Mixtures of Skew-t Factor Analyzers

by

Paula Murray

A Thesis
presented to
The University of Guelph

In partial fulfilment of requirements
for the degree of
Master of Science
in
Bioinformatics

Guelph, Ontario, Canada

ABSTRACT

Mixtures of Skew-t Factor Analyzers

Paula Murray
University of Guelph, 2012

Advisors:
Dr. P.D. McNicholas
Dr. R.P. Browne

Model-based clustering allows for the identification of subgroups in a data set through the use of finite mixture models. When applied to high-dimensional microarray data, we can discover groups of genes characterized by their gene expression profiles. In this thesis, a mixture of skew-t factor analyzers is introduced for the clustering of high-dimensional data. Notably, we make use of a version of the skew-t distribution which has not previously appeared in mixture-modelling literature. Allowing a constraint on the factor loading matrix leads to two mixtures of skew-t factor analyzers models. These models are implemented using the alternating expectation-conditional maximization algorithm for parameter estimation with an Aitken's acceleration stopping criterion used to determine convergence. The Bayesian information criterion is used for model selection and the performance of each model is assessed using the adjusted Rand index. The models are applied to both real and simulated data, obtaining clustering results which are equivalent or superior to those of established clustering methods.

Acknowledgements

I would like to express my sincerest gratitude to my advisors, Dr. Paul McNicholas and Dr. Ryan Browne, for their guidance and commitment to my success as a graduate student. It has truly been a privilege to work with and learn from both of you.

Additionally, I would like to thank my examining committee members, Dr. Lewis Lukens and Dr. Dan Ashlock, for taking the time to examine this thesis.

Finally, to my family - Words are not enough to express my appreciation for your unwavering support throughout the course of my education. This work would not have been possible without your endless love and encouragement and I am truly blessed to have you all in my life.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Most research in the fields of molecular biology and genetics calls for the analysis of high-dimensional gene expression data. Cluster analysis allows us to look for groups of genes within a data set which are characterized by their gene expression profiles. Non-parametric methods such as agglomerative hierarchal clustering, k-means, and k-medioids clustering have been used to this end. In this thesis, we will work within the framework of model-based clustering by making use of parametric finite mixture models to estimate group membership. Finite mixture models assume that a population is composed of a finite number of subpopulations, making them well-suited for use in cluster analysis. However, many model-based clustering techniques are unsuitable for modelling high-dimensional data sets without using some form of dimension reduction.

Factor analysis is a well-known method of dimension reduction which assumes that a large number of observed variables can be modelled by a smaller number of latent variables. Mixtures of factor analyzers models have previously been developed which make used of both

the multivariate Gaussian distribution as well as the multivariate-t distribution. However, no mixtures of factor analyzers models have been implemented which utilize the skew-t distribution.

This thesis introduces a mixtures of skew-t factor analyzers model with the option of constraining the loading matrix, resulting in both a constrained and an unconstrained model. Notably, this form of the skew-t distribution, which exists as a special case of the generalized hyperbolic distribution, has not previously been seen in the mixture modelling literature. This particular variant was introduced by Barndorff-Nielsen and Shephard (2001) and applied to non-Gaussian processes, suggesting it is well-suited to situations in which the standard Gaussian assumption is not met.

We implement the models developed in this thesis using an alternating expectation-conditional maximization algorithm for parameter estimation and the Bayesian information criterion for model selection. We apply them to simulated data as well as four real data sets, two toy data sets with low dimension and two high-dimensional gene expression data sets. We find equivalent or better results from applying the models developed herein to these data sets as compared to well-established clustering techniques.

1.1 Thesis Outline

1.1.1 Chapter 2

This chapter discusses previous work that has appeared in the mixture modelling literature. It gives a review of and background information on methods that will be used in the formulation of the methodology in Chapter 3.

1.1.2 Chapter 3

In this chapter, we will outline the methodology used in this work. The mathematics behind the mixtures of generalized hyperbolic distributions and the mixtures of skew-t distributions will be discussed. The implementation of the AECM algorithm for parameter estimation and the corresponding parameter updates will be covered as well as the constraint on the factor loading matrix.

1.1.3 Chapter 4

This chapter contains the results of applying the mixtures of skew-t factors models to both real and simulated data. These results will be compared to those obtained using `mclust` and `pgmm`.

1.1.4 Chapter 5

Chapter 5 will conclude this thesis with a discussion of the results and suggestions for future work.

Chapter 2

Model-based Clustering

2.1 Mixture Models

2.1.1 Finite Mixture Models

The use of finite mixture models in statistical analysis goes back as far as Pearson (1894) who fitted a mixture of two normal probability functions to data collected from crabs sampled from the Bay of Naples. This method assumes that the data are sampled from a population consisting of a finite collection of subpopulations, usually of the same statistical distribution type. More specifically, we can say that a p-dimensional random vector $\mathbf{X}$ arises from a finite mixture model if for all $\mathbf{x} \subset \mathbf{X}$,

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

such that

$$\boldsymbol{\vartheta} = (\pi_1, ..., \pi_G, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_G),\ \pi_g > 0, \text{ and } \textstyle\sum_{g=1}^{G} \pi_g = 1,$$

where π_g is the gth mixing proportion, $\boldsymbol{\theta}_g$ is a vector of parameters, and $f_g(\mathbf{x}|\boldsymbol{\theta}_g)$ is the gth component density. Typically a family of mixture models will be fit to the data with some variant of the expectation-maximization (EM) algorithm (Dempster et al., 1977) used to estimate the model parameters. A model selection criterion such as the Bayesian information criterion (BIC; Schwarz, 1978) or integrated completed likelihood criterion (ICL; Biernacki et al., 2000) is then used to select the best model. Reviews of finite mixture models can be found in Everitt and Hand (1981), Titterington et al. (1985), McLachlan and Basford (1988), and McLachlan and Peel (2000a).

Cluster analysis aims to identify groups of observations within a data set which are in some sense similar. Clustering may be performed using a variety of techniques, such as non-parametric methods including k-means clustering or agglomerative hierarchal clustering as well as by the fitting of parametric finite mixture models with maximum likelihood estimation used to obtain parameter estimates. However, Marriott (1974) writes that this mixture-likelihood based approach "is about the only clustering technique that is entirely satisfactory from the mathematical point of view. It assumes a well-defined mathematical model, investigates it by well-established statistical techniques, and provides a test of significance for the results." This sound mathematical framework provided for cluster analysis makes model-based clustering our choice as the foundation for the models to be developed in this thesis. See Fraley and Raftery (2002) and McNicholas (2011) for a detailed review of model-based clustering.

2.1.2 Gaussian Mixture Models

To date most mixture modelling has been performed with the component densities taken to be multivariate Gaussian (i.e., McNicholas and Murphy, 2008; McLachlan et al., 2002; Banfield and Raftery, 1993; Celeux and Govaert, 1995). Gaussian mixture models have become popular due to their mathematical tractability and are also useful specifically in a bioinformatics context in that the covariance structure of the Gaussian mixture model may account for correlation within the expression profile (McNicholas and Murphy, 2010). The density of a Gaussian mixture model is written as

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

with vector of parameters $\boldsymbol{\vartheta}$ and component densities given by

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

where $\phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ is the density of the multivariate Gaussian distribution with mean $\boldsymbol{\mu}_g$, covariance matrix $\boldsymbol{\Sigma}_g$, and mixing proportions π_g. A detailed review of Gaussian mixture modelling may be found in Fraley and Raftery (2002).

2.1.3 Mixtures of Generalized Hyperbolic Distributions

Mixtures of generalized hyperbolic distributions, introduced by Browne and McNicholas (2012), offer an alternative to mixtures of Gaussian distributions. This adds to the growing amount of literature on non-Gaussian approaches including mixtures of multivariate-t distributions (i.e., McLachlan and Peel, 1998; Peel and McLachlan, 2000), skew-normal distributions (i.e., Lin, 2009), skew-t distributions (i.e., Lin, 2012; Lee and McLachlan, 2011; Vrbik and McNicholas, 2012), and other approaches (i.e., Karlis and Meligkotsidou, 2007;

Browne et al., 2012). The model density is written as

$$f(\mathbf{x}|\boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{x}|\boldsymbol{\vartheta}_g), \tag{2.4}$$

where

$$\begin{aligned} f(\mathbf{x} \mid \boldsymbol{\vartheta}_g) = & \left[\frac{\chi_g + \delta(\mathbf{x}, \boldsymbol{\mu}_g|\boldsymbol{\Sigma}_g)}{\psi_g + \boldsymbol{\alpha}_g' \boldsymbol{\Sigma}_g^{-1} \boldsymbol{\alpha}_g}\right]^{(\lambda_g - p/2)/2} \\ & \times \frac{[\psi_g/\chi_g]^{\lambda_g/2} K_{\lambda_g - p/2}\left(\sqrt{[\psi_g + \boldsymbol{\alpha}_g \boldsymbol{\Sigma}_g^{-1} \boldsymbol{\alpha}_g][\chi_g + \delta(\mathbf{x}, \boldsymbol{\mu}_g|\boldsymbol{\Sigma}_g)]}\right)}{(2\pi)^{p/2} \mid \boldsymbol{\Sigma}_g \mid^{1/2} K_{\lambda_g}(\sqrt{\chi_g \psi_g}) \exp(\boldsymbol{\mu}_g - \mathbf{x})' \boldsymbol{\Sigma}_g^{-1} \boldsymbol{\alpha}_g}, \end{aligned} \tag{2.5}$$

is the density of the generalized hyperbolic distribution, $\boldsymbol{\vartheta}_g = (\lambda_g, \chi_g, \psi_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\alpha}_g)$ is a vector of parameters, and $\delta(\mathbf{x}, \boldsymbol{\mu}|\boldsymbol{\Sigma}) = (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$ is the squared Mahalanobis distance between $\mathbf{x}$ and $\boldsymbol{\mu}$. The generalized hyperbolic distribution includes several distributions such as the Gaussian, skew-normal, and variance-gamma distributions as special cases. The density of the skew-t distribution can be obtained from Equation (2.5) by setting $\lambda = -\nu/2$, $\chi = \nu$ and $\psi \to 0$. This leads to a mixture of skew-t distributions with a model density of the form

$$f(\mathbf{x}) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{x}| - \nu_g/2, \nu_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\alpha}_g). \tag{2.6}$$

We will see this model developed in Chapter 3 and applied to data in Chapter 4.

2.2 The Expectation-Maximization Algorithm

The expectation-maximization (EM) algorithm (Dempster et al., 1977) is an iterative procedure for the computation of maximum likelihood estimates. This method is applied to data that is viewed as being incomplete which makes the EM algorithm useful for clustering problems in which the component labels are unknown. The fitting of mixture distributions by the EM algorithm has been studied by Day (1969) and Wolfe (1965) on the way to becoming

a commonly used method for the fitting of finite mixture models. The algorithm alternates between two steps, an expectation-step (E-step) and a maximization-step (M-step). In the E-step we compute the value of the complete-data log-likelihood based on the current model parameters where the complete-data consist of the observed and missing data. The M-step consists of maximizing the expected value of the complete-data log-likelihood with respect to the model parameters. These two steps are repeated until some convergence criterion is reached. A review of the EM algorithm, its extensions, and various applications is given by McLachlan and Krishnan (2008).

2.2.1 Expectation-Conditional Maximization Algorithm

The M-step of the EM algorithm involves the updating of parameter values by an often computationally straightforward equation. However, when the expected complete-data likelihood is relatively complicated, the M-step can be replaced by a number of computationally simpler conditional maximization steps. This expectation-conditional maximization (ECM) algorithm (Meng and Rubin, 1993) typically requires more iterations to achieve satisfactory parameter estimates but converges more quickly in terms of overall computation time.

2.2.2 Alternating Expectation-Conditional Maximization Algorithm

Meng and van Dyk (1997) introduce an extension of the ECM algorithm known as the alternating expectation-conditional maximization (AECM) algorithm. This algorithm allows the complete-data to change between two CM-steps. We will make use of the AECM algorithm to implement the skew-t factor analyzer models developed in Chapter 3 which have varying sources of incomplete data. Refer to McLachlan and Krishnan (2008) for a detailed

overview of the AECM algorithm and its use in fitting mixtures of factor analyzers models.

2.2.3 Aitken's Acceleration Stopping Criterion

The Aitken's acceleration stopping criterion (Aitken, 1926) is one of many methods that may be used to determine convergence of the EM algorithm. Since the algorithm converges at a linear rate it may require many iterations to achieve satisfactory parameter estimates. To determine convergence, Aitken's acceleration estimates the asymptotic maximum of the log-likelihood at each iteration. Given that $l^{(k)}$ is the log-likelihood at the k^{th} iteration of the EM algorithm, the Aitken acceleration at iteration k is given by

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

The EM algorithm is considered to have converged when

$$l_{\infty}^{k} - l^{k} < \epsilon,$$

where ϵ is some small constant and

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

is the asymptotic estimate of the log-likelihood at iteration $k + 1$ (Böhning et al., 1994; Lindsay, 1995). Aitken's acceleration will be used in this thesis as the convergence criterion for the analysis in Chapter 4.

2.3 MCLUST

MCLUST is a family of mixture models in which the component densities are taken to be multivariate Gaussian. These models make use of an eigen-decomposed covariance structure

(Celeux and Govaert, 1995) such that the covariance matrix $\boldsymbol{\Sigma}_g$ in Equation (2.2) is of the form

$$\boldsymbol{\Sigma}_g = \boldsymbol{\lambda}_g \mathbf{D}_g \mathbf{A}_g \mathbf{D}_g',$$

where $\boldsymbol{\lambda}_g$ is a constant, $\mathbf{D}_g$ is a matrix consisting of the eigenvectors of $\boldsymbol{\Sigma}_g$, and $\mathbf{A}_g$ is a diagonal matrix with entries proportional to the eigenvalues of $\boldsymbol{\Sigma}_g$. Allowing various combinations of the constraints $\mathbf{A}_g = \mathbf{A}$, $\mathbf{D}_g = \mathbf{D}$, $\boldsymbol{\lambda}_g = \boldsymbol{\lambda}$, $\mathbf{A} = \mathbf{I}_p$, and $\mathbf{D} = \mathbf{I}_p$, where $\mathbf{I}_p$ is the $p \times p$ identity matrix, results in the ten models with varying numbers of covariance parameters which make up the MCLUST family. Note that in some of the models, the number of covariance parameters is quadratic in p making these models unsuitable for modelling high-dimensional data sets in which $n \ll p$. Gaussian mixture models are easily implemented using the `mclust` package (Fraley and Raftery, 1999) which is available through the software package R (R Core Team, 2012). Table 2.1 gives the various models available through `mclust` and the associated free covariance parameters.

2.4 Factor Analyzers

2.4.1 Introduction

Factor analyzers (Spearman, 1904) explain variability within a data set by replacing observed variables with a smaller number of unobserved factors. Therefore, they are useful for dimensionality reduction. However, factor analyzers are globally linear, a matter which is addressed by the mixtures of factor analyzers (MFA) model introduced by Ghahramani and Hinton (1997). This model uses a mixture of Gaussian distributions with a factor analysis

Table 2.1: The covariance structures and associated free covariance parameters available in the `mclust` software package

Model	Volume	Shape	Orientation	$\boldsymbol{\Sigma}_g$	Free covariance parameters
EII	Equal	Spherical	-	$\boldsymbol{\lambda}\mathbf{I}$	1
VII	Variable	Spherical	-	$\boldsymbol{\lambda}_g\mathbf{I}$	G
EEI	Equal	Equal	Axis-Aligned	$\boldsymbol{\lambda}\mathbf{A}$	p
VEI	Variable	Equal	Axis-Aligned	$\boldsymbol{\lambda}_g\mathbf{A}$	$p+G-1$
EVI	Equal	Variable	Axis-Aligned	$\boldsymbol{\lambda}\mathbf{A}_g$	$pG-G+1$
VVI	Variable	Variable	Axis-Aligned	$\boldsymbol{\lambda}_g\mathbf{A}_g$	pG
EEE	Equal	Equal	Equal	$\boldsymbol{\lambda}\mathbf{D}\mathbf{A}\mathbf{D}'$	$p(p+1)/2$
EEV	Equal	Equal	Variable	$\boldsymbol{\lambda}\mathbf{D}_g\mathbf{A}\mathbf{D}_g'$	$Gp(p+1)/2-(G-1)p$
VEV	Variable	Equal	Variable	$\boldsymbol{\lambda}_g\mathbf{D}_g\mathbf{A}_g\mathbf{D}_g'$	$Gp(p+1)/2-(G-1)(p-1)$
VVV	Variable	Variable	Variable	$\boldsymbol{\lambda}_g\mathbf{D}_d\mathbf{A}_g\mathbf{D}_g'$	$Gp(p+1)/2$

covariance structure which leads to a model of the form

$$f(\mathbf{x}) = \sum_{g=1}^{G} \pi_g \phi_g(\mathbf{x}|\boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}). \tag{2.7}$$

McLachlan and Peel (2000b) further generalize this covariance structure such that $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g + \boldsymbol{\Psi}_g$. Tipping and Bishop (1999) introduced a mixtures of probabilistic principal components analyzers model in which the distribution of the errors is isotropic. This leads to a covariance structure of the form $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g + \psi_g\mathbf{I}_p$. A mixture of multivariate t-factor analyzers model was introduced by McLachlan et al. (2007) with further developments by Andrews and McNicholas (2011).

2.4.2 Mixtures of Factor Analyzers

Using mixtures of factor analyzers, a p-dimensional data vector $\mathbf{X}$ can be modelled by a q-dimensional vector of latent factors $\mathbf{U}$ with $q \ll p$ such that

$$X = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{U} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\Lambda}$ is a matrix of factor loadings, $\mathbf{U} \sim N(\mathbf{0}, \mathbf{I}_q)$ is the vector of factors, and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Psi})$ with $\boldsymbol{\Psi}$ =diag$(\psi_1, \psi_2, ..., \psi_p)$. The marginal distribution of $\mathbf{X}$ is multivariate Gaussian with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$.

2.4.3 Woodbury Identity

The Woodbury Identity (Woodbury, 1950) states that given an $p \times p$ matrix $\mathbf{A}$, an $p \times q$ matrix $\mathbf{U}$, an $q \times q$ matrix $\mathbf{C}$ and a $q \times p$ matrix $\mathbf{V}$,

$$(\mathbf{A} + \mathbf{UCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{C}^{-1} + \mathbf{VA}^{-1}\mathbf{U})^{-1}\mathbf{VA}^{-1}.$$

Given the factor analysis covariance structure $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g$, we can set $\mathbf{U} = \boldsymbol{\Lambda}_g$, $\mathbf{V} = \boldsymbol{\Lambda}_g'$, $\mathbf{A} = \boldsymbol{\Psi}_g$ and $\mathbf{C} = \mathbf{I}_q$, and it follows that

$$(\boldsymbol{\Psi}_g + \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g')^{-1} = (\boldsymbol{\Psi}_g)^{-1} - (\boldsymbol{\Psi}_g)^{-1}\boldsymbol{\Lambda}_g(\mathbf{I}_q + \boldsymbol{\Lambda}_g'\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}_g)\boldsymbol{\Lambda}_g'\boldsymbol{\Psi}_g^{-1}.$$

For the modelling of high-dimensional data in which $n \ll p$, the calculation of the inverse of the $p \times p$ covariance matrix $\boldsymbol{\Sigma}_g^{-1}$ can be computationally intensive. However, the Woodbury Identity requires only the inversion of diagonal and $q \times q$ matrices to compute $(\boldsymbol{\Psi}_g + \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g')^{-1}$. Furthermore, we have the identity

$$\mid \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g \mid = \mid \boldsymbol{\Psi}_g \mid / \mid \mathbf{I}_q - (\boldsymbol{\Lambda}_g'(\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}_g)^{-1}\boldsymbol{\Lambda}_g) \mid$$

for the determinant of the covariance matrix. The Woodbury Identity will be used in this thesis to calculate both the inverse and determinant of the covariance parameter $\boldsymbol{\Sigma}_g$.

2.5 Parsimonious Gaussian Mixture Models

The family of parsimonious Gaussian mixture models (PGMM), introduced by McNicholas and Murphy (2008), consists of eight models which includes both mixtures of probabilistic principle component analyzers and mixtures of factor analyzers models. Allowing the loading and noise terms of the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$ to be equal or unequal across groups such that $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}_g$ and $\boldsymbol{\Psi} = \boldsymbol{\Psi}_g$ leads to the four PGMM models CUU, UCU, CUU, and CUU. McNicholas and Murphy (2008) further allow the isotropic constraint $\Psi_g = \psi_g\mathbf{I}_p$ on the error variance matrix, extending the number of models to eight. These models take the component densities to be multivariate Gaussian and are easily implemented using the `pgmm` package in R. The covariance structures for the PGMM models are given in Table 2.2.

In Gaussian mixture modelling, the majority of parameters which must be estimated come into the model through the covariance matrix $\mathbf{\Sigma}$. Recall that in a subset of the `mclust` models previously discussed in this chapter, the number of covariance parameters is quadratic in p. However, the number of free covariance parameters in each `pgmm` model is linear in p, making them efficient for the modelling of high-dimensional data in which p is large.

Table 2.2: The covariance structures and associated free covariance parameters of the parsimonious Gaussian mixture models.

ID	Loading Matrix	Error Variance	Isotropic	Free Covariance Parameters
CCC	Constrained	Constrained	Constrained	$[pq - q(q-1)/2] + 1$
CCU	Constrained	Constrained	Unconstrained	$[pq - q(q-1)/2] + p$
CUC	Constrained	Unconstrained	Constrained	$[pq - q(q-1)/2] + G$
CUU	Constrained	Unconstrained	Unconstrained	$[pq - q(q-1)/2] + Gp$
UCC	Unconstrained	Constrained	Constrained	$G[pq - q(q-1)/2] + 1$
UCU	Unconstrained	Constrained	Unconstrained	$G[pq - q(q-1)/2] + p$
UUC	Unconstrained	Unconstrained	Constrained	$G[pq - q(q-1)/2] + G$
UUU	Unconstrained	Unconstrained	Unconstrained	$G[pq - q(q-1)/2] + Gp$

2.6 Model Selection and Performance

2.6.1 Bayesian Information Criterion

The Bayesian information criterion (BIC) is commonly used in mixture modelling to select the number of components G and the best model from among a family of models. The BIC is given by

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

where $l(\mathbf{x}, \hat{\boldsymbol{\vartheta}})$ is the maximized log-likelihood, $\hat{\boldsymbol{\vartheta}}$ is the maximum likelihood estimate of the model parameters $\boldsymbol{\vartheta}$, ρ is the number of free parameters in the model, and n is the number of observations. Campbell et al. (1997) and Dasgupta and Raftery (1998) give support for the use of the BIC in mixture modelling. Furthermore, Leroux (1992) has demonstrated that, asymptotically, the BIC does not underestimate the true number of components. In this thesis, the BIC will be used to select the most appropriate factor analysis covariance structure as well as the number of mixture components G.

2.6.2 Rand Index

The Rand index (Rand, 1971) and the adjusted Rand index (ARI) (Hubert and Arabie, 1985) can be used to measure class agreement in model-based clustering. Calculated as a cross-tabulation between the true group memberships and the maximum *a posteriori* (MAP) classification of the observations, the Rand Index is expressed as

$$\frac{\text{number of pairwise agreements}}{\text{total number of pairs}},$$

where the total number of pairs is calculated as

$$^{n}C_{r} = \frac{n!}{(n-2)!2!}.$$

The Rand index has an expected value which is greater than zero. It is, therefore, preferable to make use of the adjusted Rand index which accounts for the fact that random classification would produce some correct class agreements. The ARI takes on a value of zero under random classification, one under perfect classification, and a negative value for classification worse than random, thus making this value easier to interpret.

Chapter 3

Methodology

3.1 Introduction

This thesis introduces a mixtures of skew-t factor analyzers model by imposing a factor structure on the covariance parameter $\boldsymbol{\Sigma}_g$ in the model given in Equation (2.6) such that $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g$. The loading matrix $\boldsymbol{\Lambda}_g$ is allowed to be equal or vary across groups such that $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g$ or $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}_g$. This leads to a fully unconstrained and a constrained model which, following McNicholas and Murphy (2008), will be referred to as UUU and CUU, respectively, herein.

Note that the mathematics in this chapter corresponding to the formulation of the mixtures of generalized hyperbolic distributions model is the same as that given by Browne and McNicholas (2012). The notation used herein will also be the same as that used in the Browne and McNicholas (2012) article.

3.2 Mixtures of Skew-t Factor Analyzers

3.2.1 Skew-t Distribution

The density of the generalized hyperbolic distribution is given by

$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \left[\frac{\chi + \delta(\mathbf{x}, \boldsymbol{\mu}|\boldsymbol{\Sigma})}{\psi + \boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}\right]^{(\lambda - p/2)/2} \times \frac{[\psi/\chi]^{\lambda/2} K_{\lambda - p/2}\left(\sqrt{[\psi + \boldsymbol{\alpha}\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}][\chi + \delta(\mathbf{x}, \boldsymbol{\mu}|\boldsymbol{\Sigma})]}\right)}{(2\pi)^{p/2} \mid \boldsymbol{\Sigma} \mid^{1/2} K_{\lambda}(\sqrt{\chi\psi}) \exp((\boldsymbol{\mu} - \mathbf{x})'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha})}, \tag{3.1}$$

where $\boldsymbol{\vartheta} = (\lambda, \chi, \psi, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$ is the vector of parameters, $\delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma}) = (\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ is the squared Mahalanobis distance between $\mathbf{x}$ and $\boldsymbol{\mu}$, and $K_{\lambda}(\omega)$ is the modified Bessel function of the third kind with index λ where $\lambda \in \mathbb{R}$ and $\sqrt{\chi\psi} > 0$. We will write $\mathbf{X} \sim \mathcal{G}_p(\lambda, \chi, \psi, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$ to denote that a p-dimensional random variable $\mathbf{X}$ follows the generalized hyperbolic distribution.

The density of the skew-t distribution can be obtained as a special case of the generalized hyperbolic distribution by setting $\lambda = -\nu/2$, $\chi = \nu$ and $\psi \to 0$. Therefore, a skew-t random variable $\mathbf{X}$ has density

$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \left[\frac{\nu + \delta(\mathbf{x}, \boldsymbol{\mu}|\boldsymbol{\Sigma})}{\boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}\right]^{\frac{(-\nu - p)}{4}} \times \frac{\nu^{\nu/2} K_{(-\nu - p)/2}\left(\sqrt{[\boldsymbol{\alpha}\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}][\nu + \delta(\mathbf{x}, \boldsymbol{\mu}|\boldsymbol{\Sigma})]}\right)}{(2\pi)^{p/2} \mid \boldsymbol{\Sigma} \mid^{1/2} \Gamma(\frac{\nu}{2}) 2^{\nu/2 - 1} \exp((\boldsymbol{\mu} - \mathbf{x})'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha})}, \tag{3.2}$$

with parameters as in Equation (3.1).

3.2.2 Alternate Forms of the Skew-t Distribution

Many alternative forms of the skew-t distribution have appeared in the literature. For example, see Sahu et al. (2003), Branco and Dey (2001), Jones and Faddy (2003), and Ma and Genton (2004). The form used in this thesis, introduced by Barndorff-Nielsen and Shephard (2001) with density function as given above, was chosen for this work as it has not previously appeared in the mixture modelling literature and has a computationally efficient E-step.

3.2.3 Relationship to the Normal Distribution

A generalized hyperbolic random variable $\mathbf{X}$ can then be generated by combining a random variable $Y \sim GIG(\psi, \chi, \lambda)$ and a multivariate Gaussian random variable $\mathbf{U} \sim N(0, \boldsymbol{\Sigma})$ such that

$$\mathbf{X} = \boldsymbol{\mu} + Y\boldsymbol{\alpha} + \sqrt{Y}\mathbf{U}.$$

Recall that the skew-t distribution is obtained from the generalized hyperbolic distribution by setting $\lambda = -\nu/2$, $\chi = \nu$, and $\psi \to 0$. By substituting these parameters into the GIG density formula, we find that

$$Y \sim \Gamma^{-1}\left(\nu_g/2, \nu_g/2\right),$$

where $\Gamma^{-1}(\cdot)$ is the inverse Gamma distribution. Therefore, we have that a skew-t random variable $\mathbf{X}$ can be obtained through the relationship

$$\mathbf{X} = \boldsymbol{\mu} + Y\boldsymbol{\alpha} + \sqrt{Y}\mathbf{U},$$

where $Y \sim \Gamma^{-1}(\nu_g/2, \nu_g/2)$ and $\mathbf{U} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$.

3.2.4 Distribution of Y Given X

If follows that the distribution of a generalized hyperbolic random variable $\mathbf{X}$ given Y is multivariate normal; i.e., $\mathbf{X} \mid (Y = y) \sim \phi(\mathbf{x} \mid \boldsymbol{\mu} + y\boldsymbol{\alpha}, y\boldsymbol{\Sigma})$, where

$$\phi(\mathbf{x} \mid \boldsymbol{\mu} + y\boldsymbol{\alpha}, y\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^p \mid \boldsymbol{\Sigma} \mid}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\} \tag{3.3}$$

with mean $\boldsymbol{\mu}$ and covariance parameter $\boldsymbol{\Sigma}$. Using Bayes' theorem, we obtain

$f(y \mid \mathbf{x}) = \frac{f(\mathbf{x}|y)h(y)}{f(\mathbf{x})} =$

$$\left[\frac{\psi + \boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}{\chi + \delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma})}\right]^{(\lambda - p/2)/2} \frac{y^{\lambda + p/2 - 1}\exp\{-[y(\psi + \boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}) + (\chi + \delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma}))/y]/2\}}{2K_{\lambda - p/2}\left(\sqrt{[\psi + \boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}][\chi + \delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma})]}\right)}$$

and thus we have $Y \mid (\mathbf{X} = \mathbf{x}) \sim \text{GIG}(\psi + \boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}, \chi + \delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma}), \lambda - p/2)$. Therefore, for a skew-$t$ random variable $\mathbf{X}$ we have that

$f(y \mid \mathbf{x}) = \frac{f(\mathbf{x}|y)h(y)}{f(\mathbf{x})} =$

$$\left[\frac{\boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}{\nu + \delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma})}\right]^{-(\nu + p)/4} \frac{y^{-\nu/2 + p/2 - 1}\exp\{-[y(\boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}) + (\nu + \delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma}))/y]/2\}}{2K_{-\nu/2 - p/2}\left(\sqrt{[\boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}][\nu + \delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma})]}\right)}$$

and $Y \mid (\mathbf{X} = \mathbf{x}) \sim \text{GIG}(\boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}, \nu + \delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma}), \nu/2 - p/2)$.

3.2.5 Generalized Inverse Gaussian Distribution

The GIG distribution has the density function

$$p(y \mid \psi, \chi, \lambda) = \frac{(\psi/\chi)^{\lambda/2} y^{\lambda - 1}}{2K_\lambda(\sqrt{\psi\chi})} \exp\left\{-\frac{\psi y + \chi/y}{2}\right\},$$

for $y > 0$ with $\psi, \chi \in \mathbb{R}$ and where K_λ is the modified Bessel function of the third kind with index λ. This distribution has the following expected values:

$$\mathbb{E}[Y] = \sqrt{\frac{\chi}{\psi}} \frac{K_{\lambda+1}(\sqrt{\psi\chi})}{K_\lambda(\sqrt{\psi\chi})},$$

$$\mathbb{E}[1/Y] = \sqrt{\frac{\psi}{\chi}}\frac{K_{\lambda+1}(\sqrt{\psi\chi})}{K_\lambda(\sqrt{\psi\chi})} - \frac{2\lambda}{\chi},$$

and

$$\mathbb{E}[\log(Y)] = \log\sqrt{\frac{\chi}{\psi}} + \frac{1}{K_\lambda(\sqrt{\psi\chi})}\frac{\delta}{\delta\lambda}K_\lambda(\sqrt{\psi\chi}).$$

The complete-data log-likelihood for this model is given by

$$l_c(\boldsymbol{\vartheta} \mid \mathbf{x}, \mathbf{y}, \mathbf{z}) = \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\Big[\log\pi_g + \log\phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g + y_i\boldsymbol{\alpha}_g, y_i\boldsymbol{\Sigma}_g) + \log h(y_i \mid \omega_g, \lambda_g)\Big], \quad (3.4)$$

where z_{ig}=1 if observation i is a member of group g and zero otherwise.

3.2.6 The AECM Algorithm for Skew-t-Factor Analyzers

E-step

On each E-step, the group membership labels, z_{ig}, are updated by

$$\mathbb{E}[Z_{ig}|\mathbf{x}_i] = \frac{\pi_g f(\mathbf{x}_i|\boldsymbol{\theta}_g)}{\sum_{h=1}^{G}\pi_h f(\mathbf{x}_i|\boldsymbol{\theta}_h)} =: \hat{z}_{ig}.$$

Following the notation of Browne and McNicholas (2012), we make use of the expectations

$$\mathbb{E}[Y_i|\mathbf{x}_i, Z_{ig}=1] = \sqrt{\frac{\chi}{\psi}}\frac{K_{\lambda+1}(\sqrt{\psi\chi})}{K_\lambda(\sqrt{\psi\chi})} =: a_{ig},$$

$$\mathbb{E}[1/Y_i|\mathbf{x}_i, Z_{ig}=1] = \sqrt{\frac{\psi}{\chi}}\frac{K_{\lambda+1}(\sqrt{\psi\chi})}{K_\lambda(\sqrt{\psi\chi})} - \frac{2\lambda}{\chi} =: b_{ig},$$

and

$$\mathbb{E}[\log(Y_i)|\mathbf{x}_i, Z_{ig}=1] = \log\sqrt{\frac{\chi}{\psi}} + \frac{1}{K_\lambda(\sqrt{\psi\chi})}\frac{\delta}{\delta\lambda}K_\lambda(\sqrt{\psi\chi}) =: c_{ig}.$$

CM Step 1

Following Browne and McNicholas (2012), we will use the notation

$$n_g = \textstyle\sum_{i=1}^{n} z_{ig},\ A_g = (1/n_g)\sum_{i=1}^{n} z_{ig}a_{ig}, \text{ and } B_g = (1/n_g)\sum_{i=1}^{n} z_{ig}b_{ig}.$$

On the first CM step of the AECM algorithm, the missing data is taken to be the latent variables Y_i and the group membership labels z_{ig}. The mixing proportions, π_g, the component means, $\boldsymbol{\mu}_g$, and the skewness parameters, $\boldsymbol{\alpha}_g$, are updated by

$$\hat{\pi}_g = \frac{\sum_{i=1}^{n} \hat{z}_{ig}}{n}, \qquad \hat{\boldsymbol{\mu}}_g = \frac{\sum_{i=1}^{n} \mathbf{x}_i(A_g b_{ig} - 1)}{\sum_{i=1}^{n} A_g b_{ig} - n},$$

and

$$\hat{\boldsymbol{\alpha}}_g = \frac{\sum_{i=1}^{n} \mathbf{x}_i(b_{ig} - B_g)}{\sum_{i=1}^{n} A_g b_{ig} - n},$$

respectively.

The update for the degrees of freedom, ν_g, does not exist in closed form. We compute the update iteratively by setting

$$\log\left(\frac{\hat{\nu}_g^{\text{new}}}{2}\right) + 1 - \varphi\left(\frac{\hat{\nu}_g^{\text{new}}}{2}\right) - \sum_{i=1}^{n}\left(z_{ig}\log(a_{ig}) + \frac{1}{a_{ig}}\right)$$

equal to zero and solving for ν_g^{new}. We utilize the `uniroot` function available through the `stats` package in R to numerically solve for this update. The range of values to be searched for the updated degrees of freedom value $\hat{\nu}_g$ was restricted to the interval $[0.5, 200]$. A lower bound of $\hat{\nu}_g = 0.5$ is sufficient to capture an underlying skew-t distribution, while at the upper end of this interval, a maximum value of $\hat{\nu}_g = 200$ will allow us to fit an underlying Gaussian distribution.

CM Step 2

On the second CM step, we update the factor loadings, $\hat{\boldsymbol{\Lambda}}_g$, and the diagonal error variance matrix, $\hat{\boldsymbol{\Psi}}_g$. Here the missing data is taken to be the group membership labels and the latent variable $\mathbf{U}$ from the factor analysis model. Following McNicholas and Murphy (2008) we will make use of the notation $\boldsymbol{\beta}_g = \hat{\boldsymbol{\Lambda}}_g'(\hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\Lambda}}_g' + \hat{\boldsymbol{\Psi}}_g)^{-1}$ and $\boldsymbol{\Theta}_g = \mathbf{I}_p - \hat{\boldsymbol{\beta}}_g\hat{\boldsymbol{\Lambda}}_g + \hat{\boldsymbol{\beta}}_g\mathbf{S}_g\hat{\boldsymbol{\beta}}_g'$. Therefore, our updates are given by

$$\hat{\boldsymbol{\Lambda}}_g = \mathbf{S}_g\hat{\boldsymbol{\beta}}_g'\hat{\boldsymbol{\Theta}}_g^{-1} \text{ and } \hat{\boldsymbol{\Psi}}_g = \text{diag}\{\boldsymbol{\Sigma}_g - \hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\beta}}_g\mathbf{S}_g\},$$

where

$$\hat{\mathbf{S}}_g = \frac{1}{n_g}\sum_{i=1}^{n} z_{ig}b_{ig}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)' - \hat{\boldsymbol{\alpha}}_g(\bar{\mathbf{x}}_g - \hat{\boldsymbol{\mu}}_g)' - (\bar{\mathbf{x}}_g - \hat{\boldsymbol{\mu}}_g)(\hat{\boldsymbol{\alpha}}_g)' + A_g\hat{\boldsymbol{\alpha}}_g(\hat{\boldsymbol{\alpha}}_g)',$$

and $\bar{\mathbf{x}}_g = (1/n_g)\sum_{i=1}^{n} z_{ig}\mathbf{x}_i$.

3.2.7 Constraining the Factor Loading Matrix

In the case that the factor loading matrix $\boldsymbol{\Lambda}$ is constrained to be equal across groups, McNicholas and Murphy (2008) give the following updates:

$$\hat{\boldsymbol{\beta}}_g = \hat{\boldsymbol{\Lambda}}'(\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}' + \hat{\boldsymbol{\Psi}}_g)^{-1}, \qquad \hat{\lambda}_i = \mathbf{r}_i\left(\sum_{g=1}^{G}\frac{n_g}{\hat{\psi}_{g(i)}}\boldsymbol{\theta}_g\right)^{-1},$$

and

$$\hat{\boldsymbol{\Psi}}_g = \text{diag}(\mathbf{S}_g - 2\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\beta}}_g\mathbf{S}_g + \hat{\boldsymbol{\Lambda}}\boldsymbol{\theta}_g(\hat{\boldsymbol{\Lambda}})')$$

where $\hat{\lambda}_i$ is the i^{th} row of $\hat{\boldsymbol{\Lambda}}$, $\hat{\boldsymbol{\Psi}}_{g(i)}$ is the i^{th} element in the diagonal of $\hat{\boldsymbol{\Psi}}_g$ and $\mathbf{r}_i$ is the i^{th} row of the matrix $\sum_{g=1}^{G} n_g\hat{\boldsymbol{\Psi}}_g^{-1}\mathbf{S}_g\hat{\boldsymbol{\beta}}_g'$ and $i = 1, ..., p$. This model is slow to fit since the matrix $\hat{\boldsymbol{\Lambda}}$ must be solved for row-by-row (cf. McNicholas and Murphy, 2008).

Table 3.1 gives the number of free covariance parameters in the mixtures of skew-t factor analyzers models. We will see both models applied to microarray gene expression data in the following chapter.

Table 3.1: The covariance structures and associated free covariance parameters of the skew-t factor analyzer models

Model	Loading Matrix	Error Variance	Free covariance parameters
CUU	Constrained	Unconstrained	$[pq - q(q-1)/2] + Gp$
UUU	Unconstrained	Unconstrained	$G[pq - q(q-1)/2] + Gp$

Chapter 4

Results

4.1 Introduction

In this chapter, the skew-t factor analyzer models developed in Chapter 3 will be applied to both real and simulated data sets. Simulated data following skew-t, skew-normal, multivariate-t, and Gaussian distributions will be used to demonstrate the ability of the models to capture each of these distribution types. Two real data sets of low-dimension are tested to illustrate the ability of the skew-t mixture models to fit skewed data which is not well modelled by `mclust`. Finally, the mixtures of skew-t factor analyzer models were run on two high-dimensional data sets consisting of gene expression data. These results are compared to the results from the equivalent `pgmm` models.

4.2 Simulated Data

4.2.1 The Data

The skew-t mixture model given in Equation (2.6) was run on data simulated from four different statistical distributions. 100 observations were randomly sampled from the skew-t model developed in the previous chapter. Additionally, 100 observations were randomly sampled from each of the multivariate Gaussian, multivariate-t, and multivariate skew-normal distributions. These samples were generated using the `mvrnorm` function in the `MASS` package, the `rmvt` function in the `mvtnorm` package, and the the `rmsn` function in the package `sn` in R, respectively.

4.2.2 Skew-t Mixture Models

The skew-t mixture model was fit to the data for $G = 1, ..., 5$. The BIC values for each of these models are given in Table 4.1. We see that the best model in terms of BIC was the 4-component model which obtained perfect clustering results. Table 4.2 gives parameter estimates obtained from this model. The contour plot in Figure 4.1 shows that skew-t mixture model is well fit to the data and is able to capture all four underlying distributions.

4.3 Old Faithful Geyser Data

4.3.1 The Data

This data set, collected from data on the Old Faithful Geyser in Yellowstone National Park, Wyoming, USA, contains 2 measured variables for 272 observations where the variables are

Table 4.1: BIC and ARI values for the skew-t mixture model fit to the simulated data for G=1,...,5.

G	BIC	ARI
1	-14288.26	0
2	-11311.83	0.4990596
3	-10338.21	0.7135174
4	$\mathbf{-8683.81}$	**1**
5	-8745.26	0.9271678

Table 4.2: Parameter estimates obtained from the four-component skew-t mixture model fit to the simulated data.

Cluster	$\hat{\nu}$	$\hat{\alpha}$
1	27.19	$(12.72, 35.44)$
2	17.00	$(4.98, 1.78)$
3	11.04	$(4.62, 6.94)$
4	83.56	$(10.71, -9.92)$

duration of eruptions in minutes and waiting time to the next eruption in minutes. Although there are no true group memberships, the data appears to comprise two skewed groups and thus this data set, available as `faithful` in the datasets library in R, is commonly used as an example of skewed data (Azzalini and Bowman, 1990).

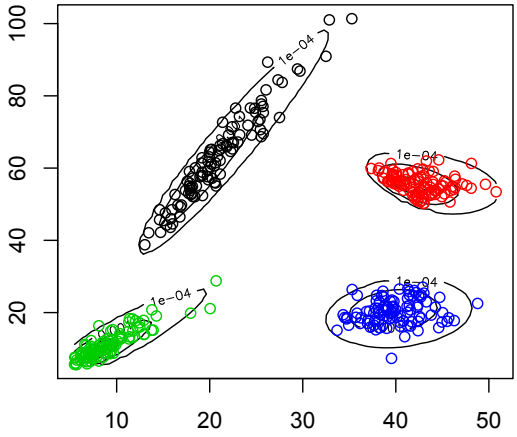

Figure 4.1: *Contour plot of the four-component skew-t mixture model fit to the simulated data.*

4.3.2 Skew-t Mixture Models

The skew-t mixture model given in Equation (2.6) was fit to the data for G=1,...,5 and the model with the highest BIC value was selected. The BIC values for each of the five models are contained in Table 4.3. The best model in terms of the BIC was the two-component model. Figure 4.2 gives the contour plot of the two-component model fit to the Old Faithful data.

4.3.3 MCLUST

For comparison, a Gaussian mixture model was fit to the data using the `mclust` package in R. The best model selected by `mclust` was the 3-component EEE model with a BIC value of -2314.386. The contour plot in Figure 4.3 shows both the MAP classification and density

Table 4.3: BIC values for the skew-t mixture model fit to the Old Faithful data for G=1,...,5.

G	BIC
1	-2593.43
2	$\mathbf{-2311.77}$
3	-2344.97
4	-2382.55
5	-2418.17

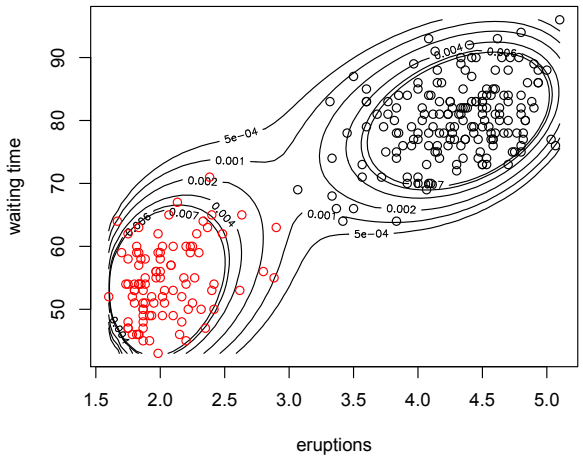

Figure 4.2: *Contour plot of the two-component skew-t mixture model fit to the Old Faithful data.*

estimate of the data.

4.3.4 Discussion

While the Old Faithful Geyser data does not contain any true groupings, we can see that the skew-t mixture model is able to both capture the bimodal nature of the data as well as

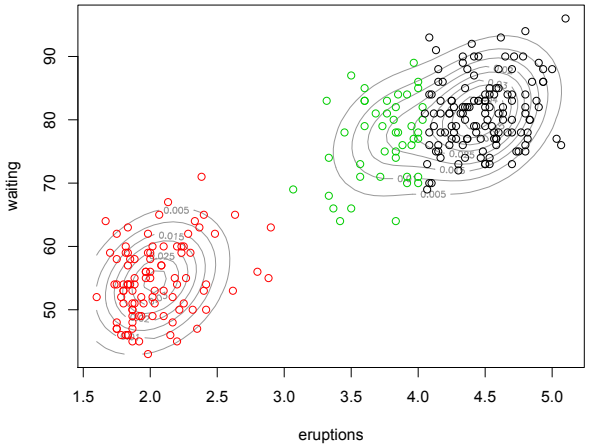

Figure 4.3: *Contour plot of the EEE* `mclust` *model fit to the Old Faithful Data for G=3.*

obtain a close fit. In contrast, the elliptical clusters fit by `mclust` are unable to capture the underlying skewness of the data, resulting in a model with a higher number of components being selected as the best model.

4.4 Australian Institute of Sport Data

4.4.1 The Data

This data set is made up of thirteen measured variables on 102 male and 100 female athletes collected at the Australian Institute of Sport (AIS). We will consider the body fat percentages and body mass index (BMI) for each athlete, taking their gender to be unknown. This data set, previously analzyed by Lin (2012), Lee and McLachlan (2011), and Vrbik and McNicholas (2012), is widely considered to exemplify skewed data.

4.4.2 Skew-t Mixture Models

As with the Old Faithful Geyser data, the skew-t mixture model was applied to the AIS data for G =1,...,5. The BIC values for all five models are contained in Table 4.4.

Table 4.4: BIC values for the skew-t mixture model fit to the AIS data for G=1,...,5.

G	BIC
1	-2286.22
2	$\mathbf{-2239.08}$
3	-2255.34
4	-2289.72
5	-2331.53

We can see that the two-component model resulted in the highest BIC value of -2239.077. A cross tabulation of the gender of the athletes versus the clustering results obtained using this model is reported in Table 4.5. This particular classification, with all but 9 of the observations correctly classified, results in an error rate of 0.045. The contour plot showing the density estimate and MAP classifications of the two-component model fitted to the AIS data can be seen in Figure 4.4.

4.4.3 MCLUST

The `mclust` models were also run on the AIS data for $G = 1, ..., 5$. The best model according to the BIC was the three-component VVV model. Figure 4.5 shows the contour plot of

Table 4.5: Classification table for the 2-component skew-t mixture model fit to the AIS data.

	1	2
Male	94	8
Female	1	99

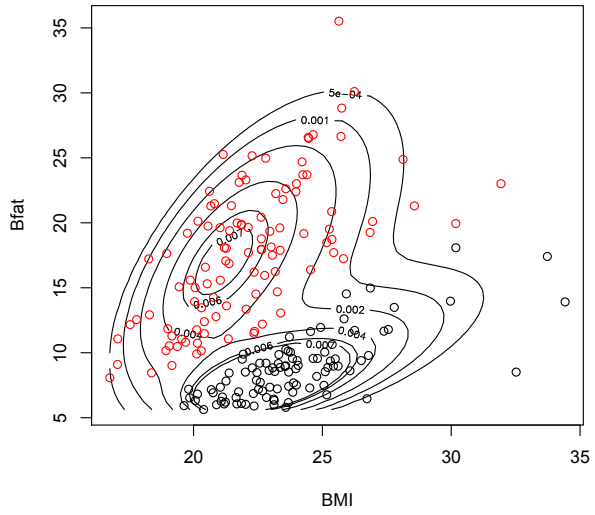

Figure 4.4: *Contour plot of the two-component skew-t mixture model fit to the AIS data.*

this model fit to the AIS data. The classification obtained from this model, which results in an ARI of 0.677, is given in Table 4.6.

4.4.4 Discussion

Similarly to the Old Faithful Geyser data, we see that applying `mclust` to the AIS data results in an overestimate of the number of true groups. It appears that `mclust` is grouping

Table 4.6: Classification table for the 3-component VVV `mclust` model fit to the AIS data.

	1	2	3
Female	90	9	1
Male	4	17	81

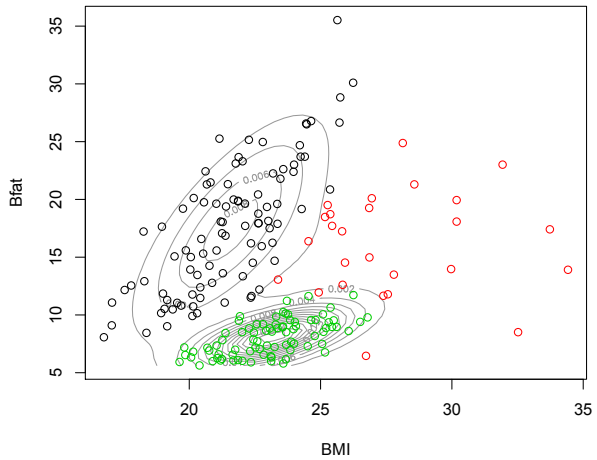

Figure 4.5: *Contour plot of the three-component VVV* `mclust` *model fit to the AIS data.*

many of the outlying observations together into a third component. However, the skew-t mixture model is able to capture the skewed nature of the data, resulting in a low misclassification rate.

4.5 Colon Data

4.5.1 The Data

Alon et al. (1999) introduce a set of gene expression data from 62 colon tissue samples, 40 of which are tumor tissue samples and 22 being from normal tissue. This data, which contains expression values for 6500 genes for each sample, was obtained using Affymetrix microarray technology. Following McNicholas and Murphy (2010), we will analyze a reduced version of this data set containing expression values for only 461 genes. This reduced data was sourced from `www.uoguelph.ca/~ pmcnicho/publications.html`.

4.5.2 Skew-t Factor Models

The PGMM package allows the option of starting from random starting values, k-means starting values, or user-specified starting values. For the most direct comparison between the `pgmm` and skew-t factor models, k-means clustering results were used to initialize the matrix of group membership labels $\hat{z}_{ig}$ each time a model was run. The component means $\boldsymbol{\mu}_g$ and component covariance matrices $\boldsymbol{\Sigma}_g$ were then initialized using the `cov.wt` function in R. We chose to initialize the skewness parameters $\boldsymbol{\alpha}_g$ as a p-dimensional vector with each entry equal to 0.1. The degrees of freedom were initialized as $\nu_g = 50$. Following McNicholas and Murphy (2008), the elements of the factor loading matrices $\boldsymbol{\Lambda}_g$ were initialized as $\lambda_{ij} = \sqrt{d_j}\rho_{ij}$ where d_j is the j^{th} largest eigenvalue of the component covariance matrix and ρ_{ij} is the i^{th} element of the eigenvector which corresponds to the j^{th} largest eigenvalue of $\boldsymbol{\Sigma}_g$ with $i \in 1, ..., p$ and $j \in 1, ..., q$. The noise matrix $\boldsymbol{\Psi}_g$ was initialized as $\boldsymbol{\Psi}_g = \text{diag}\{\boldsymbol{\Sigma}_g - \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g'\}$.

The EM algorithm is highly dependent on starting values to avoid convergence to a local rather than global maxima of the likelihood surface. For this reason, when fitting a mixture of skew-t factor analyzers to the colon data and the leukaemia data in the following section, we make use of a deterministic annealing approach (Ueda and Nakano, 1998; Zhou and Lange, 2010) which flattens the likelihood surface, slowly returning it to its original shape over several iterations of the algorithm to improve the chances of converging to the dominant mode. For the first 25 iterations of the AECM algorithm, the update for the expected value of the group membership labels $\hat{z}_{ig}$ is of the form

$$\mathbb{E}[Z_{ig}|\mathbf{x}_i] = \frac{(\pi_g f(\mathbf{x}_i|\boldsymbol{\theta}_g))^d}{\sum_{h=1}^{G}(\pi_h f(\mathbf{x}_i|\boldsymbol{\theta}_h))^d}.$$

The value of d is gradually increased at each iteration following a sequence of values from d=1e-7 to 1.

Following McNicholas and Murphy (2010), we ran the mixtures of skew-t factor analyzer models on the colon data for $G = 2$ and $q = 1, ..., 10$. This was treated as a true clustering problem in which the true tissue types were taken to be unknown. After running the models five times from k-means starting values, the best model overall in terms of BIC was the UUU model with $q = 7$ latent factors. This model produced an ARI value of 0.4502882. Table 4.7 gives the MAP classifications resulting from the application of this model. Here we see that 10 tissues were misclassified.

Table 4.7: Classification table for the UUU skew-t factor analyzer model with q=7 latent factors fit to the colon data.

	1	2
Tumour	32	8
Normal	2	20

4.5.3 Parsimonious Gaussian Mixture Models

The UUU model and the CUU model from the `pgmm` package were run on the colon data for $G = 2$ and $q = 1, ..., 10$ latent factors five times from k-means starting values. The best model in terms of BIC over the five runs was the CUU model with $q = 8$ latent factors. This model has a BIC value of -71094.67 and an ARI value of -0.016. The MAP classifications obtained using this model are given in Table 4.8. We can see that 29 tissue samples were misclassified.

Table 4.8: Classification table for the 2-component CUU `pgmm` model with q=8 latent factors fit to the colon data.

	1	2
Tumour	17	23
Normal	6	16

4.5.4 Discussion

We see that the skew-t factor analyzer models outperformed the CUU and UUU `pgmm` models on the colon data. One might argue that 5 runs is too few to capture the best

model. However, when the colon data was tested by McNicholas and Murphy (2010), 10 random starts were used and of the UUU and CUU models, the CUU model with 8 factors was still the best model in terms of BIC. Therefore, more runs would likely only lead to a further improvement in the classification results obtained from the skew-t model compared to `pgmm`.

4.6 Leukaemia Data

4.6.1 The Data

Golub et al. (1999) introduce data comprising gene expression values from 47 acute lymphoblastic leukaemia (ALL) tissue samples and 25 acute myeloid leukaemia (AML) tissue samples. This data was gathered using Affymetrix arrays and contains expression values for 7129 genes for all 72 samples. McNicholas and Murphy (2010) reduce the number of genes in this data set to 2030 and a subset of 400 of the remaining genes were used in this work. As with the colon data, this data was sourced from `www.uoguelph.ca/~pmcnicho/publications.html`.

4.6.2 Skew-t Factor Models

The mixtures of skew-t factor analyzers models were run on the leukaemia data for $G = 2$ and $q = 1, ..., 6$ latent factors. The algorithm was initialized as for the colon data. The best model in terms of BIC was the UUU model with 6 latent factors which obtained an ARI value of 0.786. Table 4.9 gives the classification results obtained from this model. Here we see that four tissue samples were misclassified.

Table 4.9: Classification table for the 2-component UUU mixtures of skew-t factors model with q=6 latent factors fit to the leukaemia data.

	1	2
ALL	44	1
AML	3	24

4.6.3 Parsimonious Gaussian Mixture Models

The UUU and CUU models from the `pgmm` package were also run on the leukaemia data for $G = 2$ and $q = 1, ..., 6$ latent factors. The best model in terms of BIC was the CUU model with one latent factor. This model obtained an ARI value of 0.786. Table 4.9 gives the classification results from this model which were the same as those obtained from the best mixtures of skew-t factors model.

Table 4.10: Classification table for the 2-component CUU `pgmm` model with q=1 latent factors fit to the leukaemia data.

	1	2
ALL	44	1
AML	3	24

4.6.4 Discussion

For the leukaemia data, we attained the same classification results from the best model for both the mixtures of skew-t factor models and the `pgmm` models. We can expect that given data which is relatively normally distributed, we will get comparable results from both sets

of models since we have shown that the skew-t models are able to capture an underlying Gaussian distribution. The advantage in using these models comes from applying them to data that is skewed or contains outlying observations. However, when the mixtures of skew-t factor models were applied to the leukaemia data, the estimates of the degrees of freedom for the two groups were ν_1=13.1 and ν_2=4.8. While there was not clear evidence of skewness based on our parameter estimates, these results do suggest outlying data and thus we would have expected superior performance from the mixtures of skew-t factor models.

This may indicate that the two groups are well separated, making it easier for `pgmm` to model the data. Alternatively, this may indicate that the parsimonious Gaussian mixture models are more flexible in capturing groups with outlying data than we might expect considering the component densities are multivariate Gaussian. Further evidence of this is given by the fact that when the `pgmm` models are fit to the Old Faithful data, a two-component model is selected as the best model in terms of BIC. However, when these models are fit to the AIS data, a three-component model is selected, and given the superior results of the skew-t factor models on the colon data, we suggest that it is indeed advantageous to fit the mixtures of skew-t factor models given a true clustering problem in which the underlying distribution type is unknown.

Chapter 5

Conclusion

5.1 Summary

This thesis introduces a family of skew-t mixture models with a factor analysis covariance structure. Allowing the factor loading matrix to be equal or vary across groups leads to two mixtures of skew-t factor analyzer models. Mixtures of factor analyzer models have previously been introduced which take the component densities to be multivariate Gaussian or multivariate-t. However, the mixture models developed in this thesis are the first to make use of the skew-t distribution in the context of factor analysis. Furthermore, this form of the skew-t distribution, which stems from the generalized hyperbolic distribution, has not previously appeared within the mixture-modelling literature.

After the models are developed, we apply them to both real and simulated data. The results are compared to `mclust` and `pgmm` where appropriate. We find that these models prove to be equivalent or superior to the established clustering methods when applied to these data

sets.

5.2 Discussion

The results from Chapter 4 show that the skew-t factor models outperformed the equivalent `pgmm` models when run on the colon data with identical classification results when applied to the leukaemia data. However, when McNicholas and Murphy (2010) ran the 12 models in the EPGMM family on these data sets, the UUU and CUU equivalent models did not perform best overall. For example, for the colon data, the best EPGMM model in terms of BIC was the CCUC model with 6 latent factors. This model, which is equivalent to the CUC model in the PGMM family, obtained an ARI value of 0.697. This suggests that while our models were able to outperform or match their PGMM counterparts, neither the UUU or CUU models may be the best fit for this data.

When running the mixtures of skew-t factor analyzers models on both the colon data and leukaemia data, it was noted that the model which obtained the highest BIC value frequently did not result in the highest ARI value. Alternative selection criterions could be tested, such as the integrated completed likelihood (ICL) criterion, which may prove to be more suitable.

5.3 Future Work

This thesis introduces two mixtures of skew-t factor analyzers models which are equivalent to two of the eight parsimonious Gaussian mixture models. Future research will include

extending the skew-t factor models to include constraints on the error variance matrix including the isotropic constraint. This would result in a family of eight models equivalent to the PGMM family. We can also consider constraining both the degrees of freedom and skewness parameters to be equal across groups. Therefore, the family of skew-t factor analyzer models can be extended to include as many as 32 models.

Furthermore, an eigen-decomposed covariance structure will be imposed on the covariance matrix $\boldsymbol{\Sigma}_g$ in the skew-t mixture model given in Equation (2.6) such that $\boldsymbol{\Sigma}_g = \boldsymbol{\lambda}_g \mathbf{D}_g \mathbf{A}_g \mathbf{D}_g'$. Allowing constraints on the components of the eigen-decomposed covariance matrix, we will introduce a family of skew-t mixture models which will be suitable for the clustering of low-dimensional data. These models could be considered to be a skew-t equivalent of the `mclust` models. A comparison of sorts may then be conducted to give insight into the performance of these skew-t models as compared to similar models which use another variant of this distribution.

The results from this thesis suggest that the full family of skew-t factor analyzer models which remain to be developed will prove to be a useful addition to the current collection of clustering methods. The ability of these models to cluster high-dimensional data makes them especially useful to the rapidly developing field of bioinformatics, while their flexibility in capturing several underlying distributions makes them particularly advantageous when applied to true clustering problems.

Bibliography

Aitken, A. (1926). On Bernoulli's numerical solution of algebraic equations. *Proceedings of the Royal Society of Edinburgh 46*, 289–305. 9

Alon, U., N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America 96*, 6745–6750. 34

Andrews, J. and P. McNicholas (2011). Mixture of modified t-factor analyzers for model-based clustering, classification, and discriminant analysis. *Journal of Statistical Planning and Inference 141*(4), 1479–1486. 12

Azzalini, A. and A. Bowman (1990). A look at some data on the Old Faithful geyser. *Applied Statistics 39*, 357–365. 27

Banfield, J. and A. E. Raftery (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics 49*(3), 803–821. 6

Barndorff-Nielsen, O. and N. Shephard (2001). Non-Gaussian Ornstein-Uhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society B 63*, 167–241. 2, 19

Biernacki, C., G. Celeux, and G. Govaert (2000). Assessing a mixture model for clustering with the integrated complete likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*(7), 719–725. 5

Böhning, D., E. Dietz, R. Schaub, P. Schlattmann, and B. Lindsay (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics 46*, 373–388. 9

Branco, M. and D. Dey (2001). A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis 79*, 99–113. 19

Browne, R. and P. McNicholas (2012). A mixture of generalized hyperbolic distributions. Unpublished manuscript. 6, 17, 21

Browne, R., P. McNicholas, and M. Sparling (2012). Model-based learning using a mixture of mixtures of Gaussian and uniform distributions. *IEEE Transactions of Pattern Analysis and Machine Intelligence 34*(4), 814–817. 7

Campbell, J., C. Fraley, F. Murtagh, and A. Raftery (1997). Linear flaw detection in woven textiles using model-based clustering. *Pattern Recognition Letters 18*, 1539–1548. 15

Celeux, G. and G. Govaert (1995). Gaussian parsimonious clustering models. *Pattern recognition 28*, 781–793. 6, 10

Dasgupta, A. and A. Raftery (1998). Detecting features in spatial point processed with clutter via model-based clustering. *Journal of the American Statistical Association 93*, 294–302. 15

Day, N. (1969). Estimating the components of a mixture of two normal distributions. *Biometrika 56*, 463–474. 7

Dempster, A. P., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B 39*(1), 1–38. 5, 7

Everitt, B. and D. Hand (1981). *Finite Mixture Distributions*. London: Chapman and Hall. 5

Fraley, C. and A. E. Raftery (1999). MCLUST: software for model-based cluster analysis. *Journal of Classification 16*, 297–306. 10

Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association 97*(458), 611–631. 5, 6

Ghahramani, Z. and G. Hinton (1997). The EM algorithm for factor analyzers. Technical Report CRG-TR-96-1, University of Toronto, Toronto. 10

Golub, T., D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science 286*, 531–537. 37

Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of Classification 2*, 193–218. 15

Jones, M. and M. Faddy (2003). A skew extension of the t-distribution, with applications. *Journal of the Royal Statistical Society B 65*, 159–174. 19

Karlis, D. and L. Meligkotsidou (2007). Finite mixtures of multivariate Poisson distributions with application. *Journal of Statistical Planning and Inference 137*(6), 1942–1960. 6

Lee, S. and G. McLachlan (2011). On the fitting of mixtures of multivariate skew t-distributions via the EM algorithm. 6, 30

Leroux, B. (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics 20*, 1350–1360. 15

Lin, T.-I. (2009). Maximum likelihood estimation for multivariate skew normal mixture models. *Journal of Multivariate Analysis 100*, 257–265. 6

Lin, T.-I. (2012). Robust mixture modelling using multivariate skew t distributions. *Statistics and Computing 20*, 343–356. 6, 30

Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry, and Applications* (vol. 5 ed.). California: Institute of Mathematical Statistics: Hayward: in NSF-CBMS Regional Conference Series in Probability and Statistics. 9

Ma, Y. and M. Genton (2004). A flexible class of skew-symmetric distributions. *Scandinavian Journal of Statistics 31*, 459–468. 19

Marriott, F. (1974). *Interpretation of multiple observations.* London, United Kingdom: Academic Press. 5

McLachlan, G. and K. Basford (1988). *Mixture Models: Inference and Applications to Clustering*. New York, USA: Marcel Dekker. 5

McLachlan, G., R. Bean, and L. Ben-Tovim Jones (2007). Extension of the mixture of factor analyzers model to incorporate the multivariate t-distribution. *Computational Statistics & Data Analysis 51*(11), 5327–5338. 12

McLachlan, G., R. Bean, and D. Peel (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics 18*(3), 413–422. 6

McLachlan, G. and T. Krishnan (2008). *The EM Algorithm and Extensions* (Second ed.). New York: Wiley. 8

McLachlan, G. and D. Peel (2000a). *Finite Mixture Models* (8th ed.). New York, USA: John Wiley & Sons. 5

McLachlan, G. and D. Peel (2000b). Mixtures of factor analyzers. In *In Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 599–606. 12

McLachlan, G. J. and D. Peel (1998). Robust cluster analysis via mixtures of multivariate t-distributions. In *Lecture Notes in Computer Science*, pp. 658–666. Springer-Verlag. 6

McNicholas, P. (2011). On model-based clustering, classification, and discriminant analysis. *Journal of the Iranian Statistical Society 10*(2), 181–199. 5

McNicholas, P. D. and T. B. Murphy (2008). Parsimonious Gaussian mixture models. *Statistics and Computing 18*, 285–296. 6, 13, 17, 22, 23, 34

McNicholas, P. D. and T. B. Murphy (2010). Model-based clustering of microarray expression data via latent Gaussian mixture models. *Bioinformatics 26*(21), 2705–2712. 6, 34, 35, 37, 41

Meng, X.-L. and D. Rubin (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika 80*, 267–278. 8

Meng, X.-L. and D. van Dyk (1997). The EM algorithm- an old folk song sung to a fast new tune (with discussion). *Journal of the Royal Statistical Society Series B 59*, 511–567. 8

Pearson, K. (1894). Contributions to the theory of mathematical evolution. *Philosophical Transactions of the Royal Society of London A 185*, 71–110. 4

Peel, D. and G. J. McLachlan (2000). Robust mixture modelling using the t distribution. *Statistics and Computing 10*(4), 339–348. 6

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. 10

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association 66*, 846–850. 15

Sahu, S., D. Dey, and M. Branco (2003). A new class of multivariate skew distributions with application to Bayesian regression models. *Canadian Journal of Statistics 31*, 129–150. 19

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*, 461–464. 5

Spearman, C. (1904). The proof and measurement of association between two things. *Journal of Statistical Planning and Inference 15*, 72–101. 10

Tipping, T. and C. Bishop (1999). Mixtures of probabilistic component analyzers. *Neural Computation 11*(2), 443–482. 12

Titterington, D., A. Smith, and U. Makov (1985). *Statistical Analysis of Finite Mixture Distributions*. New York, USA: John Wiley & Sons. 5

Ueda, N. and R. Nakano (1998). Deterministic annealing EM algorithm. *Neural Networks 11*, 271–282. 35

Vrbik, I. and P. McNicholas (2012). Analytic calculations for the EM algorithm for multivariate skew-mixture models. *Statistics and Probability Letters 82*, 1169–1174. 6, 30

Wolfe, J. (1965). A computer program for the computation of maximum likelihood analysis of types. *Communications in Statistics-Theory and Methods 25*, 1799–1824. 7

Woodbury, M. (1950). Inverting modified matrices. Technical Report 42, Princeton University, Princeton, N.J. 12

Zhou, H. and K. Lange (2010). On the bumpy road to the dominant mode. *Scandinavian Journal of Statistics 37*, 612–631. 35