U	C	$\rho + \bar{\tau} + 2G + s + 1$
UCUUC	U	C	U	U	C	$\rho + \bar{\tau} + G + s + 2$
DUUUC	D	U	U	U	C	$\rho + \bar{\tau} + 3G + 1$
CUUUC	C	U	U	U	C	$\rho + \bar{\tau} + 2G + 2$
DCUUC	D	C	U	U	C	$\rho + \bar{\tau} + 2G + 2$
CCUUC	C	C	U	U	C	$\rho + \bar{\tau} + G + 3$
UUUCC	U	U	U	C	C	$\rho + G(\tau + d + 1) + 2$
UCUCC	U	C	U	C	C	$\rho + G(\tau + d) + 3$
DUUCC	D	U	U	C	C	$\rho + G(\tau + 2) + 2$
CUUCC	C	U	U	C	C	$\rho + G(\tau + 1) + 3$
DCUCC	D	C	U	C	C	$\rho + G(\tau + 1) + 3$
CCUCC	C	C	U	C	C	$\rho + G\tau + 4$
GCCCC	G	C	C	C	C	$\rho + \tau + d + 3$
CCCCC	C	C	C	C	C	$\rho + \tau + 4$

Chapter 4

Application

4.1 Computational Considerations

The data analyses will be treated as genuine clustering problems, where the true classifications are not known. Because we do have the true class labels, the ARI will be used to assess class agreement between the true and predicted class labels. In all applications, the best fitting models will be chosen using the BIC.

All analyses are performed in R version 3.3.2 (R Core Team, 2016) for Linux 6.5[1]. Both the HDDC and tHDDC models are fit with $G = 1, \ldots, 10$. It is important to note that the only HDDC models considered are ones with a monotonic likelihood.

[1]Using a 32-core Intel Xeon E5 server with 256GB RAM running 64-bit CentOS

4.2 Simulation Studies

We use a simulation study to highlight the aforementioned drawback of the considered mixture of multivariate Gaussian distributions. Ten data sets were simulated from a two-component multivariate-t distribution with $\nu_1 = 2$ and $\nu_2 = 3$. Figure 4.1 provides an illustration of the first three dimensions of one of the simulated data sets. In each component, observations are scattered from the mean, with many outliers on far ends of the clusters. Table 4.1 gives the classification results for the tHDDC and HDDC models when fitted to the simulated data set. As expected, the tHDDC approach outperforms the HDDC approach, achieving a near perfect classification. The relatively small standard deviation reveals that the selected tHDDC models are

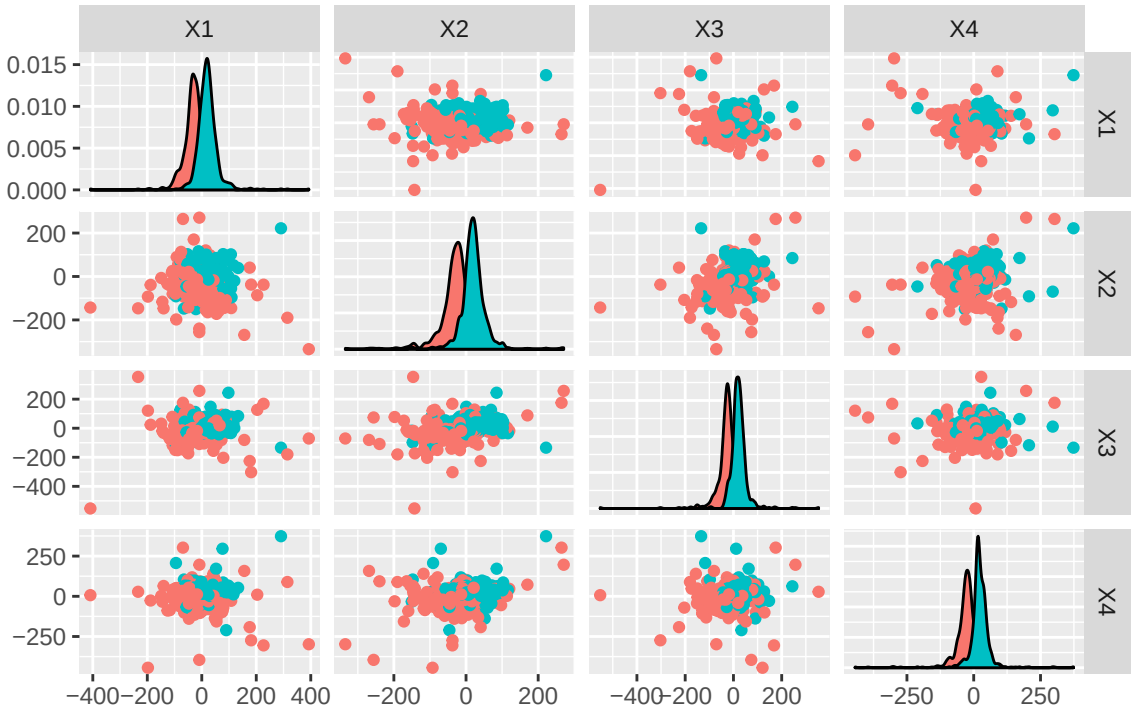

Figure 4.1: Pairs and density plots of the first three dimensions of one multivariate-t simulated data set, coloured by true groups.

consistently returning an exceptional classification performance, whereas the selected HDDC models are using extra components to account for the increased variation in the simulated data sets.

Table 4.1: Mean and standard deviation of ARI values returned by the best fitting HDDC and tHDDC models found by the BIC for the simulated multivariate-t data sets.

	Mean of ARI	Standard Deviation of ARI
HDDC	0.021	0.012
tHDDC	0.995	0.005

4.3 Fisher's Irises

In our first real data analysis, we consider Fisher's famous iris data set, which is available in the R package `datasets` and `gclus` (Hurley, 2012). It is composed of four explanatory variables: sepal length, sepal width, petal length, and petal width, measured in centimetres. There are three species of the plant: Setosa, Versicolour and Virginica. Table 4.2 gives the classification results.

Both the best fitting HDDC and tHDDC models return a very good classification of the irises, with the best fitting tHDDC model outperforming the corresponding HDDC model. Across the three selected components, both the best fitting HDDC and tHDDC models use a varying number of eigenvalues and eigenvectors with a

Table 4.2: Model decomposition, number of components, BIC and ARI values for the best fitting tHDDC and HDDC models when fitted for $G = 1, \ldots, 10$ components for the Fisher's irises.

	Model	G	BIC	ARI
HDDC	UUUC	3	-588.01	0.868
tHDDC	UUUCC	3	-646.33	0.904

common intrinsic dimension. Specifically, the selected tHDDC models uses a common dimension of one and constrained degrees of freedom. In total, this model misclassifies only 5 irises (see Table 4.3).

Figure 4.2 shows the data in the one dimensional subspace for all three groups. The clusters found by the selected tHDDC model are clear and well separated.

Table 4.3: A classification table showing the results for the selected three-component UUUCC tHDDC model for the iris data.

	A	B	C
Setosa	50	0	0
Versicolor	0	45	5
Virginica	0	0	50

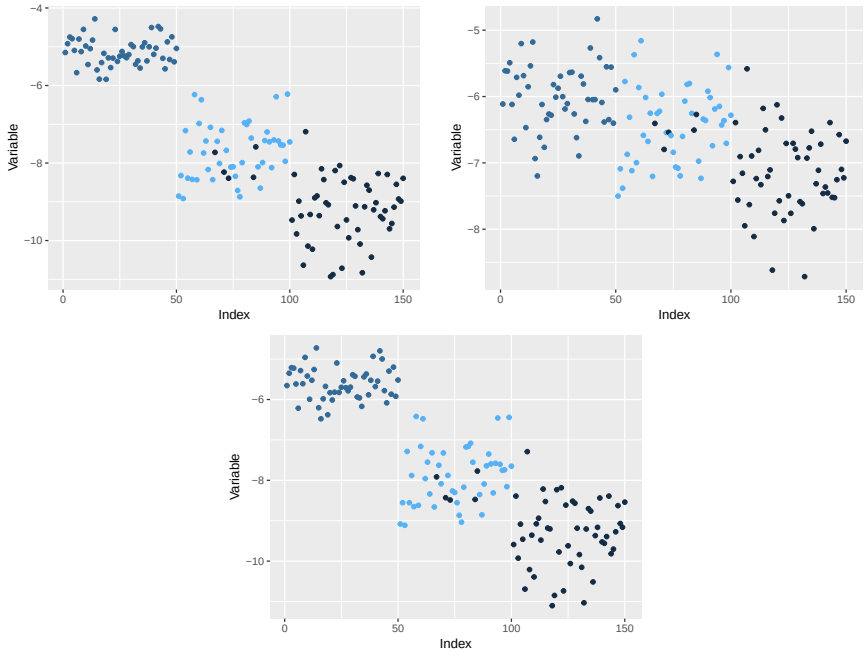

Figure 4.2: The iris data coloured by the classes found by the best fitting tHDDC model and projected onto the one-dimensional spaces for each component.

4.4 Italian Wines

The Italian wines data set, available as `wine` in the R package `pgmm` (McNicholas *et al.*, 2015), is composed of 178 Italian wines on which 27 measurements are taken (Forina *et al.*, 1986). The wines come from three different cultivars and are classified based on which one they come from: Barolo, Grignolino and Barbera. The *t*HDDC models are fit using $G = 1, \ldots, 5$, as the BIC will select a four-component mixture model. Table 4.4 gives the classification results. Note that when fitting only three-component HDDC and tHDDC models, the selected GCCC and GCCCC models gave the same classification result (ARI = 0.933).

We can see that the best fitting HDDC model overfits by selecting a model with eight components. Although the best *t*HDDC model did not have three groups, the four group solution gives a superior classification performance (see Table 4.4). The model selected by the BIC was the GCCCC model, with a common intrinsic dimension of 4. Looking at Figure 4.3, in the four dimensional subspace, we can see that there are no clear and well separated clusters.

Table 4.4: Model decomposition, number of components, BIC and ARI values for the best fitting HDDC and *t*HDDC models when fitted for $G = 1, \ldots, 10$ components for the Italian wines data.

	Model	G	BIC	ARI
HDDC	GCCC	8	$-12,071.63$	0.658
*t*HDDC	GCCCC	4	$-11,965.26$	0.758

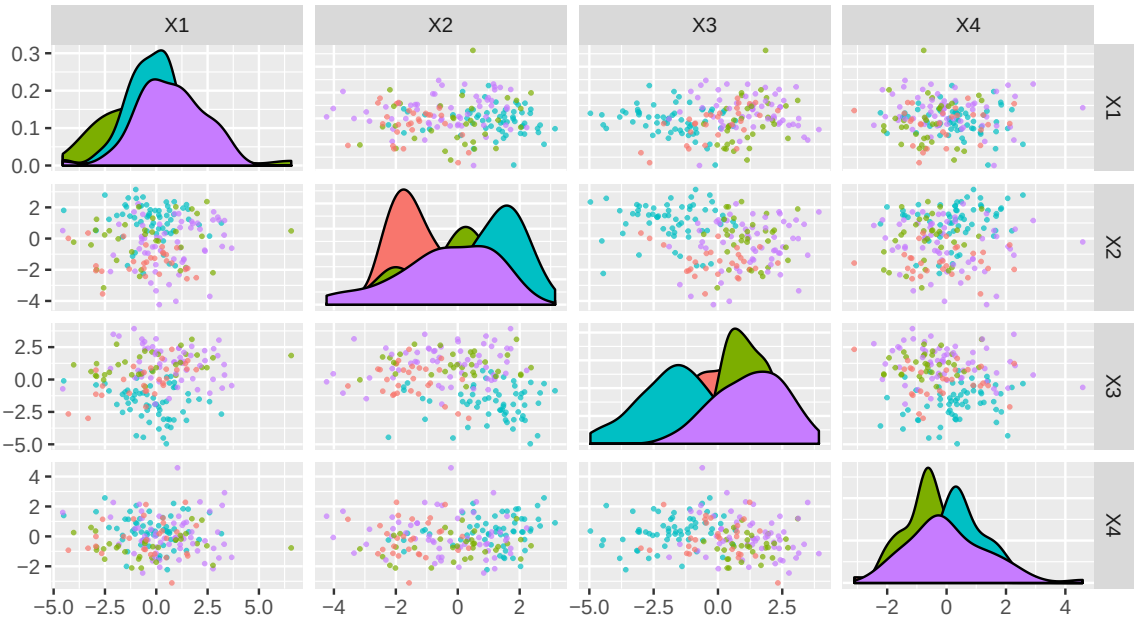

Figure 4.3: Pairs and density plots of the Italian wines data colored by classes found by the best fitting tHDDC model and projected onto the four-dimensional subspace.

4.5 Martian Surface

This data set was retrieved by the OMEGA instrument (Mars Express, ESA; Bibring *et al.*, 2004). The OMEGA instrument is used for characterization of the Martian surface based on physical and chemical composition. This can include classes of silicates, hydrated minerals, ices and more. The data used is based on one 300×128 raw image. It contains 255 variables on 38,400 observations. With a physical model, eight groups were found and for the purpose of this analysis, these will be treated as true groups; however, the best determination of model performance here is based on efficacy for image reconstruction. For $G = 8$ components, tHDDC does a little better than HDDC; however, neither performs well (Table 4.5).

Although the physical model suggests eight groups, experts in the field are interested in exploring a five group solution (Bouveyron *et al.*, 2007). Both HDDC and *t*HDDC models are applied to this data with $G = 5$, and the selected *t*HDDC model recovers the clusters better than the selected HDDC model (see Table 4.6).

In Table 4.7, we can see that the classification results returned by the selected *t*HDDC and HDDC models are quite different. Furthermore, comparing the recovered image based on the predicted classes to the original image (see Figure 4.4), the utility of the model becomes clear, i.e., the physical details are generally recovered very well.

Table 4.5: Model decomposition, BIC and ARI values for the selected eight-component HDDC and *t*HDDC models for the Martian surface data.

	G	Model	BIC	ARI
HDDC	8	CUUU	62,460,591	0.319
*t*HDDC	8	CCCCC	64,249,591	0.351

Table 4.6: Model decomposition, BIC and ARI values for the selected five-component HDDC and *t*HDDC models for the Martian surface data.

	G	Best Model	BIC	ARI
HDDC	5	UUUU	61,956,344	0.472
*t*HDDC	5	UCUUC	70,120,085	0.645

Table 4.7: A classification table comparing the best fitting five-component *t*HDDC and HDDC models for the Martian surface data.

		HDDC				
		A	B	C	D	E
*t*HDDC	1	10744	22	0	0	2973
	2	1019	1598	0	99	785
	3	0	39	7807	4372	4
	4	2	319	1	4871	716
	5	605	914	4	283	1223

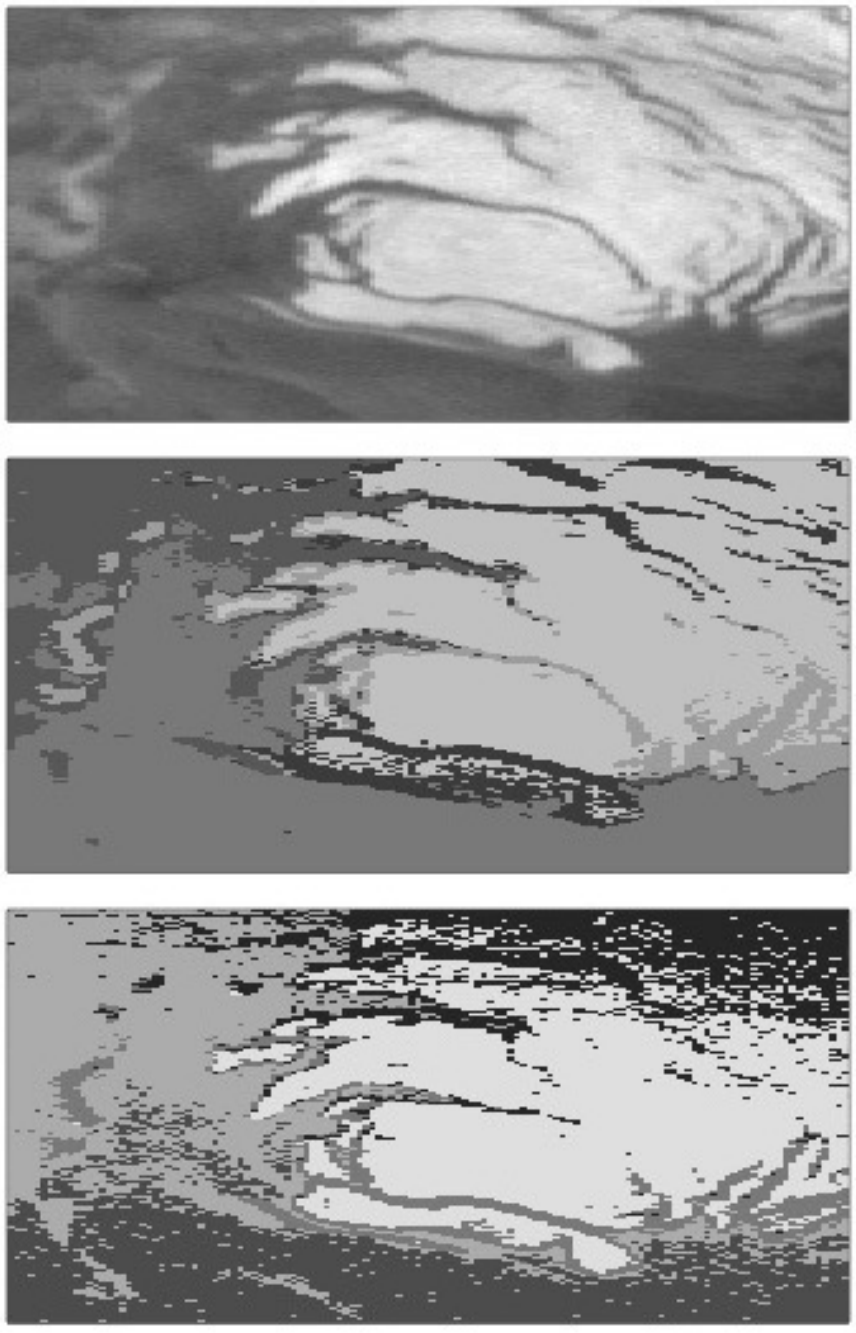

Figure 4.4: Three images of the Martian surface constructed using spectral data collected by the OMEGA instrument (top), the parameter estimates of the best-fitting tHDDC model (middle) and the parameter estimates of the best-fitting HDDC model (bottom).

Chapter 5

Conclusions and Future Work

A new family of multivariate-t mixture models has been proposed. The tHDDC approach is an extension of the HDDC approach that incorporates the multivariate-t distribution, allowing for a more robust clustering scheme. A total of 28 models have been developed and the need for these models was shown through a simulation study which demonstrated their flexibility in recognizing outliers. The models were tested on both simulated and real data sets and show superior results when compared to the HDDC family. In particular, the results on the high-dimensional Martian surface data show that image recovery can be greatly improved. Overall, the added degrees of freedom parameter allows for more flexible clusters and a more flexible modelling structure than HDDC.

A further direction to this research would be to explore different model and dimension selection criteria. Some include the integrated completed likelihood criterion (Biernacki *et al.*, 2000) and the approximate weight of evidence (Banfield and Raftery, 1993). In addition, this method can be extended to include skewed mixture-models. Examples include the mixture of multivariate skew-t distributions (Lin, 2010; Murray

et al., 2014, 2017), the mixture of shifted asymmetric Laplace distributions (Franczak *et al.*, 2014), the mixture of variance-gamma distributions (McNicholas *et al.*, 2017), and the mixture of generalized hyperbolic distributions (Browne and McNicholas, 2015).

Appendix A

Cost Function Derivations

The cost function for a mixture of multivariate-t distributions can be derived as follows:

$$\log \xi_p\left(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu\right) = \log\left[\frac{\Gamma\left[(\nu+p)/2\right]|\boldsymbol{\Sigma}|^{-1/2}(\nu\pi)^{-p/2}}{\Gamma\left[\nu/2\right]\left[1+\delta\left(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma}\right)/\nu\right]^{(\nu+p)/2}}\right],$$

$$\log \xi_p\left(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu\right) = \log\Gamma\left[(\nu+p)/2\right] - \log\left[|\boldsymbol{\Sigma}|^{1/2}\right] - \log\left[(\nu\pi)^{p/2}\right] - \log\Gamma\left[\nu/2\right] - \log\left[\left(1+\frac{1}{\nu}\delta\left(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma}\right)\right)^{(\nu+p)/2}\right],$$

$$\log \xi_p\left(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu\right) = \log\Gamma\left[(\nu+p)/2\right] - \frac{p}{2}\log\left[\nu\pi\right] - \log\Gamma\left[\nu/2\right] - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{(\nu+p)}{2}\log\left[1+\frac{1}{\nu}\delta\left(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma}\right)\right],$$

$$-2\log \xi_p\left(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu\right) = -2\log\Gamma\left[(\nu+p)/2\right] + 2\log\Gamma\left[\nu/2\right] + p\left(\log\nu + \log\pi\right) + \log\left|(\mathbf{D}\boldsymbol{\Delta}\mathbf{D}')\right| + (\nu+p)\log\left[1+\frac{1}{\nu}\left(\mathbf{x}-\boldsymbol{\mu}\right)'\left(\mathbf{D}\boldsymbol{\Delta}\mathbf{D}'\right)^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right],$$

$$-2\log\xi_p\left(\mathbf{x}\mid\boldsymbol{\mu},\boldsymbol{\Sigma},\nu\right)=-2\log\Gamma\left[(\nu+p)/2\right]+2\log\Gamma\left[\nu/2\right]+p\left(\log\nu+\log\pi\right)+\log|\boldsymbol{\Delta}|$$
$$+(\nu+p)\log\left[1+\frac{1}{\nu}\left(||\tilde{\mathbf{D}}\tilde{\mathbf{D}}'\left(\mathbf{x}-\boldsymbol{\mu}\right)||_{\mathbf{A}}^2+\frac{1}{b}||\bar{\mathbf{D}}\bar{\mathbf{D}}'\left(\mathbf{x}-\boldsymbol{\mu}\right)||^2\right)\right],$$

$$\begin{aligned}-2\log\xi_p\left(\mathbf{x}\mid\boldsymbol{\mu},\boldsymbol{\Sigma},\nu\right)=&-2\log\Gamma\left[(\nu+p)/2\right]+2\log\Gamma\left[\nu/2\right]+p\left(\log\nu+\log\pi\right)\\&+(\nu+p)\log\left[1+\frac{1}{\nu}\left(||\boldsymbol{\mu}-P\left(\mathbf{x}\right)||_{\mathbf{A}}^2+\frac{1}{b}||\mathbf{x}-P'\left(\mathbf{x}\right)||^2\right)\right]\\&+\sum_{j=1}^{d}\log a_{jg}+(p-d)\log b.\end{aligned}$$

Therefore, we can write

$$z_{ig}=\frac{1}{\sum_{h=1}^{G}\exp\left\{\frac{1}{2}\left(K_g\left(\mathbf{x}_i\right)-K_h\left(\mathbf{x}_i\right)\right)\right\}},$$

where

$$\begin{aligned}K_g\left(\mathbf{x}\right)=&-2\log\Gamma\left[(\nu_g+p)/2\right]+2\log\Gamma\left[\nu_g/2\right]+p\left(\log\nu_g+\log\pi\right)-2\log\pi_g\\&+(\nu_g+p)\log\left[1+\frac{1}{\nu_g}\left(||\boldsymbol{\mu}_g-P_g\left(\mathbf{x}\right)||_{\mathbf{A}_g}^2+\frac{1}{b_g}||\mathbf{x}-P_g'\left(\mathbf{x}\right)||^2\right)\right]\\&+\sum_{j=1}^{d_g}\log a_{jg}+(p-d_g)\log b_g.\end{aligned}$$

Bibliography

Aitken, A. C. (1926). On Bernoulli's numerical solution of algebraic equations. *Proceedings of the Royal Society of Edinburgh*, **46**, 289–305.

Andrews, J. L. and McNicholas, P. D. (2011a). Extending mixtures of multivariate t-factor analyzers. *Statistics and Computing*, **21**(3), 361–373.

Andrews, J. L. and McNicholas, P. D. (2011b). Mixtures of modified t-factor analyzers for model-based clustering, classification, and discriminant analysis. *Journal of Statistical Planning and Inference*, **141**(4), 1479–1486.

Andrews, J. L. and McNicholas, P. D. (2012). Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions. *Statistics and Computing*, **22**(5), 1021–1029.

Andrews, J. L. and McNicholas, P. D. (2014). teigen: Model-based clustering and classification with the multivariate t-distribution. *R package version*, **1**, 24–46.

Andrews, J. L., McNicholas, P. D., and Chalifour, M. (2015). *mmtfa: Model-Based Clustering and Classification with Mixtures of Modified t Factor Analyzers*. R package version 0.1.

Andrews, J. L., Wickins, J. R., Boers, N. M., and McNicholas, P. D. (2017). teigen: An R package for model-based clustering and classification via the multivariate t distribution. *Journal of Statistical Software*. To appear.

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

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49**(3), 803–821.

Barndorff-Nielsen, O., Kent, J., and Sørensen, M. (1982). Normal variance-mean mixtures and z distributions. *International Statistical Review / Revue Internationale de Statistique*, **50**(2), pp. 145–159.

Berge, L., Bouveyron, C., and Girard, S. (2012). HDclassif: An R package for model-based clustering and discriminant analysis of high-dimensional data. *Journal of Statistical Software*, **46**(6), 1–29.

Bibring, J.-P., Soufflot, A., Berthé, M., Langevin, Y., Gondet, B., Drossart, P., Bouyé, M., Combes, M., Puget, P., Semery, A., *et al.* (2004). Omega: Observatoire pour la minéralogie, l'eau, les glaces et l'activité. In *Mars Express: The Scientific Payload*, volume 1240, pages 37–49.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, **22**(7), 719–725.

Bouveyron, C. and Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: A review. *Computational Statistics and Data Analysis*, **71**, 52–78.

Bouveyron, C., Girard, S., and Schmid, C. (2007). High-dimensional data clustering. *Computational Statistics & Data Analysis*, **52**(1), 502–519.

Browne, R. P. and McNicholas, P. D. (2015). A mixture of generalized hyperbolic distributions. *Canadian Journal of Statistics*, **43**(2), 176–198.

Browne, R. P., ElSherbiny, A., and McNicholas, P. D. (2015). *mixture: Mixture Models for Clustering and Classification*. R package version 1.4.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research*, **1**(2), 245–276.

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

Dasgupta, A. and Raftery, A. E. (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association*, **93**, 294–302.

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

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

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, **97**(458), 611–631.

Franczak, B. C., Browne, R. P., and McNicholas, P. D. (2014). Mixtures of shifted asymmetriclaplace distributions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **36**(6), 1149–1157.

Ghahramani, Z. and Hinton, G. E. (1997). The EM algorithm for mixtures of factor analyzers. Technical report, Technical Report CRG-TR-96-1, University of Toronto.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, **2**(1), 193–218.

Hurley, C. (2012). *gclus: Clustering Graphics*. R package version 1.3.1.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**(430), 773–795.

Kass, R. E. and Wasserman, L. (1995). A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the American Statistical Association*, **90**(431), 928–934.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 49–66.

Kotz, S. and Nadarajah, S. (2004). *Multivariate t-distributions and their applications*. Cambridge University Press.

Leroux, B. G. *et al.* (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics*, **20**(3), 1350–1360.

Lin, T.-I. (2010). Robust mixture modeling using multivariate skew t distributions. *Statistics and Computing*, **20**(3), 343–356.

Lin, T.-I., McNicholas, P. D., and Hsiu, J. H. (2014). Capturing patterns via parsimonious t mixture models. *Statistics and Probability Letters*, **88**, 80–87.

Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry, and Applications*. Institute for Mathematical Statistics: Hayward, CA.

McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. Wiley, New York, 2nd edition.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. John Wiley and Sons, New York.

McLachlan, G. J., Bean, R., and Jones, L. B.-T. (2007). Extension of the mixture of factor analyzers model to incorporate the multivariate t-distribution. *Computational Statistics & Data Analysis*, **51**(11), 5327–5338.

McNicholas, P. D. (2016a). *Mixture Model-Based Classification*. Chapman and Hall/CRC Press, Boca Raton.

McNicholas, P. D. (2016b). Model-based clustering. *Journal of Classification*, **33**(3), 331–373.

McNicholas, P. D. and Murphy, T. B. (2008). Parsimonious Gaussian mixture models. *Statistics and Computing*, **18**(3), 285–296.

McNicholas, P. D. and Murphy, T. B. (2010). Model-based clustering of microarray expression data via latent Gaussian mixture models. *Bioinformatics*, **26**(21), 2705–2712.

McNicholas, P. D., Murphy, T. B., McDaid, A. F., and Frost, D. (2010). Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. *Computational Statistics and Data Analysis*, **54**(3), 711–723.

McNicholas, P. D., ElSherbiny, A., McDaid, A. F., and Murphy, T. B. (2015). *pgmm: Parsimonious Gaussian Mixture Models*. R package version 1.2.

McNicholas, S. M., McNicholas, P. D., and Browne, R. P. (2017). A mixture of variance-gamma factor analyzers. In S. E. Ahmed, editor, *Big and Complex Data Analysis: Methodologies and Applications*, pages 369–385. Springer International Publishing, Cham.

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, **80**(2), 267–278.

Murray, P. M., Browne, R. P., and McNicholas, P. D. (2014). Mixtures of skew-t factor analyzers. *Computational Statistics and Data Analysis*, **77**(0), 326–335.

Murray, P. M., Browne, R. P., and McNicholas, P. D. (2017). A mixture of SDB skew-t factor analyzers. *Econometrics and Statistics*, **3**, 160–168.

Peel, D. and McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing*, **10**(4), 339–348.

Puts, M., Daas, P., and de Waal, T. (2015). Finding errors in big data. *Significance*, **12**(3), 26–29.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**, 846–850.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**(2), 461–464.

Scott, D. W. and Thompson, J. R. (1983). Probability density estimation in higher dimensions. In *Computer Science and Statistics: Proceedings of the fifteenth symposium on the interface*, volume 528, pages 173–179. North-Holland, Amsterdam.

Steane, M. A., McNicholas, P. D., and Yada, R. (2012). Model-based classification via mixtures of multivariate t-factor analyzers. *Communications in Statistics – Simulation and Computation*, **41**(4), 510–523.

Vrbik, I. and Mcnicholas, P. D. (2014). Parsimonious skew mixture models for model-based clustering and classification. *Computational Statistics & Data Analysis*, **71**, 196–210.

Wolfe, J. H. (1963). *Object Cluster Analysis of Social Areas*. Master's thesis, University of California, Berkeley.