Dimension Reduction and Clustering of High Dimensional Data using a Mixture of Generalized Hyperbolic Distributions

DIMENSION REDUCTION AND CLUSTERING OF HIGH DIMENSIONAL DATA USING A MIXTURE OF GENERALIZED HYPERBOLIC DISTRIBUTIONS

BY

THINESH PATHMANATHAN, 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 (2018) McMaster University

(Mathematics & Statistics) Hamilton, Ontario, Canada

TITLE:	Dimension Reduction and Clustering of High Dimensional Data using a Mixture of Generalized Hyperbolic Distributions
AUTHOR:	Thinesh Pathmanathan B.Sc., (Mathematics and Statistics) University of Toronto, Toronto, Canada
SUPERVISOR:	Dr. Sharon McNicholas
NUMBER OF PAGES:	viii, 35

To my family and friends for their endless support throughout this endeavor.

Abstract

Model-based clustering is a probabilistic approach that views each cluster as a component in an appropriate mixture model. The Gaussian mixture model is one of the most widely used model-based methods. However, this model tends to perform poorly when clustering high-dimensional data due to the over-parametrized solutions that arise in high-dimensional spaces. This work instead considers the approach of combining dimension reduction techniques with clustering via a mixture of generalized hyperbolic distributions. The dimension reduction techniques, principal component analysis and factor analysis along with their extensions were reviewed. Then the aforementioned dimension reduction techniques were individually paired with the mixture of generalized hyperbolic distributions in order to demonstrate the clustering performance achieved under each method using both simulated and real data sets. For a majority of the data sets, the clustering method utilizing principal component analysis exhibited better classification results compared to the clustering method based on the extending the factor analysis model.

Acknowledgements

First and foremost, I have to thank my research supervisor, Dr. Sharon McNicholas, without her assistance and dedicated involvement in every step throughout the process, this paper would have never been accomplished. I would like to thank you very much for your support and understanding over the past year.

I would also like to thank Dr. Roman Viveros-Aguilera and Dr. Paul McNicholas for taking the time to be to serve as members of my examining committee. Additionally, I would like to thank Dr. Paul McNicholas for providing most of the R code that I used for this analysis.

Contents

Abstract **iv**

Acknowledgements **v**

1 Introduction **1**

2 Content **4**

2.1 Mixture Models 4

2.2 Mixtures of Multivariate Gaussian distributions 5

2.3 Generalized Inverse Gaussian Distributions 5

2.4 Mixtures of Generalized Hyperbolic Distributions 7

2.5 Principal Component Analysis 9

2.6 Mixtures of Generalized Hyperbolic Factor Analyzers 10

2.6.1 Factor Analysis 10

2.6.2 Mixtures of Factor Analyzers 11

3 Methodology **12**

3.1 Likelihood 12

3.2 Parameter Estimation 13

3.2.1 EM Algorithm 13

3.2.2 AECM Algorithm 14

3.2.3 Method Initialization 14

3.2.4 Convergence 15

3.3 Predicted Classifications 16

3.4 Model Selection 16

3.5 Performance Assessment 17

4 Application **18**

4.1 Simulated Studies 19

4.1.1 Twonorm Data 19

4.1.2 Ringworm Data 20

4.1.3 Waveform Data 22

4.2 Real Data 23

4.2.1 Wine Data 23

4.2.2 Satellite data 24

5 Conclusions **26**

A Bessel Functions **28**

Bibliography **30**

List of Figures

2.1 Densities of GIG distributions with various parameterizations. 6

4.2 Projection of the clustered twonorm data in the 2 first principal components of PCA. 20

4.3 Projection of the clustered ringnorm data in the 2 first principal components of PCA. 21

4.4 Projection of the clustered waveform data in the 2 first principal components of PCA. 23

Chapter 1

Introduction

Classification is a procedure in which group membership labels are assigned to unlabeled observations. A group can be a class or a cluster. Having prior knowledge of observation labels and the degree in which this information is used, can lead to the following three types of classification: supervised, semi-supervised, and unsupervised (McNicholas, 2016). Cluster analysis is unsupervised classification, where the labels for all observations are missing or are treated as such.

A typical way to define a cluster is as a group of observations such that the observations in a particular group are more similar to one another than they are to the observations present in any alternative groups. Using such a definition is problematic because taken at the extreme each observation is defined as its own cluster (McNicholas, 2016). Instead it is more prudent if one thinks of a cluster as a component in a suitable mixture model (Tiedeman, 1955; Wolfe, 1963). The use of a mixture model or a family of mixture models for clustering is known as model-based clustering (McNicholas, 2016).

The onset of the information age has resulted in large improvements to the amount of

data that can be stored and a rapid increase in the rate at which this data can be generated. Both these developments have resulted in the term "Big Data" being coined. Big Data sources are typically associated with high-dimensional variables of mixed type which are produced in rapid succession (Daas *et al.*, 2015). The focus herein will be on model-based clustering of high-dimensional data. Clustering high-dimensional data using traditional model-based approaches has proven to be difficult due to a phenomenon known in literature as the *curse of dimensionality* (Bellman, 1957). This "curse" refers to the over-parametrized solutions that are returned from estimating parameters over a high-dimensional space (Bouveyron and Brunet-Saumard, 2014). The works by (Scott and Thompson, 1983; Huber, 1985) have demonstrated it is more pragmatic to implement clustering techniques in a lower dimensional space as opposed to the original space. Consequently, an appropriate measure to address the issue of dimensionality is considering dimension reduction techniques.

This thesis is a continuation of the earlier work begun by Bouveyron and Brunet-Saumard (2014), who used Gaussian mixture approaches to cluster high-dimensional data. This work further develops their contributions by pairing dimension reduction techniques alongside clustering using the mixture of hyperbolic distributions (MGHD). In particular, the approach of using principal component analysis for dimension reduction in addition to clustering using MGHD will be compared against the mixture of generalized hyperbolic factor analyzers (MGHFA) model. The mixture of generalized factor analyzers (MFA) model simultaneously clusters the data and reduces the dimensionality within each cluster locally. The MGHFA model is an extension of the MFA model which incorporates the generalized hyperbolic distribution. Model-based clustering will be carried out using the MGHD model in favor of

the conventional Gaussian mixture model because it is able to produce more favorable results when the clusters are asymmetrical or skewed. Furthermore, the MGHD exhibits a great deal of modelling flexibility as a result of its index parameter (see Chapter 2 for details).

The remainder of this thesis will be organized as follows: in Chapter 2, the theory behind mixture models, relevant distributions, and popular dimension reduction techniques will be discussed. In Chapter 3, parameter estimation, model selection, and classification assessment will be overviewed. In Chapter 4, R software (R Core Team, 2017) will be used to demonstrate the classification performance achieved when using dimension reduction approaches in conjunction with model-based clustering methods. Lastly, Chapters 5, will end with a general discussion on the highlighted methods and suggestions for future work.

Chapter 2

Content

2.1 Mixture Models

A finite mixture model is a convex linear combination of two or more probability density functions. Let $\mathbf{X}$ be a p-dimensional random vector. The probability density function of a mixture model is given by

$$f(\mathbf{x}|\boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{x}|\boldsymbol{\theta}_g), \tag{2.1}$$

where $f_g(\mathbf{x}|\boldsymbol{\theta}_g)$ is the density function for the gth component, π_g are the mixing proportions, and $\boldsymbol{\vartheta} = (\pi_1, \pi_2, ..., \pi_G, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_G)$ is the vector of parameters. The following assumptions are made: $\pi_g > 0$ for $g \in \{1, 2, ..., G\}$, and $\sum_{g=1}^{G} \pi_g = 1$. By convention, the component densities $f_1(\mathbf{x}|\boldsymbol{\theta}_1), f_2(\mathbf{x}|\boldsymbol{\theta}_2), ..., f_G(\mathbf{x}|\boldsymbol{\theta}_G)$ come from the same family of distributions and $f(\mathbf{x}|\boldsymbol{\vartheta})$ is known as the G-component finite mixture density (McNicholas, 2016).

2.2 Mixtures of Multivariate Gaussian distributions

Gaussian mixture models (GMMs) are widely used in clustering applications because data following a mixture of multivariate normal densities have spherical clusters centered around their means with higher density for points closer to the mean (Fraley and Raftery, 2002). The density function for a GMM can be written as follows:

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

where

$$\phi(\mathbf{x}|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) = \frac{1}{\sqrt{(2\pi)^p |\boldsymbol{\Sigma}_g|}} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_g)' {\boldsymbol{\Sigma}_g}^{-1} (\mathbf{x} - \boldsymbol{\mu}_g) \right\} \tag{2.3}$$

is the probability density function of a multivariate Gaussian distribution with mean $\boldsymbol{\mu}_g$ and variance-covariance matrix $\boldsymbol{\Sigma}_g$. Parameter estimation for the GMM model is carried out using the expectation-maximization algorithm. GMMs perform poorly when the data exhibits departures from the assumptions of normality such as skewness or outliers (Juárez and Steel, 2010).

2.3 Generalized Inverse Gaussian Distributions

The generalized inverse Gaussian (GIG) distribution was first formalized by Good (1953) but its development can be traced back to Halphen (1941). At the time of its development, the GIG was known as the Halphen Type A distribution. Barndorff-Nielsen and Halgreen (1977) outlined the statistical properties that resulted in the

GIG distribution that is used today. Consider a random variable W following a GIG distribution with parameters $a, b \in \mathbb{R}^+$, $\lambda \in \mathbb{R}$, then its probability density function for $w > 0$ can be written as:

$$q(w|a,b,\lambda) = \frac{(a/b)^{\lambda/2} w^{\lambda-1}}{2\eta K_\lambda \sqrt{ab}} \exp\left\{ -\frac{aw + b/w}{2} \right\}, \tag{2.4}$$

where K_λ is the modified Bessel function of the third kind (see Appendix A for details) with index parameter λ, and the parameters a and b control concentration via $\sqrt{ab}$ and scaling via $\sqrt{a/b}$ (Koudou *et al.*, 2014; Browne and McNicholas, 2015).

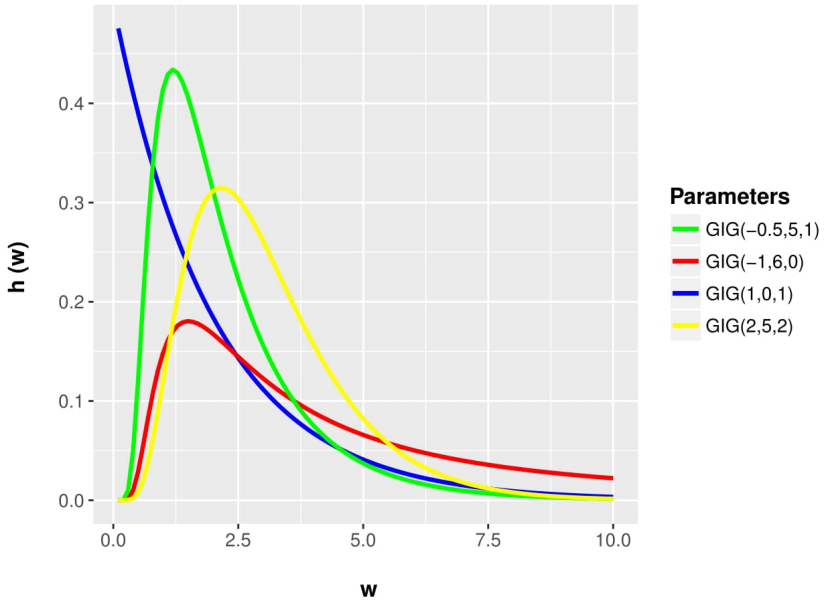

Figure 2.1: Densities of GIG distributions with various parameterizations.

Setting $\psi = \sqrt{ab}$ and $\eta = \sqrt{a/b}$ can result in a more interpretable parameterization which expresses the concentration and scale parameters explicitly (Koudou *et al.*,

2014; Browne and McNicholas, 2015). This parameterization is represented by:

$$h(w|\psi, a, b, \lambda) = \frac{(w/\eta)^{\lambda-1}}{2\eta K_\lambda(\psi)} \exp\left\{ -\frac{\psi}{2}\left(\frac{w}{\eta} + \frac{\eta}{w}\right)\right\}. \tag{2.5}$$

The GIG is a highly versatile distribution with special cases consisting of widely used distributions (Jorgensen, 2012; Browne and McNicholas, 2015), i.e., the gamma distribution ($b > 0$, $\lambda > 0$) and the inverse Gaussian distribution ($\lambda = -1/2$).

2.4 Mixtures of Generalized Hyperbolic Distributions

The generalized hyperbolic (GH) distribution was introduced by Barndorff-Nielsen (1977) with the intention of modeling aeolian sand deposits and dune movements. McNeil *et al.* (2015) has written the probability density of the GH distribution as follows:

$$\begin{aligned} f(\mathbf{x}|\boldsymbol{\vartheta}) = &\left[\frac{\chi + \delta(\mathbf{x}, \boldsymbol{\mu}|\boldsymbol{\Delta})}{\psi + \boldsymbol{\gamma}'\boldsymbol{\Delta}^{-1}\boldsymbol{\gamma}}\right]^{(\lambda-p/2)/2} \\ &\times \frac{[\psi/\chi]^{\lambda/2} K_{\lambda-p/2}\left(\sqrt{[\psi + \boldsymbol{\gamma}'\boldsymbol{\Delta}^{-1}\boldsymbol{\gamma}][\chi + \delta(\mathbf{x}, \boldsymbol{\mu}|\boldsymbol{\Delta})]}\right)}{(2\pi)^{p/2}|\boldsymbol{\Delta}|^{1/2} K_\lambda(\sqrt{\chi\psi}) \exp\{(\boldsymbol{\mu} - \boldsymbol{x})'\boldsymbol{\Delta}^{-1}\boldsymbol{\gamma}\}}, \end{aligned} \tag{2.6}$$

where $\boldsymbol{\vartheta} = (\boldsymbol{\mu}, \boldsymbol{\gamma}, \boldsymbol{\Delta}, \lambda, \chi, \psi,)$ is the set of parameters. K_λ is the modified Bessel function of the third kind with index parameter λ. The location and skewness parameters are given by $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$ respectively, while the concentration parameters are given by χ and ψ. The squared Mahalanobis distance between $\mathbf{x}$ and $\boldsymbol{\mu}$ is denoted by $\delta(\mathbf{x}, \boldsymbol{\mu}|\boldsymbol{\Delta}) = (\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Delta}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$, where $\boldsymbol{\Delta}$ is the positive definite scale matrix with

the constraint $|\boldsymbol{\Delta}| = 1$. The GH distribution can be defined in terms of a normal variance-mean mixture with a GIG mixing distribution using the following relation:

$$\mathbf{X} = \boldsymbol{\mu} + W\boldsymbol{\gamma} + \sqrt{W}\boldsymbol{V}, \tag{2.7}$$

where $W \sim \text{GIG}(\psi, \chi, \lambda)$ is a GIG random variable and $\mathbf{V} \sim \text{N}(\mathbf{0}, \boldsymbol{\Delta})$ is a latent multivariate Gaussian random variable (Hammerstein, 2010; McNicholas, 2016). Analogous to the GIG distribution several distributions can be recovered as special or limiting cases of the GH distribution. Some well known examples include the skew-t (Murray *et al.*, 2014) the variance-gamma (McNicholas *et al.*, 2017), and the normal-inverse Gaussian (Karlis and Santourian, 2009). Browne and McNicholas (2015) as well as Hu (2005) noted that using the GH distribution for the purposes of model-based clustering requires relaxing the constraint $|\boldsymbol{\Delta}| = 1$. However, relaxing this assumption results in the model being unidentifiable. This means that two or more parametrizations correspond to the same probability distribution making inference about the true parameters difficult. Browne and McNicholas (2015) introduced the parametrization $\omega = \sqrt{\psi\chi}$ and $\eta = \sqrt{\chi/\psi}$. This leads to another parametrization of the GH distribution:

$$\begin{aligned} f(\mathbf{x}|\boldsymbol{\vartheta}) = &\left[\frac{\omega + \delta(\mathbf{x}, \boldsymbol{\mu}|\boldsymbol{\Sigma})}{\omega + \boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}\right]^{(\lambda - p/2)/2} \\ &\times \frac{K_{\lambda - p/2}\left(\sqrt{[\omega + \boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}][\chi + \delta(\mathbf{x}, \boldsymbol{\mu}|\boldsymbol{\Sigma})]}\right)}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}K_{\lambda}(\sqrt{\omega})\exp\{(\boldsymbol{\mu} - \boldsymbol{x})'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}\}}, \end{aligned} \tag{2.8}$$

where $\boldsymbol{\Sigma}$ is the scale matrix and $\boldsymbol{\alpha}$ is the skewness parameter. Browne and McNicholas (2015) use the parametrization in (2.8) for the mixture of generalized hyperbolic

distributions (MGHD). Parameter estimation for the MGHD model is carried out using the expectation-maximization algorithm.

2.5 Principal Component Analysis

Principal component analysis (PCA) was first proposed by Pearson (1901) and was further developed by Hotelling (1933). Hotelling described it as a method which aimed at reducing the dimensionality of the data while retaining as much of the variability as possible. The classical view of principal components is as orthogonal linear transformations of the original data (Bouveyron and Brunet-Saumard, 2014). Consider a random vector $\mathbf{X} = (X_1, X_2, ..., X_p)$ with covariance matrix $\mathbf{\Sigma}$. Let $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$ be the ordered eigenvalues of $\mathbf{\Sigma}$, and let $\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_p$ be the corresponding eigenvectors. Then the ith principal component can be written as:

$$Y_i = e_{i1}X_1 + e_{i2}X_2 + ... + e_{ip}X_p, \tag{2.9}$$

for $i = 1, 2, ..., p$. Where the ith principal component is the linear combination of the X_i's with maximum variance conditional on it being orthogonal to the first $i-1$ components (Johnson and Wichern, 2007). The variance of the kth principal component is given by λ_k. Consequently, the proportion of total variation explained by the first k principal components is given by:

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i}. \tag{2.10}$$

An important consideration when using PCA for statistical analysis is determining how many principal components to retain in the model. This is typically addressed by considering the amount of total sample variance explained (Johnson and Wichern, 2007).

2.6 Mixtures of Generalized Hyperbolic Factor Analyzers

2.6.1 Factor Analysis

Factor analysis is another popular dimension reduction technique that was introduced by Spearman (1904) and its statistical properties were outlined by Bartlett (1953), and Lawley and Maxwell (1962). Factor analysis sets out to reduce the dimensionality of the p observed variables by substituting them with q latent factors where $q < p$. An important consideration is that the q latent factors must account for a sufficient amount of the variability originally explained by the p observed variables (McNicholas, 2016). Consider p-dimensional random variables $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$. Assume there are q-dimensional latent factors $\mathbf{U}_1, \mathbf{U}_2, ..., \mathbf{U}_n$, then the factor analysis model is given by:

$$\mathbf{X}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{U}_i + \boldsymbol{\epsilon}_i, \tag{2.11}$$

for $i = 1, 2, ..., n$, where $\boldsymbol{\Lambda}$ is a $p \times q$ matrix of factor loadings. The latent factor is denoted by $\mathbf{U}_i \sim \text{N}(\mathbf{0}, \mathbf{I}_q)$, with error terms given by $\boldsymbol{\epsilon}_i \sim \text{N}(\mathbf{0}, \boldsymbol{\Psi})$, where Ψ is a $p \times p$ diagonal positive definite matrix. Furthermore, it is assumed that both the $\mathbf{U}_i$ and $\boldsymbol{\epsilon}_i$ are independently distributed, moreover, they are independent of each other. Hence,

the marginal distribution of $\mathbf{X}_i$ under the factor analysis model is $N(\boldsymbol{\mu}, \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{'} + \boldsymbol{\Psi})$.

2.6.2 Mixtures of Factor Analyzers

The factor analysis approach can be extended to the mixture of factor analyzers model which assumes that:

$$\mathbf{X}_i = \boldsymbol{\mu}_g + \boldsymbol{\Lambda}_g \mathbf{U}_{ig} + \boldsymbol{\epsilon}_{ig}, \tag{2.12}$$

with probability π_g, for $g = 1, ..., G$ (Ghahramani and Hinton, 1997; McLachlan and Peel, 2000). This approach was further extended to develop the mixture of generalized hyperbolic factor analyzers model (MGHFA, see Tortora *et al.* (2015) for details). The density can be described by:

$$f(\mathbf{x}|\boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_h(\mathbf{x}|\boldsymbol{\theta}_g), \tag{2.13}$$

where $f_h(\mathbf{x}|\boldsymbol{\theta}_g)$ is as defined in equation (2.8) and $\boldsymbol{\theta}_g = (\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\alpha}_g, \omega_g, \lambda_g)$ with scale matrix for component g defined by, $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g$. The parameters for the MGHFA model are estimated using the alternating expectation-conditional maximization algorithm.

Chapter 3

Methodology

3.1 Likelihood

The likelihood function serves as a basis for making inferences about the unknown model parameters through utilizing maximum likelihood estimators. In general, the log-likelihood function is used to find the maximum likelihood estimates because it is easier to differentiate than the original likelihood function. Since $\log(x)$ is an increasing monotonic function, maximizing the log-likelihood is equivalent to maximizing the likelihood function. Consider n p-dimensional unlabelled observations $(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$ that we wish to separate into G similar groups. Let $(\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n)$ represent the unobserved labels for each observation where $\mathbf{z}_i = (z_{i1}, z_{i2}, ..., z_{iG})$. Then the model-based clustering log-likelihood for the set of observations $(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$ is given by,

$$\mathcal{L}(\boldsymbol{\vartheta}|\mathbf{x}) = \sum_{i=1}^{n} \log \left(\sum_{g=1}^{G} \pi_g f(\mathbf{x}_i|\boldsymbol{\theta}_g) \right). \tag{3.1}$$

In this instance the goal of clustering is to determine the value of the missing label $\mathbf{z}_i$ for each observation $\mathbf{x}_i$. This makes maximizing the log-likelihood difficult since information about the group labels $(\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n)$ is missing. Together the sets $(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$ and $(\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n)$ form the complete-data set (Bouveyron and Brunet-Saumard, 2014). The model-based clustering complete-data log-likelihood is given by,

$$l_c(\boldsymbol{\vartheta}; \mathbf{x}, \mathbf{z}) = \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \log(\pi_g f(\mathbf{x}_i|\boldsymbol{\theta}_g)), \tag{3.2}$$

where $z_{ig} = 1$ if the ith observation belongs to the gth cluster and $z_{ig} = 0$ otherwise. The parameters in the proposed model can be estimated using the expectation-maximization (EM) algorithm (Dempster *et al.*, 1977). The EM algorithm is one of the most popular methods for parameter estimation in the presence of missing data or unobserved variables.

3.2 Parameter Estimation

3.2.1 EM Algorithm

The EM algorithm is an iterative process that alternates sequentially between the E-step and the M-step, until convergence. The E-step consists of determining the expected value of the complete data log-likelihood given the observed data and the current parameter estimates. The expected log-likelihood of complete data is denoted by:

$$E[l_c(\boldsymbol{\vartheta}; \mathbf{x}, \mathbf{z})|\boldsymbol{\vartheta}^{(h)}] = \sum_{g=1}^{G} \sum_{i=1}^{n} {z_{ig}}^{(h)} \log(\pi_g f(\mathbf{x}_i|\boldsymbol{\theta}_g)), \tag{3.3}$$

where $\boldsymbol{\vartheta}^{(h)}$ denotes the estimate for $\boldsymbol{\vartheta}$ after the hth iteration and ${z_{ig}}^{(h)} = E[z_{ig} = 1|\boldsymbol{x}_i, \boldsymbol{\vartheta}^{(h)}]$ (Bouveyron and Brunet-Saumard, 2014). During the M-step, the expected log-likelihood from the previous step is maximized and the current estimate of $\boldsymbol{\vartheta}$ is updated for the next iteration. A remarkable feature about the EM algorithm is that each iteration guarantees the log-likelihood function must remain the same or show improvement until the desired stopping criterion is reached.

3.2.2 AECM Algorithm

A variation of the EM algorithm is the expectation-conditional maximization (ECM) algorithm which was first established by Meng and Rubin (1993). Overall the ECM is similar to the EM algorithm but with one key distinction: the maximization step is replaced by multiple conditional maximization (CM) steps (McNicholas, 2016). A modified version of the ECM algorithm is alternating expectation-conditional maximization (AECM) algorithm (Meng and Van Dyk, 1997). This version of the algorithm allows the basis for incomplete data to vary for each conditional maximization step (Murray *et al.*, 2014). The AECM algorithm is particularly useful for models like the mixture of hyperbolic factor analyzers model, which has its source of incomplete data stem from both the unknown group labels, $\{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n\}$ and the unobserved factors, $\{\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_n\}$.

3.2.3 Method Initialization

In consideration of detecting the global optimum it is necessary to start with a good choice the initial values ($\boldsymbol{\vartheta}^0$) for the model and for several iterations to be run; in order to compensate for the effects of random initializations (McLachlan *et al.*, 2002). A

commonly used initialization method is the k-means clustering algorithm (Hartigan and Wong, 1979). The objective of the k-means algoritm is to partition a set of observations into k clusters such that the distances between points within clusters and the cluster centers are minimized.

3.2.4 Convergence

One widely used stopping criterion for the EM algorithm is a measure based on lack of improvement in the log-likelihood. That is:

$$l^{r+1} - l^r < \epsilon, \tag{3.4}$$

where l^r is the observed log-likelihood value for the rth iteration and ϵ is some arbitrarily small value. The stopping criterion for the AECM algorithm uses Aitken's acceleration (Aitken, 1926), which estimates the asymptotic maximum of the log-likelihood. Aitken's acceleration at iteration r can be written as:

$$a^r = \frac{l^{(r+1)} - l^{(r)}}{l^{(r)} - l^{(r-1)}} \tag{3.5}$$

and the asymptotic estimate of the log-likelihood at iteration the $r + 1$ is:

$$l_\infty^{(r+1)} = l^{(r)} + \frac{1}{1 - a^{(r)}}(l^{(r+1)} - l^{(r)}), \tag{3.6}$$

McNicholas *et al.* (2010) considers the algorithm to have reached its exiting condition when,

$$l_\infty^{r+1} - l^r < \epsilon. \tag{3.7}$$

3.3 Predicted Classifications

Once the desired stopping criterion is achieved and parameter estimation is complete, the predicted group labels are retrieved from the *a posteriori* probability that observation $\boldsymbol{x}_i$ belongs to the gth component:

$$\hat{z}_{ig} = \frac{\hat{\pi}_g f_g(\mathbf{x}_i|\hat{\boldsymbol{\vartheta}}_g)}{\sum_{h=1}^{G} \hat{\pi}_h f_h(\mathbf{x}_i|\hat{\boldsymbol{\vartheta}}_h)}, \tag{3.8}$$

for $i = 1, ..., n$ and $g = 1, ..., G$. The numerical values for these *a posteriori* classifications can be designated as "soft" or as "hard" depending on the nature of the problem. "Soft" classifications consider the $\hat{z}_{ig}$ as each observation's probability of belonging to each component; where as "hard" classifications restrict the $\hat{z}_{ig} \in \{0, 1\}$ (McNicholas, 2016). In practice, if one wishes to convert the *a posteriori* classification from "soft" to "hard", this can be done so using maximum *a posteriori* (MAP) classifications, where

$$\text{MAP}\{\hat{z}_{ig}\} = \begin{cases} 1 & \text{if} \quad g = \text{argmax}_h\{\hat{z}_{ih}\}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.9}$$

3.4 Model Selection

Penalized likelihood approaches such as the Bayesian information criterion (BIC; Schwarz (1978)) are used for determining the number of components used for model-based clustering. BIC is given by:

$$\text{BIC} = 2l(\hat{\boldsymbol{\vartheta}}) - \rho \log n \tag{3.10}$$

where ρ is the number of free parameters in the model, n is the number of observations, and $l(\hat{\boldsymbol{\vartheta}})$ is the optimized log-likelihood value. The use of the BIC for selecting the number of components in a mixture model has been supported by Leroux (1992) and Keribin (2000). The BIC can be used for selecting the number of latent factors in the factor analysis model. Simulation studies by Lopes and West (2004) corroborate the use of BIC in selecting the number of latent factors.

3.5 Performance Assessment

The adjusted Rand index (ARI), introduced by Hubert and Arabie (1985) is the typically used method for assessing classification performance in model-based clustering. The ARI stems from the Rand index (Rand, 1971), which is the ratio of pairwise agreements to the total number of pairs. The Rand index can take on any value from 0 to 1, where 1 is indicative of perfect class agreement. The main drawback with the Rand index is that it has a positive expected value under random classification. The ARI was designed to compensate for this and it is defined as:

$$\text{ARI} = \frac{\text{index} - \text{expected index}}{\text{maximum index} - \text{expected index}}. \tag{3.11}$$

Analogous to the Rand index, an ARI value of 1 refers to a perfect class agreement. However, this time the expected value of the ARI under random classification is 0. Negative values for the ARI are possible if the classifications are worse than what would be expected by random classification.

Chapter 4

Application

In this chapter, model-based clustering experiments were conducted utilizing both real and simulated data. The class labels were treated as missing and all data sets were scaled. The R package MixGHD created by Tortora et al., (2015) contains the functions: MGHD and MGHFA, which were used to implement the MGHD and MGHFA methods respectively. These clustering methods were initialized using the k-means algorithm with the best fitting models being chosen via the BIC. The number of principal components retained for each analysis was contingent upon a threshold of 85 percent of the proportion of variance being explained. The classification performance for these approaches was then assessed using the ARI.

4.1 Simulated Studies

4.1.1 Twonorm Data

The twonorm data set contained 20 numerical attributes over 7400 observations and it was simulated using two classes from the multivariate distribution with unit covariance (Breiman, 1996). The first class had mean vector $(a, a, ..., a)$ while the second class had mean vector $(-a, -a, ..., -a)$, where $a = \frac{2}{\sqrt{20}}$. The MGHFA model was fitted to data for $G = 1, 2, ..., 5$ components and $q = 1, 2, ..., 5$ latent factors. The BIC selected $G = 2$ components with $q = 2$ latent factors, and the associated model had an excellent classification performance (Table 4.2; ARI = 0.914). Next, the MGHD model was applied to the first 16 principal components with BIC selecting $G = 2$ components. The aforementioned components explained roughly 85 percent of the total variation in the data. The corresponding model had an excellent classification performance (Table 4.2; ARI = 0.915).

Table 4.1: Cross-tabulation of the true versus predicted class labels for model-based clustering on twonorm data.

	cluster	
label	A	B
Type I	3630	73
Type II	87	3610

(a) MGHFA

	cluster	
label	A	B
Type I	3641	62
Type II	101	3596

(b) PCA & MGHD

Table 4.2: ARI values and misclassification rates (MCR), based on predicted classifications for the unlabelled observations, for the MGHFA and PCA & MGHD models.

	MGHFA	PCA & MGHD
ARI	0.914	0.915
MCR	0.022	0.022

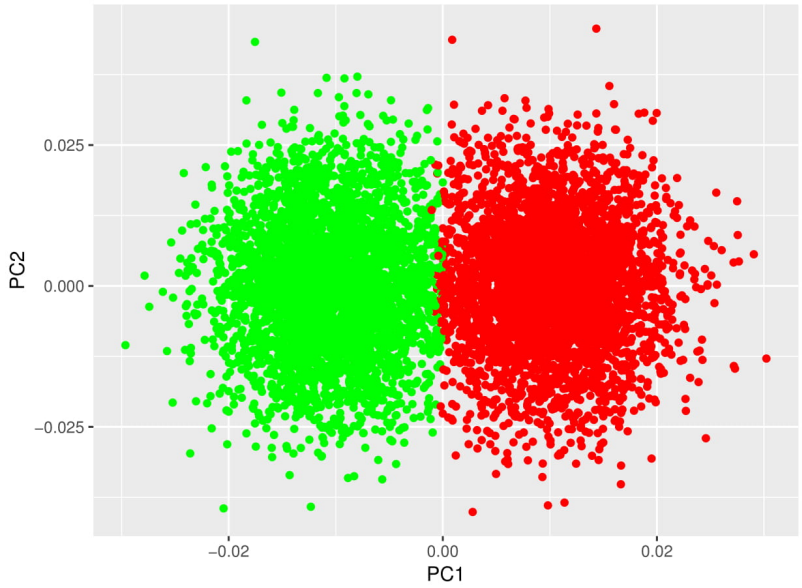

Figure 4.2: Projection of the clustered twonorm data in the 2 first principal components of PCA.

4.1.2 Ringworm Data

The ringworm data set contained 20 quantitative features over 7400 instances and it was simulated using two classes from the multivariate distribution (Breiman, 1996). The first class had a zero mean vector and a covariance matrix with 4's along its main diagonal, and 0's on the off diagonal entries. The second class had mean vector $(a, a, ..., a)$ and a unit covariance matrix. The MGHFA model was fitted to the data resulting in the BIC selecting $G = 2$ components and $q = 1$ latent factors. The corresponding model had an excellent classification performance (Table 4.4; ARI = 0.934). Roughly 87 percent of the total variation in the data is accounted for by the first 17 principal components. As a results, the MGHD model was applied to the first 17 principal components, resulting in a model with excellent classification performance (Table 4.4; ARI = 0.927).

Table 4.3: Cross-tabulation of the true versus predicted class labels for model-based clustering on ringworm data.

	cluster	
label	A	B
Type I	3558	106
Type II	18	3718

(a) MGHFA

	cluster	
label	A	B
Type I	3565	99
Type II	38	3698

(b) PCA & MGHD

Table 4.4: ARI values and misclassification rates (MCR), based on predicted classifications for the unlabelled observations, for the MGHFA and PCA & MGHD models.

	MGHFA	PCA & MGHD
ARI	0.934	0.927
MCR	0.017	0.018

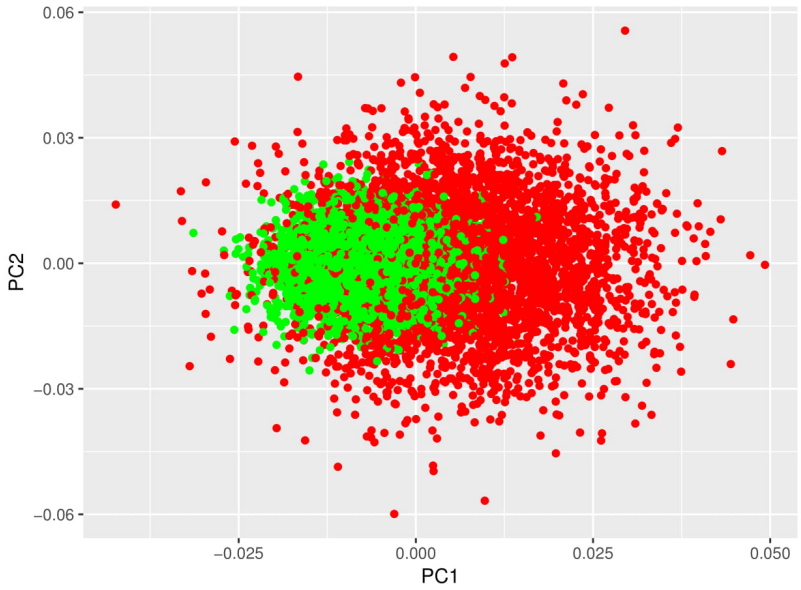

Figure 4.3: Projection of the clustered ringnorm data in the 2 first principal components of PCA.

4.1.3 Waveform Data

This data was retrieved from the UCI machine learning repository. The data contained 21 attributes measured over 5000 observations (Breiman, 1996). Each observation was generated from added noise attributes with mean 0 and variance 1. The observations can be split into three equal partitions according to wave type. The MGHFA model was fitted to this data resulting in the BIC selecting $G = 3$ components and $q = 1$ latent factors. The associated model had a poor classification performance (Table 4.6; ARI = 0.531). Roughly 86 percent of the total variation in the data was explained by the first 12 components. Consequently, the MGHD model was fitted to the first 12 principal components with BIC selecting $G = 3$ components. The resulting model had a better classification performance (Table 4.6; ARI = 0.601) compared to the approach using MGHFA.

Table 4.5: Cross-tabulation of the true versus predicted class labels for model-based clustering on waveform data.

	cluster		
label	A	B	C
Type I	1261	251	145
Type II	79	1420	148
Type III	189	101	1406

(a) MGHFA

	cluster		
label	A	B	C
Type I	1247	199	211
Type II	79	1454	114
Type III	55	95	1546

(b) PCA & MGHD

Table 4.6: ARI values and misclassification rates (MCR), based on predicted classifications for the unlabelled observations, for the MGHFA and PCA & MGHD models.

	MGHFA	PCA & MGHD
ARI	0.531	0.601
MCR	0.183	0.151

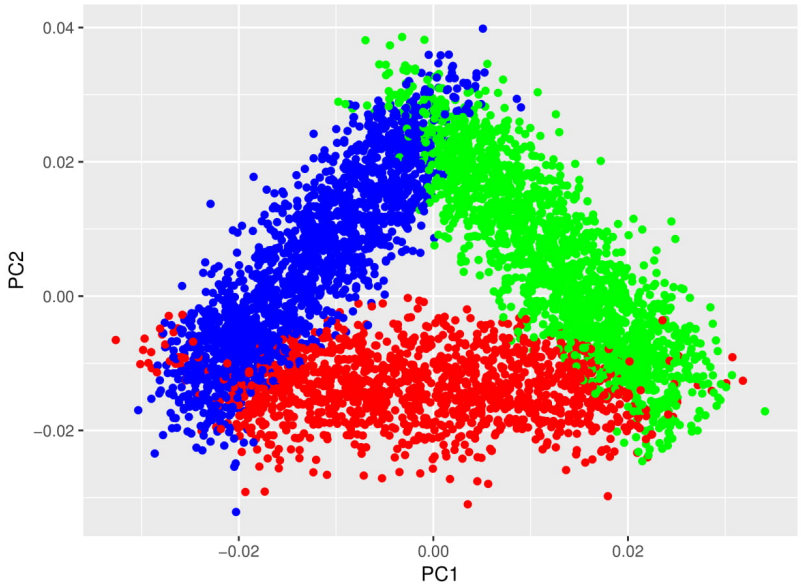

Figure 4.4: Projection of the clustered waveform data in the 2 first principal components of PCA.

4.2 Real Data

4.2.1 Wine Data

This data set is available in the pgmm library (McNicholas *et al.*, 2011) for R. The data contained 27 chemical and physical measurements on three types of wine, namely Barolo, Grignolino, and Barbera. There were 178 observations of these three types of wine which were all cultivated in the same region of Italy (Forina et al., 1986). Fitting the MGHFA model to this data resulted in the BIC selecting $G = 3$ componenets and $q = 1$ latent factors. The corresponding model has a fairly good classification performance (Table 4.8; ARI = 0.714). Implementing PCA revealed that the first 12 principal components accounted for roughly 85 percent of the total variation in the wine data. Based on this the MGHD model was initiated using the first 12 principal

components and this resulted in a model with a very poor classification performance (Table 4.8; ARI = 0.365) compared to the MGHFA approach.

Table 4.7: Cross-tabulation of the true versus predicted class labels for model-based clustering on wine data.

	cluster		
label	A	B	C
Barolo	50	9	0
Grignolino	2	62	7
Barbera	0	0	48

(a) MGHFA

	cluster	
label	A	B
Barolo	57	2
Grignolino	53	18
Barbera	0	48

(b) PCA & MGHD

Table 4.8: ARI values and misclassification rates (MCR), based on predicted classifications for the unlabelled observations, for the MGHFA and PCA & MGHD models.

	MGHFA	PCA & MGHD
ARI	0.714	0.365
MCR	0.101	0.579

4.2.2 Satellite data

The satellite data was collected from the UCI machine learning repository. The data contained 36 quantitative attributes measured over 4435 observations. Observations were collected with the goal of being able to classify 3×3 neighbourhoods based on a central pixel (Srinivasan, 1993). This data set contained six classes: red soil, cotton crop, grey soil, damp grey soil, soil with vegetation stubble, and very damp grey soil. Applying the MGHFA model to this data leads to the BIC selecting $G = 6$ components and $q = 1$ latent factors. The corresponding model gives a poor classification performance (Table 4.11; ARI = 0.397). Using PCA we found that about 89 percent of the total variation in the data is explained by the first three

principal components. As a result the MGHD model was applied using the first three principal components with BIC selecting $G = 6$. The associated model had a better classification performance (Table 4.11; ARI = 0.496) than the MGHFA approach.

Table 4.9: Cross-tabulation of the true versus predicted class labels for model-based clustering on Landsat Satellite data using MGHFA.

	cluster					
label	A	B	C	D	E	F
Red Soil	910	0	0	12	141	9
Cotton Crop	0	324	0	151	4	0
Grey Soil	0	0	721	9	104	127
Damp Soil	0	0	255	17	72	71
Very Damp Soil	58	29	0	68	261	54
Vegetation Stubble	0	1	560	19	131	327

Table 4.10: Cross-tabulation of the true versus predicted class labels for model-based clustering on Landsat Satellite data using MGHD & PCA.

	cluster					
Red Soil	526	0	15	511	20	0
Cotton Crop	0	416	0	0	63	0
Grey Soil	6	0	875	4	75	1
Damp Soil	1	0	125	4	262	23
Very Damp Soil	23	30	1	0	393	23
Vegetation Stubble	0	0	29	0	430	579

Table 4.11: ARI values and misclassification rates (MCR), based on predicted classifications for the unlabelled observations, for the MGHFA and PCA & MGHD models.

	MGHFA	PCA & MGHD
ARI	0.397	0.496
MCR	0.423	0.370

Chapter 5

Conclusions

In this thesis, the clustering of high-dimensional data using the MGHFA approach was compared against the procedure utilizing PCA for feature extraction followed by clustering using the MGHD. This analysis was conducted by applying the aforementioned methods to both simulated and real data.

For the two data sets simulated from the multivariate normal distribution the approach of using PCA & MGHD demonstrated ARI's comparable to that of the MGHFA approach. Executing either approach on these data resulted in excellent classification performances with over 95 percent class agreement. For the waveform data set, both the associated models had relatively poor classification performances with the PCA & MGHD approach displaying a slightly higher ARI.

For the wine data set, using the MGHFA model resulted in a good classification performance while the quality of the partition obtained by using the PCA & MGHD model was disappointing. For the satellite data set the PCA & MGHD model performed slightly better than the MGHFA model but the respective classification performances for both approaches was quite poor.

Overall, the PCA & MGHD method performed just as well or better than the MGHFA method for most of the data sets. However, caution must be exercised when utilizing PCA for dimension reduction in conjunction with clustering tasks. This is because, as noted by Bouveyron and Brunet-Saumard (2014), PCA was not designed to take into account the clustering scheme and this may lead to sub-optimal partitioning of the data. Possible future research directions could consist of the using MGHD together with other dimension reduction techniques such as variable selection or subspace methods. These approaches can simultaneously find partitions and reduce the dimensionality of the data.

Appendix A

Bessel Functions

In general, Bessel functions are the canonical solutions $y(x)$ of the Bessel differential equation:

$$x^2\frac{d^2y}{dx^2} + x\frac{dy}{dx} + (x^2 - \lambda^2)y = 0,$$

for an arbitrary complex number λ. The solutions of the Bessel equation are called modified Bessel functions of the first and second kind when the domain of x includes the complex numbers (Abramowitz and Stegun, 1968). The modified Bessel function of the second kind, also known as the modified Bessel function of the third kind, can be expressed as follows:

$$K_\lambda(x) = \frac{\pi}{2}\frac{I_{-\lambda}(x) - I_\lambda(x)}{\sin(\lambda\pi)},$$

where $I_\lambda(x)$ is the modified Bessel function of the first kind obtained from the Bessel differential equation using the power series approach.

$$I_\lambda(x) = \sum_{m=0}^{\infty} \frac{1}{m!\,\Gamma(m + \lambda + 1)} \left(\frac{x}{2}\right)^{2m+\lambda}.$$

$K_\lambda(x)$ is an exponentially decaying function with a singularity at $x = 0$ (Abramowitz and Stegun, 1968) that are an important component of the generalized inverse Gaussian and the generalized hyperbolic distributions.

Bibliography

Abramowitz, M. and Stegun, I. A. (1968). *Handbook of Mathematical Tables*. New York: Dover Publications.

Aitken, A. (1926). A series formula for the roots of algebraic and transcendental equations. *Proceedings of the Royal Society of Edinburgh*, **45**(1), 14–22.

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.

Barndorff-Nielsen, O. (1977). Exponentially decreasing distributions for the logarithm of particle size. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 353, pages 401–419.

Barndorff-Nielsen, O. and Halgreen, C. (1977). Infinite divisibility of the hyperbolic and generalized inverse gaussian distributions. *Probability Theory and Related Fields*, **38**(4), 309–311.

Bartlett, M. S. (1953). Factor analysis in psychology as a statistician sees it. In *Uppsala Symposium on Psychological Factor Analysis*, number 3, pages 23–43. Copenhagen: Ejnar Mundsgaards.

Bellman, R. (1957). *Dynamic programming*. Princeton University Press.

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

Breiman, L. (1996). Bias, variance, and arcing classifiers (technical report 460). *Statistics Department, University of California*.

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

Daas, P. J., Puts, M. J., Buelens, B., and van den Hurk, P. A. (2015). Big data as a source for official statistics. *Journal of Official Statistics*, **31**(2), 249–262.

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.

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.

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.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**(3-4), 237–264.

Halphen, E. (1941). Sur un nouveau type de courbe de fréquence. *Comptes Rendus de l'Académie des Sciences*, **213**, 633–635.

Hammerstein, E. A. v. (2010). *Generalized hyperbolic distributions: Theory and Applications to CDO Pricing*. Ph.D. thesis, PhD thesis, Universität Freiburg.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **28**(1), 100–108.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, **24**(6), 417.

Hu, W. (2005). *Calibration of Multivariate Generalized Hyperbolic Distributions Using the EM algorithm, with Applications in Risk Management, Portfolio Optimization and Portfolio Credit Risk*. Ph.D. thesis, The Florida State University, Tallahassee.

Huber, P. J. (1985). Projection pursuit. *The Annals of Statistics*, **13**(2), 435–475.

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

Johnson, R. and Wichern, D. (2007). *Applied multivariate correspondence analysis*, pages 430–444. Prentice-Hall New Jersey, sixth edition.

Jorgensen, B. (2012). *Statistical properties of the generalized inverse Gaussian distribution*, volume 9. Springer Science & Business Media.

Juárez, M. A. and Steel, M. F. (2010). Model-based clustering of non-gaussian panel data based on skew-t distributions. *Journal of Business & Economic Statistics*, **28**(1), 52–66.

Karlis, D. and Santourian, A. (2009). Model-based clustering with non-elliptically contoured distributions. *Statistics and Computing*, **19**(1), 73–83.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A*, **62**(1), 49–66.

Koudou, A. E., Ley, C., *et al.* (2014). Characterizations of gig laws: A survey. *Probability Surveys*, **11**, 161–176.

Lawley, D. N. and Maxwell, A. E. (1962). Factor analysis as a statistical method. *Journal of the Royal Statistical Society. Series D (The Statistician)*, **12**(3), 209–229.

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

Lopes, H. F. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, **14**(1), 41–67.

Mangasarian, O. L., Street, W. N., and Wolberg, W. H. (1995). Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, **43**(4), 570–577.

McLachlan, G. and Peel, D. (2000). Mixtures of factor analyzers. In *Proceedings of the Seventh International Conference on Machine Learning*, pages 599–606.

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

McNeil, A. J., Frey, R., and Embrechts, P. (2015). *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press.

McNicholas, P., Jampani, K., McDaid, A., Murphy, T., and Banks, L. (2011). pgmm: Parsimonious gaussian mixture models. *R package version*, **1**(1).

McNicholas, P. D. (2016). *Mixture model-based classification*. CRC Press.

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 & Data Analysis*, **54**(3), 711–723.

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*, pages 369–385. Springer.

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

Meng, X.-L. and Van Dyk, D. (1997). The EM algorithm-an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **59**(3), 511–567.

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

Pearson, K. (1901). Principal components analysis. *The London, Edinburgh and Dublin Philosophical Magazine and Journal*, **6**(2), 566.

R Core Team, R. (2017). R: A Language and Environment for Statistical Computing. Version 3.3.3. R Foundation for Statistical Computing, Vienna, Austria. *URL: https://www. R-project. org*.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**(336), 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.

Sigillito, V. G., Wing, S. P., Hutton, L. V., and Baker, K. B. (1989). Classification of radar returns from the ionosphere using neural networks. *Johns Hopkins APL Technical Digest*, **10**(3), 262–266.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, **15**(1), 72–101.

Tiedeman, D. (1955). On the study of types. In S. B. Sells, editor, *Symposium on Pattern Analysis*, pages 1–14, Randolph Field, TX: Air University. U.S.A.F. School of Aviation Medicine.

Tortora, C., McNicholas, P. D., and Browne, R. P. (2015). A mixture of generalized hyperbolic factor analyzers. *Advances in Data Analysis and Classification*, **10**(4), 423–440.

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