Unsupervised Classification for Skewed and Mixed-Type Data

UNSUPERVISED CLASSIFICATION FOR SKEWED AND MIXED-TYPE DATA

BY

EMAN MOHAMMED S. ALAMER, M.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

DOCTOR OF PHILOSOPHY

Doctor of Philosophy (2022) McMaster University

(Mathematics & Statistics) Hamilton, Ontario, Canada

TITLE:	Unsupervised Classification for Skewed and Mixed-Type Data
AUTHOR:	Eman Mohammed S. Alamer M.Sc. (Statistics) McMaster University, Hamilton, Canada
SUPERVISOR:	Dr. Paul D. McNicholas
NUMBER OF PAGES:	xv, 104

To my parents, Radheya and Mohammed

Abstract

Clustering, also known as unsupervised classification, is a foundational machine learning technique and is used to find underlying group structures in data. There are many well-established model-based techniques to analyze either categorical or continuous data in the clustering paradigm. However, there is a relative paucity of work for mixed-type data, especially mixed data where the continuous variables exhibit skewness and heavy tails. In this thesis, different methodologies and models are presented for analyzing asymmetric and mixed-typed data. The first method is a mixture model for analyzing asymmetric mixed-type data. The second is modelling contaminated mixed-type data and identifying potential outliers. Lastly, model averaging techniques are developed for skewed-data based on Occam's window and parsimonious mixture models. The expectation-maximization algorithm is used here to estimate the model parameters. Both real and simulated data are used for illustration.

Acknowledgements

I would like to acknowledge and express my sincerest thanks to my supervisor, Dr. Paul McNicholas, for his support and guidance.

A special thank you for Dr. Michael Gallaugher for his collaboration.

I would also like to acknowledge the fund provided by King Salman AL Saud Foreign Scholarship program and the Saudi Culture Bureau in Canada.

I would like to thank my parents, Radheya and Mohammed, and my brothers and sisters for their love, support and belief in me.

Thanks to my supervisory committee members Dr. Noah Forman and Dr. Shui Feng. I would also like to thank my external examiner Dr. Patrick Flaherty and the chair of my defence Dr. Michael Noseworthy.

Publications

The following publications based on the work presented in this thesis have been published, submitted, or are in preparation for submission for publication:

[1] Alamer, E.M.S., Gallaugher, M.P.B., McNicholas, P.D. (2022), 'Model-based clustering for mixed type data using Skewed-t distributions', *In preparation.*

[2] Alamer, E.M.S., Gallaugher, M.P.B., McNicholas, P.D. (2022) 'Mixture model for contaminated mixed type data', *In preparation.*

[3] Alamer, E.M.S., Gallaugher, M.P.B., McNicholas, P.D. (2022), 'Model averaging methods for skewed data', *In preparation.*

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

With the amount of data collected every day, data mining techniques such as clustering and classification are rapidly developed. Clustering, also known as unsupervised classification, is a fundamental machine learning — or statistical learning — technique that is used in many fields of science. The goal of cluster analysis is to gather n observations into clusters or groups based on similarities. In the clustering paradigm, there are many methods for performing cluster analysis. In general, these methods are either distance-based or model-based methods. In the former, the objects will be put in one group if the distance between them is small, whereas, in the latter case, the objects will be assigned to a certain group if they have the same distribution.

1.2 Overview

A lot of work has been done to model either categorical or continuous data using clustering and classification. However, many real data sets are of mixed-type. Mixed-type data contains two or more types of variables, where a type might be categorical, ordinal, count, continuous, etc. There has been a little work done on mixed data. Browne and McNicholas (2012) used a latent variable model to model mixed data. That approach is based on Bartholomew and Knott (1999) and considers as a generalization of the latent model.

Most of the studies that proposed to model mixed-type data, including Browne and McNicholas (2012), use the normal distribution to perform classification or clustering methods. In addition, in real data applications, there are many situations where the normal distribution is not an appropriate model. That is because the normal distribution is unable to capture the potentially sizeable skewness and/or the tail-heaviness characteristic within real data.

Due to the lack of work in the literature in the area of mixed-type data cluster analysis and the move to non-Gaussian mixture model, we present in this thesis the following approaches:

- A mixture approach for skewed mixed-type data is introduced. This model is an extension of the Gaussian mixed-type mixture model for unsupervised classification. In this mixture, an asymmetric distribution, namely the skew-t distribution is used. This model provides a better fit than Gaussian mixed-type mixture and results in a more accurate statistical analysis when skewness is exhibited in the data.

- A mixture of contaminated distributions to identify "outliers" or atypical points in mixed-type data is proposed. The contaminated Gaussian distribution is used along with latent models to develop the proposed model. This mixture gives promising results in different simulation scenarios and real applications to detect atypical points and fits better than the competing Gaussian mixture model for mixed-type data.

- A family of parsimonious variance-gamma mixture models is developed based on the eigenvalue decomposition of the scale component matrix. Then, two model averaging approaches are introduced for averaging a set of models. In the first method, we average the *a posteriori* probabilities and, in the second, we average the parameter estimates of models within Occam's window.

1.3 Outline

1.3.1 Chapter 2

Chapter 2 will present background information on latent variable models and model-based clustering. Next, the expectation-maximization (EM) algorithm, where different convergence criteria — or stopping rules — are discussed. In addition, methods for model selection and performance assessment for model-based clustering are discussed.

1.3.2 Chapters 3

In this chapter, we introduce and discuss in detail a mixture model for clustering skewed mixed-type data. In our model, we assume there is skewness in the continuous variable, and we use the skew-t distribution to model the continuous variable.

Parameter estimation and derivations of the E-step and M-step for each type of variable are outlined. The mixture of the mixed typed skew-t model is applied to simulated and real data.

1.3.3 Chapter 4

A mixture model for mixed-type data that can handle potential outliers or atypical points is proposed. The contaminated Gaussian distribution, along with factor analysis and mixture models, are used together to develop contaminated mixture models for mixed-type data. Parameter estimation and model performance assessment methods are carried out through the EM algorithm and adjusted Rand index (ARI), respectively. To demonstrate the performance of our model, we use simulated and real data.

1.3.4 Chapter 5

We discuss model averaging methods for skewed data using the variance-gamma parsimonious models and Occam's window. Herein, we used two modelling averaging techniques that are based on averaging the model parameters and averaging posterior probabilities. We use the ARI and the misclassification rate for merging mixture components and matching components, respectively, as needed.

1.3.5 Chapter 6

The final chapter summarizes this thesis and outlines directions for future work.

Chapter 2

Background

2.1 Latent Variable Models

There are two types of variables in statistics: manifest (observed) variables and latent (unobserved) variables. The latter is hidden or not directly observed but can be inferred by a statistical model called a latent variable model. Latent variable models are classified into four different types (Bartholomew and Knott, 1999). Table 2.1 shows the four types of model, which are based on the nature of the manifest and the latent variables.

Table 2.1: Summary table of latent variable models similar to that given by Bartholomew and Knott (1999).

		Manifest	
		Metrical	Categorical
Latent	Metrical	Factor analysis	Latent trait analysis
	Categorical	Latent profile analysis	Latent class analysis

By using latent variable models, we solve two problems. One is to reduce the

dimensionality of data, and the other is to explain the underlying structure. The latter problem can be seen more in areas such as psychology, medicine, and social sciences.

Now, assume that the data has p-dimensional variables $\mathbf{x} = (x_1, x_2, ..., x_p)'$ which are conditionally independent given q latent variables $\mathbf{y}$ where $q < p$. Then, the density function of the p-dimensional random vector $\mathbf{x}$ can be expressed as

$$f(\mathbf{x}) = \int \prod_{j=1}^{p} g_i(x_i|\mathbf{y}) h(\mathbf{y}) d\mathbf{y}. \tag{2.1}$$

2.1.1 Factor Analysis

Factor analysis was first used in psychology by Spearman (1904). Later, Bartlett (1953) and Lawley and Maxwell (1962) introduced it in statistical terms. It is a technique that is used to reduce the dimension of a large number of observed variables when the manifest variables are continuous. Suppose that we have p-dimensional observed variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$; then we can reduce the number of variables by using q-dimensional latent variables, where $q < p$. The generative model of factor analysis can be written as

$$\mathbf{X}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{U}_i + \boldsymbol{\epsilon}_i \tag{2.2}$$

for $i = 1, ..., n$, where $\boldsymbol{\Lambda}$ is $p \times q$ matrix called the factor loading matrix, $\mathbf{U}_i \sim N(\mathbf{0}, \mathbf{I}_q)$, $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Psi})$, for some $p \times p$ diagonal matrix $\boldsymbol{\Psi}$, and $\mathbf{U}_i$ and $\boldsymbol{\epsilon}_i$ are independent of each other. Under the factor analysis model, the observed $\mathbf{X}_i$ is normally distributed with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{'} + \boldsymbol{\Psi}$. It can shown the joint distribution between

the factors $\mathbf{U}_i$ and the observed variable $\mathbf{X}_i$ can be written as

$$\begin{bmatrix} \mathbf{X}_i \\ \mathbf{U_i} \end{bmatrix} \sim N \left(\begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{'} + \boldsymbol{\Psi} & \boldsymbol{\Lambda} \\ \boldsymbol{\Lambda}^{'} & \mathbf{I}_q \end{bmatrix} \right) \tag{2.3}$$

and thus the the conditional distribution of the factors $\mathbf{U}_i$ given the observed variable $\mathbf{X}_i$ can be given by

$$\mathbf{U}_i \mid \mathbf{X}_i \sim \mathcal{N}(\boldsymbol{\beta}(\mathbf{X}_i - \boldsymbol{\mu}), \mathbf{I} - \boldsymbol{\beta}\boldsymbol{\Lambda}) \tag{2.4}$$

where $\boldsymbol{\beta} = \boldsymbol{\Lambda}^{'}(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{'} + \boldsymbol{\Psi})^{-1}$.

2.1.2 Latent Trait Analysis

Latent trait analysis is a model that can be used to model binary or multivariate categorical data. Latent trait analysis assumes that there are q-dimensional continuous latent variables $\mathbf{y}$ that can describe the underlying behaviour of K categorical levels within each observation (Bartholomew and Knott, 1999). The generative model can be written as

$$p(\mathbf{x}_i) = \int p(\mathbf{x}_i|\mathbf{y}_i)p(\mathbf{y}_i)d\mathbf{y}_i, \tag{2.5}$$

where $p(\mathbf{x}_i|\mathbf{y}_i)$ is the conditional distribution of $\mathbf{x}_i$ given $\mathbf{y}_i$, i.e.,

$$p(\mathbf{x}_i|\mathbf{y}_i) = \prod_{k=1}^{K} (\pi_k)^{x_{ik}} (1 - \pi_k(\mathbf{y}_i))^{1-x_{ik}}.$$

Note that the response function is the logistic function, i.e.,

$$\pi_k(\mathbf{y}_i) = p(x_{ik} = 1|\mathbf{y}_i) = \frac{1}{1 + \exp[-b_k + \mathbf{w}_k'\mathbf{y}_i]},$$

where b_k and $\mathbf{w}_k$ are the model parameters that are known as the intercept and the slope parameters, respectively.

2.1.3 Latent Class Analysis and Latent Profile Analysis

Latent class analysis (LCA) and latent profile analysis (LPA) are used to model categorical and continuous data, respectively (Jason and Glenwick, 2016). LCA and LPA are both used to identify the hidden subgroups from observed data. In both approaches, we assume that there is a categorical latent variable with different levels where each level represents a class that consists of all observations that share a similarity. These classes are known as latent profiles in LPA and latent classes in LCA.

One major difference between LCA and LPA is the shape of the latent classes. In LCA, the shape of the latent classes depends on the local independence assumption (i.e., within each class, the variables are independent). In contrast, this assumption is not necessary in LPA; however, imposing a restriction on the variance-covariance matrix can have an effect on the shape of latent classes. (Vermunt and Magidson, 2002). Overall, both LPA and LCA have a lot of similarities and are considered a special case of finite mixture models in the cluster analysis paradigm. The LCA model can be seen as a binomial mixture model, whereas the LPA is a Gaussian mixture model (Robertson and Kaptein, 2016). There has been a great amount of work to introduce the concepts of LCA and LPA as well as applications and extensions, e.g., Lazarsfeld and Henry (1968), Clogg and Goodman (1984), McCutcheon (1987), Bartholomew and Knott (1999), Vermunt and Magidson (2002), Vermunt (2003), Collins and Lanza (2009).

2.2 Mixture Models in Cluster Analysis

2.2.1 Finite Mixture Model

Cluster analysis is an unsupervised technique that gathers observations with similarities such as distribution and distance into the same group (cluster). In model-based clustering, the data are modelled assuming that each observation belongs to a cluster and each cluster is a probability density. Using a finite mixture model allows tractable parameter estimation as well as the ultimate classifications of each observation into a cluster (component). A finite mixture model is convex linear combination of a finite number of probability density functions. The density function of the finite mixture model can be written as

$$f(\mathbf{x}|\boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{x}|\boldsymbol{\theta}_g), \tag{2.6}$$

where $\boldsymbol{\vartheta} = (\pi_1, ..., \pi_G, \boldsymbol{\theta}_1, ...\boldsymbol{\theta}_G)$, G is the number of components, $\pi_1, ..., \pi_G$ is a vector of positive mixing proportions that sum to 1, and $f_g(\mathbf{x}|\boldsymbol{\theta}_g)$ is the gth component density. The component distributions can be of any form, continuous or discrete, such as normal (Gaussian), t, Bernoulli, Poisson, etc.

2.2.2 Gaussian Mixture Model

The most well-known mixture model in cluster analysis is the Gaussian mixture model. It has received a lot of attention in literature and its application is well-developed in the field of model-based clustering (e.g., Wolfe, 1963; Banfield and Raftery, 1993; McNicholas and Murphy, 2008). In the Gaussian mixture model, we assume that each cluster is normally distributed, each with a distinct mean and

covariance matrix. Thus (2.6) becomes

$$f(\mathbf{x}|\boldsymbol{\theta}_g) = \sum_{g=1}^{G} \pi_g \frac{1}{\sqrt{(2\pi)^p |\boldsymbol{\Sigma}_g|}} \exp\Big\{ -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_g)'\boldsymbol{\Sigma}_g^{-1}(\mathbf{x}-\boldsymbol{\mu}_g)\Big\} \tag{2.7}$$

where $\boldsymbol{\theta}_g = \{\pi_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\}_{g=1}^{G}$.

2.2.3 Mixture of Factor Analyzers Model

After the introduction of the Gaussian mixture model for clustering, it showed promising results in comparison with the distance-based clustering methods such as k-means or hierarchical clustering (e.g., MacQueen *et al.*, 1967; Hartigan and Wong, 1979; Fowlkes and Mallows, 1983). However, the Gaussian mixture model is inadequate when clusters are skewed and/or heavy tailed. Furthermore, when the number of variables is large, the Gaussian mixture model can lead to over-parametrization. With these limitations, the mixtures of factor analyzers was introduced. Within each cluster, a local factor analysis model is used, which results in local dimension reduction. In each group g, we assume that the observations $\mathbf{X}_i$ can be modelled using a q-dimension vector of latent variables $\mathbf{U}_{ig}$ analogously to the factor analysis model. The mixture of factor analyzers model is given by

$$\mathbf{X}_i = \boldsymbol{\mu}_g + \boldsymbol{\Lambda}_g \mathbf{U}_{ig} + \boldsymbol{\epsilon}_{ig} \tag{2.8}$$

with probability π_{ig}, for $i = 1, ..., n$, where $\boldsymbol{\Lambda}_g$ is $p \times q$ matrix called the factor loadings matrix, $\mathbf{U}_{ig} \sim N(\mathbf{0}, \mathbf{I}_q)$, $\boldsymbol{\epsilon}_{ig} \sim N(\mathbf{0}, \boldsymbol{\Psi}_g)$, for some $p \times p$ diagonal matrix $\boldsymbol{\Psi}_g$, and $\mathbf{U}_{ig}$ and $\boldsymbol{\epsilon}_{ig}$ are independently distributed and independent of each other. The density of

the mixture of factor analyzers model is

$$f(\mathbf{x}_i|\boldsymbol{\theta}_g) = \sum_{g=1}^{G} \pi_g \phi(\mathbf{x}_i|\boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g^{'} + \boldsymbol{\Psi}_g). \tag{2.9}$$

There has been a lot work done on the mixture of factor analyzers model and related approaches, e.g., Ghahramani *et al.* (1996), Tipping and Bishop (1999), Yoshida *et al.* (2004), Lin (2010), Andrews and Mcnicholas (2011), and Murray *et al.* (2014).

2.2.4 Mixture of Latent Trait Analyzers Model

Latent trait analyzers have been widely used for the analysis of categorical data in cluster analysis. This method is based on the latent trait model approach where the latent variable is continuous. Gollini and Murphy (2014) introduced mixture of latent trait analyzers (MLTA) model for binary categorical data. Herein, the conditional distribution of an observation $\mathbf{x}_i$ belonging to group g can be represented by the latent trait model with parameters b_{kg} and $\mathbf{w}_{kg}$ and the latent variable $\mathbf{Y}_i$ has Gaussian distribution. This model not only helps to identify the group membership of an observation but also helps to accommodate the dependency within each cluster. The MLTA model can be written as

$$p(\mathbf{x}_i|\mathbf{y}_i) = \sum_{g=1}^{G} \eta_g \int_{\mathbf{y}} p(\mathbf{x}_i|\mathbf{y}_i, z_{ig} = 1) p(\mathbf{y}_i) d\mathbf{y}_i \tag{2.10}$$

where

$$p(\mathbf{x}_i|\mathbf{y}_i, z_{ig} = 1) = \prod_{k=1}^{K} [\pi_{kg}(\mathbf{y}_i)]^{x_{ik}} [1 - \pi_{kg}(\mathbf{y}_i)]^{1-x_{ik}}$$

and

$$\pi_{kg}(\mathbf{y}_i) = p(x_{ik} = 1|\mathbf{y}_i, z_{ig} = 1) = \frac{1}{1 + \exp[-b_{kg} + \mathbf{w}'_{kg}\mathbf{y}_i]}.$$

Tang *et al.* (2015) introduced the mixture of latent trait models with common slope parameters (MLCT), where the model is developed using latent traits, and then applied it on binary data and high-dimensional binary data. The MLCT model assume that, for K binary response variables, there is a q-dimensional latent variable $\mathbf{Y}$ for each component g and all latent traits have the same slop parameters $\mathbf{W} = (w_1, ..., w_k)$. The MLCT can be written as

$$p(\mathbf{x}_i|\mathbf{y}_{ig}) = \sum_{g=1}^{G} \eta_g \int_{\mathbf{y}_g} p(\mathbf{x}_i|\mathbf{y}_{ig}, z_{ig} = 1)p(\mathbf{y}_{ig})d\mathbf{y}_{ig}, \tag{2.11}$$

where η_g is the gth mixing proportion,

$$p(\mathbf{x}_i|\mathbf{y}_{ig}, z_{ig} = 1) = \prod_{k=1}^{K} [\pi_{kg}[\mathbf{y}_{ig}]^{x_{ik}}[1 - \pi_{kg}(\mathbf{y}_{ig})]^{1-x_{ik}},$$

and

$$\pi_{kg}(\mathbf{y}_{ig}) = p(x_{ik} = 1|\mathbf{y}_{ig}, z_{ig} = 1) = \frac{1}{1 + \exp[-\boldsymbol{w}'_k\mathbf{y}_{ig}]}.$$

2.2.5 Mixture Model for Mixed-type Data

Most work that has been done in the clustering paradigm to model the mixed-type data is based on combing two latent models into one single model. This idea comes from generalizing the approach of Bartholomew and Knott (1999) by assuming the manifest $X_1, \ldots, X_n$ are conditionally independent given the latent variables. The

approaches of Muthen and Asparouhov (2006), Vermunt (2007), Browne and McNicholas (2012), McParland and Gormley (2016) and Amiri *et al.* (2018) are based on either combining latent trait and factor analysis for the analysis of mixed-type data or replacing the latent trait with latent class model and choosing the latent variable to be standard Gaussian distribution. Suppose that there are p observed variables, where the first c columns are metrical variables and $p-c$ categorical variables (binary, ordinal or nominal), then mixture model for mixed-type has the form

$$f(\mathbf{x}_i|\boldsymbol{\theta}_g) = \sum_{g=1}^{G} \pi_g \left[\int_{\mathbf{y}} \prod_{j=1}^{p} f_1(x_{ij}|\mathbf{y}_{ig}, \boldsymbol{\theta}_{jg}) h(\mathbf{y}_{ig}) d\mathbf{y}_{ig} \right] \tag{2.12}$$

for $i = 1, ..., n$ where $f_1(\cdot)$ is the density of the conditional distribution of the observed variable given the latent variables y and $h(\cdot)$ is the density of the latent variable.

2.3 Parameter Estimation

2.3.1 Expectation-Maximization Algorithm

The expectation-maximization (EM) algorithm (Dempster *et al.*, 1977) is an iterative method to estimate parameters when the data are not complete. It has two steps: the expectation step (E-step) and the maximization step (M-step). In both steps, the complete-data log-likelihood, which includes the observed and unobservable values, is used. The E-step calculates the conditional expected value of the complete-data log-likelihood, whereas the M-step consists of maximizing the conditional expectation from the E-step with respect to the model parameters. The two steps are repeated until convergence. In cluster analysis, the incompleteness in the data comes from

the missing labels (group memberships) and, in some cases, the latent variables. For $i = 1, \ldots, n$, and $g = 1, 2, \ldots, G$, the group memberships denoted by $\mathbf{z}_1, \ldots, \mathbf{z}_n$, where $\mathbf{z}_i = z_{i1}, \ldots, z_{iG}$ is an indicator to the group membership of observation i, where $z_{ig} = 1$ if the observation $\mathbf{x}_i$ is in component g and $z_{ig} = 0$ otherwise.

2.3.2 Stopping Criterion

Some of the well-known methods for determining the convergence of the EM algorithm are based on Aitken's acceleration (Aitken, 1926). In one such stopping criterion, the log-likelihood at each iteration is estimated to determines the convergence of the EM algorithm. At iteration t, the Aitken acceleration is

$$a^{(t)} = \frac{l^{(t+1)} - l^{(t)}}{l^{(t)} - l^{(t-1)}}, \tag{2.13}$$

where $l^{(t)}$ is the log-likelihood at iteration t. Böhning *et al.* (1994) define the asymptotic estimate of log-likelihood by

$$l_{\infty}^{(t+1)} = l^{(t)} + \frac{l^{(t+1} - l^{(t)}}{1 - a^{(t)}}. \tag{2.14}$$

At iteration t, Lindsay (1995) suggested that the EM algorithm will converge when the difference between the asymptotic log-likelihood and the observed log-likelihood is small such that

$$l_{\infty}^{(t)} - l^{(t)} < \epsilon, \tag{2.15}$$

where ϵ is very small and the difference is positive.

2.4 Model Selection and Performance Assessment

2.4.1 Bayesian Information Criterion

There are different techniques to choose the best model in cluster analysis. The Bayesian information criterion (BIC; Schwarz *et al.*, 1978) is the most widely used model selection criteria in model-based clustering. It can be calculated from the following equation

$$\text{BIC} = 2l(\hat{\boldsymbol{\vartheta}}) - p \log n, \tag{2.16}$$

where $l(\hat{\boldsymbol{\vartheta}})$ is the maximized log-likelihood, p is the number of free parameters to be estimated, and n is the number of observations. When we compare different models, the model with the largest BIC is chosen. The BIC is not only used as a model selection criterion but also helps to select the number of components G and the number of latent variables q. Keribin (2000) and Kass and Wasserman (1995) discussed the theoretical justifications to use the BIC as a model selection criterion in model-based clustering, whereas Fraley and Raftery (2002), Raftery and Dean (2006) and McNicholas and Murphy (2008) provide practical justifications.

2.4.2 Adjusted Rand Index

In cluster analysis, the labels are unknown or we assume that the (true) labels are unknown. However, in illustrative examples, we have usually known the labels and we can use them to assess the performance of clustering techniques. Hubert and Arabie (1985) proposed the adjusted Rand index (ARI) to measure the performance of the clustering. It compares two different partitions in the dataset and has the following

general form

$$\frac{\text{index - expected index}}{\text{maximum index - expected index}},$$

where "index" refers to the Rand index (Rand, 1971). The expected value of ARI under random classification is 0 and value of 1 for perfect classification.

2.5 Variance-Mean Mixtures

Suppose there is a k-variate random vector $\mathbf{X}$ defined in terms of a variance-mean mixture. Then $\mathbf{X}$ has a probability density function of the form

$$f(\mathbf{x}) = \int_0^\infty \phi_k(\mathbf{x}|\boldsymbol{\mu} + w\boldsymbol{\alpha}, w\boldsymbol{\Sigma})h(w|\boldsymbol{\theta})dw,$$

where W is a random variable $W > 0$ that has density function $h(w|\boldsymbol{\theta})$, and $\phi_k(\cdot)$ is the density function of the k-variate Gaussian distribution. Equivalently, $\mathbf{X}$ can be written in the form

$$\mathbf{X} = \boldsymbol{\mu} + W\boldsymbol{\alpha} + \sqrt{W}\mathbf{V}, \tag{2.17}$$

where $\boldsymbol{\mu}$ is a location parameter, $\boldsymbol{\alpha}$ is the skewness, $\mathbf{V}$ has a multivariate Gaussian distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$, W has density function $h(w|\boldsymbol{\theta})$, and W and $\mathbf{V}$ are independent. By changing the distribution of W, we can use the variance-mean mixture representation to obtain other multivariate distributions (McNicholas, 2016). For example, when $W \sim \text{IG}(\nu/2, \nu/2)$, we can obtain the multivariate skew-t distribution with ν degrees of freedom, where IG$(\cdot)$ denotes the inverse

Gamma distribution and has density function

$$f(w|\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} w^{-\alpha-1} \exp\left\{-\frac{\beta}{w}\right\}.$$

Similarly, we can use to obtain the variance-gamma distribution by letting the distribution of W in (2.17) be from a gamma with shape and scale parameter $\lambda, \psi/2$, where the density function of gamma distribution is

$$f(w|\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} w^{\alpha-1} \exp\{-\beta w\}.$$

The variance-mean mixture representation used widely in model-based clustering and classification methods, e.g., Murray *et al.* (2014) use skew-t, and McNicholas *et al.* (2017) use variance-gamma distribution.

2.6 Non-Gaussian Distributions

2.6.1 Generalized Inverse Gaussian Distribution

A random variable X follows a Generalized Inverse Gaussian (GIG) distribution (Good, 1953), if its density function written as

$$f(x \mid a,b,\lambda) = \frac{(a/b)^{\frac{\lambda}{2}}\, x^{\lambda-1}}{2K_{\lambda}(\sqrt{ab})} \exp\left\{-\frac{ax + b/x}{2}\right\},$$

where $(a,b) \in \mathbb{R}^{+}$, $\lambda \in \mathbb{R}$ and $K_{\lambda}(u)$ is the modified Bessel function of the third kind defined as

$$K_{\lambda}(u) = \frac{1}{2}\int_{0}^{\infty} x^{\lambda-1} \exp\left\{-\frac{u}{2}\left(x + \frac{1}{x}\right)\right\} dx.$$

The GIG distribution has important tractable expectations that is used for parameter estimation in different skewed data e.g.,

$$\mathbb{E}(X) = \sqrt{\frac{b}{a}}\frac{K_{\lambda+1}(\sqrt{ab})}{K_\lambda(\sqrt{ab})}, \tag{2.18}$$

$$\mathbb{E}\left(1/X\right) = \sqrt{\frac{a}{b}}\frac{K_{\lambda+1}(\sqrt{ab})}{K_\lambda(\sqrt{ab})} - \frac{2\lambda}{b}, \tag{2.19}$$

$$\mathbb{E}(\log X) = \log\left(\sqrt{\frac{b}{a}}\right) + \frac{1}{K_\lambda(\sqrt{ab})}\frac{\partial}{\partial\lambda}K_\lambda(\sqrt{ab}). \tag{2.20}$$

2.6.2 Generalized Hyperbolic Distribution

Following McNeil *et al.* (2015) parametrizations for the generalized hyperbolic distribution (GHD), a p-dimensional random variable $\mathbf{X}$ said to have GHD if it has density function of form

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

where $\boldsymbol{\vartheta} = (\lambda, \chi, \psi, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$, $\lambda > 0$ is the index parameter, χ, ψ are concentration parameters, $\boldsymbol{\mu}, \boldsymbol{\alpha} \in \mathbb{R}^p$ are the location and skewness parameter, and $K(\cdot)$ is the modified Bessel function of the third kind.

Chapter 3

Mixture for Skewed Mixed-type Data

3.1 Introduction

In this chapter, we will introduce a mixture approach for skewed mixed-type data by extending the mixture of mixed-type data approach that has been introduced by Browne and McNicholas (2012). This extension allows for the presence of skewness in the continuous variables. This model uses the skew-t factor analysis model and the latent trait model. The factor analysis model is used to model the continuous variables, whereas the latent trait model is for the categorical variables. This chapter is arranged as follows. Details of the development of density and the likelihoods of the proposed model are presented. Next, we outline the parameter estimation procedure using the EM algorithm and computational consideration. The chapter concludes with the application of our model that is applied to simulated and real data and a summary section.

3.2 Skew-t Factor Analyzers

The density of a p-dimensional random vector $\mathbf{X}$ following a skew-t distribution (St) is

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

where $\boldsymbol{\mu}$ is the location parameter , $\boldsymbol{\Sigma}$ is the scale matrix, $\boldsymbol{\alpha}$ is the skewness parameter, ν is the degrees of freedom, and $K(\cdot)$ is the modified Bessel function of the third kind.

Let $\mathbf{X} \sim \text{St}(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\alpha},\nu)$ denotes that the p-dimensional random vector $\mathbf{X}$ follows a skew-t distribution. Then we can generate the random variable $\mathbf{X}$ via combining a random variable $W_i \sim \text{IG}(\nu/2,\nu/2)$, where IG denotes the inverse-gamma distribution with the latent $\mathbf{V} \sim \mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})$ via

$$\mathbf{X} = \boldsymbol{\mu} + W\boldsymbol{\alpha} + \sqrt{W}\mathbf{V}.$$

It is easy to show that $\mathbf{X}_i \mid w_i \sim \mathcal{N}(\boldsymbol{\mu}+w_i\boldsymbol{\alpha}, w_i\boldsymbol{\Sigma})$. Now, from Bayes' theorem we get

$$\begin{aligned} f(w \mid \mathbf{x}) &= \frac{f(\mathbf{x} \mid w)g(w)}{f(\mathbf{x})} \\ &= \left[\frac{\psi+\boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}{\nu+\delta(\boldsymbol{x},\boldsymbol{\mu}|\boldsymbol{\Sigma})}\right]^{(\nu-p/2)/2} \frac{1}{2K_{\nu-p/2}\left(\sqrt{[\nu+\boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}][\nu+\delta(\boldsymbol{x},\boldsymbol{\mu}|\boldsymbol{\Sigma})]}\right)} \\ &\quad \times w^{\nu-p/2-1}\exp\{-\left[w(\psi+\boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}+(\nu+\delta(\boldsymbol{x},\boldsymbol{\mu}|\boldsymbol{\Sigma}))/w\right]/2\} \end{aligned}$$

so that $W_i \mid (\mathbf{X}_i = \mathbf{x}) \sim \text{GIG}(\boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}, \nu+\delta(\boldsymbol{x},\boldsymbol{\mu}|\boldsymbol{\Sigma}), -(\nu+p)/2)$ where GIG denotes

the generalized inverse Gaussian distribution. Using a factor analysis model, we can rewrite $\mathbf{X}$ as

$$\mathbf{X}_i = \boldsymbol{\mu} + W_i\boldsymbol{\alpha} + \mathbf{V}_i^* \tag{3.1}$$

where

$$\mathbf{V}_i^* = \boldsymbol{\Lambda}\mathbf{Y} + \boldsymbol{\epsilon}_i \tag{3.2}$$

where $\boldsymbol{\Lambda}$ is a matrix of factor loadings, $\mathbf{Y} \sim t(\nu)$ is a q-dimensional latent variable and $\mathbf{Y}_i \mid w_i \sim \mathcal{N}(\mathbf{0}, w_i\mathbf{I}_q)$, $\boldsymbol{\epsilon} \mid w_i \sim \mathcal{N}(\mathbf{0}, w_i\boldsymbol{\Psi})$ where is $\boldsymbol{\Psi} = \text{dig}(\psi_1, ..., \psi_p)$. Substituting (3.2) in (3.1), the skew-t factor analysis model can be written as

$$\mathbf{X}_i = \boldsymbol{\mu} + W_i\boldsymbol{\alpha} + \boldsymbol{\Lambda}\mathbf{Y}_i + \boldsymbol{\epsilon}_i. \tag{3.3}$$

Then it is easy to show $\mathbf{X}_i \sim \text{St}(\boldsymbol{\mu}, \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}, \boldsymbol{\alpha}, \nu)$ and $\mathbf{X}_i \mid \mathbf{y}_i, w_i \sim \mathcal{N}(\boldsymbol{\mu} + w_i\boldsymbol{\alpha} + \boldsymbol{\Lambda}\mathbf{y}_i, w_i\boldsymbol{\Psi})$.

3.3 Mixture Model for Skewed-t Mixed-type Data

Suppose that we have mixed-type data with i rows and p columns. Without loss of generality, let the first c columns be the categorical variables, and the remaining $p - c$ columns be continuous variables. Assume we have p-dimensional observed variables $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})$ and q-dimensional latent variables $\mathbf{Y}_i$ with $q < p$. In our model, we assume that the observed variables are independent given latent variables. Furthermore, let W_i as defined before be another latent variable and then assume that $\mathbf{Y}_i \mid w_i \sim \mathcal{N}(\mathbf{0}, w_i\mathbf{I}_q)$.

Now, if the manifest variable X_{ij} is categorical with levels $0, 1, 2, \ldots K_j - 1$, then

the conditional distribution is Bernoulli with success probability

$$g_1(x_{ij} = k|\mathbf{y}_i, w_i, \boldsymbol{\theta}_j) = \frac{\prod_{k=0}^{K_j-1}[\exp\{\eta_{jk} + \boldsymbol{\tau}'_{jk}\mathbf{y}_i\}]^{x_{ij}(k)}}{1 + \sum_{k=0}^{K_j-1}\exp\{\eta_{jk} + \boldsymbol{\tau}'_{jk}\mathbf{y}_i\}}, \tag{3.4}$$

where $x_{ij}(k) = 1$ if x_{ij} is in category k and 0 otherwise.

If the observed variable X_{ij} is continuous and follows a skew-t distribution, the conditional density is similar to conditional density that is derived from skew-t factor analysers

$$g_1(x_{ij}|\mathbf{y}_i, w_i, \boldsymbol{\theta}_j) = \frac{1}{\sqrt{2\pi w_i\psi_j}}\exp\left\{-\frac{1}{2w_i\psi_j}(x_{ij} - \mu_j - \boldsymbol{\lambda}_j\mathbf{y}_i - w_i\alpha_j)^2\right\}, \tag{3.5}$$

where μ_j and α_j are the jth element of $\boldsymbol{\mu}$ and $\boldsymbol{\alpha}$, respectively, and $\boldsymbol{\lambda}_j$ is the jth row of $\boldsymbol{\Lambda}$.

Combining both density functions from the continuous and the categorical observed variables, we can extend (2.1) as follows

$$f(\boldsymbol{x}|\boldsymbol{\vartheta}) = \int_{\mathbf{y}}\int_{w}\prod_{j=1}^{p} g_1(x_{ij}|\mathbf{y}_i, w_i, \boldsymbol{\theta}_j)g_2(\mathbf{y}_i|w_i)g_3(w_i)dw_i d\mathbf{y}_i. \tag{3.6}$$

Now applying model-based cluster analysis methodology, we can write the mixture model of latent variables for mixed-type data as

$$f(\boldsymbol{x}|\boldsymbol{\vartheta}) = \sum_{g=1}^{G}\pi_g\left[\int_{\mathbf{y}}\int_{w}\prod_{j=1}^{p} g_1(x_{ij}|\mathbf{y}_{ig}, w_{ig}, \boldsymbol{\theta}_{jg})g_2(\mathbf{y}_{ig}|w_{ig})g_3(w_{ig})dw_{ig}d\mathbf{y}_{ig}\right] \tag{3.7}$$

where $\boldsymbol{\vartheta}$ denotes all model parameters.

3.3.1 Parameter Estimation

Parameter estimation is carried out within the EM algorithm. We start by initializing the parameters $\boldsymbol{\mu}_g, \boldsymbol{\alpha}_g, \boldsymbol{\Lambda}_g\boldsymbol{\Psi}_g, \nu_g, \eta_g,$ and $\boldsymbol{\tau}_g$ for $g = 1, \ldots, G$. Then, the EM algorithm alternates between E-step and the M-steps until convergence. In the E-step, we calculate the conditional expectation of the complete-data log-likelihood. The complete-data consists of the labels $\mathbf{z}_1, \ldots, \mathbf{z}_n$, the latent variables $\mathbf{y}_i, \ldots, \mathbf{y}_n$ and $w_1, \ldots, w_n$ along with the observed data. The complete-data log-likelihood has the form

$$l_c(\boldsymbol{\vartheta}) = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 + \mathcal{L}_4 + \mathcal{L}_5, \tag{3.8}$$

where

$$\mathcal{L}_1 = \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig} \log \pi_g,$$

$$\mathcal{L}_2 = \sum_{i=1}^{n}\sum_{g=1}^{G}\sum_{j=1}^{c} z_{ig} \left[\log \left(\prod_{k=0}^{K_j-1} [\exp\{\eta_{jkg} + \boldsymbol{\tau}'_{jkg}\mathbf{y}_{ig}\}]^{x_{ij}(k)} \right) - \log \left(1 + \sum_{k=0}^{K_j-1} \exp\{\eta_{jkg} + \boldsymbol{\tau}'_{jkg}\mathbf{y}_{ig}\} \right) \right],$$

$$\mathcal{L}_3 = \sum_{i=1}^{n}\sum_{g=1}^{G}\sum_{j=c+1}^{p} z_{ig} \left[\frac{1}{2}\log 2\pi + \frac{1}{2}\log\left(\frac{1}{w_{ig}}\right) - 2\log\psi_{jg} - \frac{1}{2w_{ig}\psi_{jg}}(x_{ij} - \mu_{jg} - \boldsymbol{\lambda}_{jg}\mathbf{y}_{ig} - w_{ig}\alpha_{jg})^2 \right],$$

$$\mathcal{L}_4 = \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig} \left[\frac{q}{2}\log\left(\frac{1}{w_{ig}}\right) + \frac{q}{2}\log\left(\frac{1}{2\pi}\right) + \frac{1}{2w_{ig}}\mathbf{y}_{ig}\mathbf{y}'_{ig} \right],$$

$$\mathcal{L}_5 = \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\left[\frac{\nu_g}{2}\log\frac{\nu_g}{2} + \log\frac{1}{\Gamma(\frac{\nu_g}{2})} + (\frac{\nu_g}{2}-1)\log w_{ig} - \frac{1}{w_{ig}}\frac{\nu_g}{2}\right].$$

The following expectations are needed to estimate and update the latent and the labels:

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

2. $\mathbb{E}[W_{ig}|\mathbf{x}_i, Z_{ig}=1] =: a_{ig}$

$$\mathbb{E}[W_{ig}|\mathbf{x}_i, Z_{ig}=1] = \int_w w_{ig}\frac{f(\mathbf{x}_i, w_{ig})}{f(\mathbf{x}_i)}dw_{ig}.$$

3. $\mathbb{E}[1/W_{ig}|\mathbf{x}_i, Z_{ig}=1] =: b_{ig}$ and $\mathbb{E}[\log W_{ig}|\mathbf{x}_i, Z_{ig}=1] =: c_{ig}$ can be calculated from the conditional expectation of a function of W

$$\mathbb{E}[g(W_{ig})|\mathbf{x}_i, Z_{ig}=1] = \int_w g(w_{ig})\frac{f(\mathbf{x}_i, w_{ig})}{f(\mathbf{x}_i)}dw_{ig}.$$

4. $\mathbb{E}[\mathbf{Y}_{ig}|\mathbf{x}_i, Z_{ig}=1] =: \boldsymbol{e}_{1ig}$

$$\mathbb{E}[Y_{ig}|\mathbf{x}_i, Z_{ig}=1] = \int_y \mathbf{y}_{ig}\frac{f(\mathbf{x}_i, \mathbf{y}_{ig})}{f(\mathbf{x}_i)}d\mathbf{y}_{ig}.$$

5. $\mathbb{E}[(1/W_{ig})\mathbf{Y}_{ig}|\mathbf{x}_i, Z_{ig}=1] = \mathbb{E}[g(W_{ig})\mathbb{E}[\mathbf{Y}_{ig}|\mathbf{x}_i, w_{ig}, Z_{ig}=1]|\mathbf{x}_i] =: \boldsymbol{e}_{2ig}$

$$= \int_w \left\{g(w_{ig})\frac{f(\mathbf{x}_i, w_{ig})}{f(\mathbf{x}_i)}\left[\int_y \mathbf{y}_{ig}\frac{f(\mathbf{x}_i, \mathbf{y}_{ig}, w_{ig})}{f(\mathbf{x}_i, w_{ig})}d\mathbf{y}_{ig}\right]\right\}dw_{ig}.$$

6. $\mathbb{E}[(1/W_{ig})\mathbf{Y}_{ig}\mathbf{Y}'_{ig}|\mathbf{x}_i, Z_{ig}=1] = \mathbb{E}[g(W_{ig})\mathbb{E}[\mathbf{Y}_{ig}\mathbf{Y}'_{ig}|\mathbf{x}_i, w_{ig}, Z_{ig}=1]|\mathbf{x}_i] =: \boldsymbol{E}_{3ig}$

$$= \int_w \left\{g(w_{ig})\frac{f(\mathbf{x}_i, w_{ig})}{f(\mathbf{x}_i)}\left[\int_y \mathbf{y}_{ig}\mathbf{y}'_{ig}\frac{f(\mathbf{x}_i, \mathbf{y}_{ig}, w_{ig})}{f(\mathbf{x}_i, w_{ig})}d\mathbf{y}_{ig}\right]\right\}dw_{ig},$$

where

$$f(\mathbf{x}_i, \mathbf{y}_{ig}, w_{ig}) = \prod_{j=1}^{p} g_1(\mathbf{x}_i|\mathbf{y}_i, w_{ig}, \boldsymbol{\theta}_{jg}) g_2(\mathbf{y}_{ig}|w_{ig}) g_3(w_{ig}),$$

$$f(\mathbf{x}_i, \mathbf{y}_{ig}) = \int_w f(\mathbf{x}_i, \mathbf{y}_{ig}, w_{ig}) dw_{ig},$$

$$f(\mathbf{x}_i, w_{ig}) = \int_y f(\mathbf{x}_i, \mathbf{y}_{ig}, w_{ig}) d\mathbf{y}_{ig}.$$

Note that all the integrals do not have a closed form when the data are mixed-type and must be solved numerically.

In the M-step, we maximize the expected value of the complete log-likelihood to update the parameters in our model. The update for the mixing proportions is given by

$$\widehat{\pi}_g = \frac{n_g}{n},$$

where $n_g = \sum_{i=1}^{n} \hat{z}_{ig}$.

If the manifest variable X_i is continuous, then the update for the parameters $\boldsymbol{\mu}_g, \boldsymbol{\alpha}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g, \nu_g$ are analogous to the updates in factor analysis methodology.

$$\widehat{\boldsymbol{\mu}}_g = \frac{\sum_{i=1}^{n} \mathbf{x}_i \hat{z}_{ig}(\bar{a} b_{ig} - 1)}{\sum_{i=1}^{n} \hat{z}_{ig}(\bar{a} b_{ig} - 1)}, \qquad \widehat{\boldsymbol{\alpha}}_g = \frac{\sum_{i=1}^{n} \hat{z}_{ig} \mathbf{x}_i (\bar{b}_g - b_{ig})}{\sum_{i=1}^{n} \hat{z}_{ig}(\bar{a} b_{ig} - 1)},$$

where $n_g = \sum_{i=1}^{n} \hat{z}_{ig}$,

$$\bar{a}_g = \frac{\sum_{i=1}^{n} \hat{z}_{ig} a_{ig}}{n_g}, \qquad \bar{b}_g = \frac{\sum_{i=1}^{n} \hat{z}_{ig} b_{ig}}{n_g},$$

$$\hat{\boldsymbol{\Lambda}}_g = \left\{ \sum_{i=1}^{n} \hat{z}_{ig} [(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g) \boldsymbol{e}'_{2ig} - \hat{\boldsymbol{\alpha}}_g \boldsymbol{e}'_{1ig}] \right\} \left\{ \sum_{i=1}^{n} \hat{z}_{ig} \boldsymbol{E}'_{3ig} \right\}^{-1},$$

$$\hat{\boldsymbol{\Psi}}_g = \frac{1}{\sum_{i=1}^{n} \hat{z}_{ig}} \text{dig} \left\{ \sum_{i=1}^{n} \hat{z}_{ig} [b_{ig}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)^{'} - 2\hat{\boldsymbol{\alpha}}_g(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)^{'} + a_{ig}\hat{\boldsymbol{\alpha}}_g\hat{\boldsymbol{\alpha}}_g^{'} \right.$$
$$\left. -2(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)\boldsymbol{e}_{2ig}^{'}\hat{\boldsymbol{\Lambda}}_g^{'} + 2\hat{\boldsymbol{\alpha}}_g\boldsymbol{e}_{1ig}^{'}\hat{\boldsymbol{\Lambda}}_g^{'} + \hat{\boldsymbol{\Lambda}}_g\boldsymbol{E}_{3ig}\hat{\boldsymbol{\Lambda}}_g^{'}] \right\}.$$

The update for the degree of freedom ν can be calculated by solving the following equation:

$$\left(\sum_{i=1}^{n} \hat{z}_{ig}\right) \left[\log\left(\frac{\hat{\nu}_g^{\text{new}}}{2}\right) - \varphi\left(\frac{\hat{\nu}_g^{\text{new}}}{2}\right) + 1\right] - \sum_{i=1}^{n} \hat{z}_{ig}\left(c_{ig} - b_{ig}\right) = 0$$

where $\varphi(\cdot)$ is the digamma function.

Now, if the observed variable X_i is categorical, then the update for the parameters $\eta_g, \boldsymbol{\tau}_g$ can be obtained by by numerically solving the following system of non-linear equations:

$$\sum_{i=1}^{n} \hat{z}_{ig} \int_{y} (1, \mathbf{y}_{ig}) \left[x_{ij}(k) - g(x_{ij} = k|\mathbf{y}_{ig}, w_{ig}, \boldsymbol{\theta}_{jg})\right] h(\mathbf{y}_i|\mathbf{x}_i) d\mathbf{y}_{ig} = \mathbf{0}.$$

3.3.2 Computational Considerations

Numerical calculation problems

We encounter two problems, one related to calculating the expectations in the E-step and one in the M-step, specifically in the update for the categorical variable parameters. To overcome the first problem, we use some R packages to calculate and approximate the numerical integral such as `cubature` and `calculus` (Narasimhan *et al.*, 2021; Guidotti, 2020). For the latter issue, we use numerical methods such as Newton–Raphson from `nleqslv` package to compute the solution of the non-linear

system of the equations (Hasselman and Hasselman, 2018).

Initialization

As the EM algorithm is sensitive to the starting value, it is important to choose a starting value that helps to get a better estimator and speed the convergence. For the group membership z_{ig}, we use soft values that are randomly generated from a uniform over the interval (0,1). For the continuous variable parameters: $\boldsymbol{\mu}_g$ are based on calculating the weighted mean of z_{ig}, the degrees of freedom $v_g = 30$, the loading factor matrix $\boldsymbol{\Lambda}_g$ a $p \times q$ matrix of ones, and $p \times p$ identity matrix for $\boldsymbol{\Psi}_g$. We follow the initialization recommendation for the logistic and multinomial regression to initialize categorical variable parameters so that $\eta_g = 0, \boldsymbol{\tau}_g = \mathbf{0}$.

Number of Free Parameters

Number of free parameters that will be used to calculate the BIC is based on the number of free parameters in p_1-dimensional continuous and p_2-dimensional categorical variables. Hence, the number of free parameters in the MMSM model is

$$(G-1) + 2Gp_1 + G + G\left[p_1 q - \frac{1}{2}q(q-1)\right] + Gp_1 + G\left[p_2 q - \frac{1}{2}q(q-1)\right] + Gp_2.$$

3.4 Simulation

Two simulations are performed to illustrate the accuracy of our model. We first simulate data sets from our mixture model with two groups and $n = 100, 200, 400$ observations per group. In the first simulation, the two mixture components are very well separated whereas in the second simulation there is an overlap between clusters.

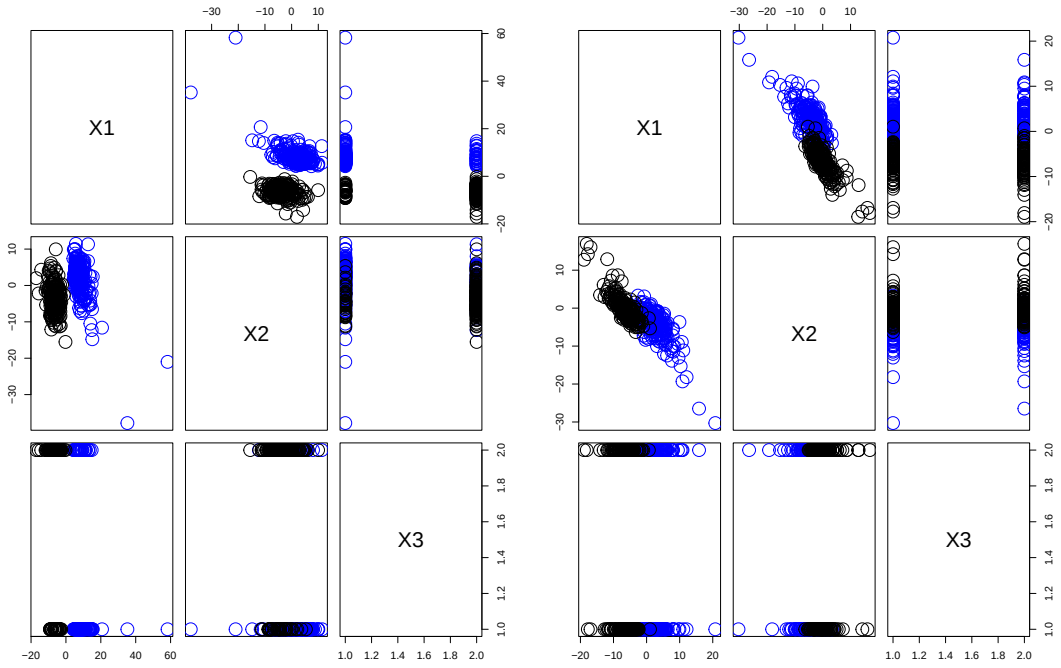
Figure 3.1 shows an illustration of one of the simulated data sets from each simulation. We fit different models to the simulated data by varying the number of groups G and the dimension of the latent variables q. We then calculate the ARI and BIC for each scenario and choose the best model based on the highest BIC.

Table 3.1 and Table 3.2 display the result for the the true value of the model parameters as well as their mean and standard error for simulation studies 1 and 2, respectively. We can notice that the sample size has an effect on the parameter estimation. As the sample size increases, the mean of the parameter gets closer to the true parameter and the standard deviation decreases.

Table 3.3 presents the number of groups that is chosen by the BIC and the number of latent factors q. From Table 3.3 , we notice that the average ARI values are higher when the groups are well separated. It also picks the correct number of groups and the number of latent factors more often than when the groups are overlap. From both simulations, we can see there is an effect of sample size on ARI. As we increased the sample size, the ARI values increase too.

3.5 Australian Institute of Sport Data

The Australian Institute of Sport (AIS) data is a well-known data set that is used to illustrate the performance of mixtures of skewed distributions (e.g., Murray *et al.*, 2014). The data frame has 202 observations on 13 continuous variables. The variables represent 13 body measurements for 102 males and 100 females. Since all the variables are numerical, we categorize the hemoglobin concentration (HGB) variable as follows: If the HGB is between 12 and 17.4, it is normal; otherwise, it is not. Our analysis also considers body fat percentage (PBF) and body mass index (BMI).



(a) (b)

Figure 3.1: Example of one of the simulated data set from (a) Simulation 1 , (b) Simulation 2.

Table 3.1: The true parameters, value the means and the standard deviations from simulation 1.

Parameters	True values	Means			Standard deviations		
		$n = 100$	$n = 200$	$n = 400$	$n = 100$	$n = 200$	$n = 400$
η_1	1	1.96	1.20	1.30	0.60	0.45	0.40
η_2	-1	-1.38	-1.08	-1.07	0.81	0.40	0.39
τ_1	0	0.17	-0.18	-0.19	0.73	0.30	0.22
τ_2	0	-0.05	-0.15	-0.05	1.27	0.58	0.42
$\boldsymbol{\mu}_1$	$(5,5)'$	$(4.30, 5.33)'$	$(5.70, 4.43)'$	$(5.85, 4.39)'$	$(1.10, 1.82)'$	$(0.68, 0.94)'$	$(0.64, 0.81)'$
$\boldsymbol{\mu}_g$	$(-5,-5)'$	$(-4.35,-5.10)'$	$(-4.86,-5.21)'$	$(-5.09,-5.28)'$	$(1.59, 1.96)'$	$(0.80, 1.18)'$	$(0.40, 0.65)'$
$\boldsymbol{\alpha}_1$	$(1,-1)$	$(0.61,-0.84)'$	$(1.24,-1.35)'$	$(1.17,-1.33)'$	$(1.45, 1.86)'$	$(0.53, 0.71)'$	$(0.52, 0.66)'$
$\boldsymbol{\alpha}_2$	$(-1,1)$	$(-0.82, 1.35)'$	$(-1.09, 1.19)'$	$(-0.90, 1, 16)'$	$(1.16, 1.76)'$	$(0.72, 1.07)'$	$(0.33, 0.56)'$
ν_1	6	6.42	6.12	6.08	2.55	1.24	1.08
ν_g	8	8.29	7.64	7.12	6.20	2.66	1.44
$\boldsymbol{\Sigma}_1$	$\begin{bmatrix} 2 & -0.1 \\ -0.1 & 11 \end{bmatrix}$	$\begin{bmatrix} 2.09 & -0.01 \\ -0.01 & 8.56 \end{bmatrix}$	$\begin{bmatrix} 1.41 & -0.01 \\ -0.01 & 10.35 \end{bmatrix}$	$\begin{bmatrix} 1.79 & -0.26 \\ -0.26 & 10.20 \end{bmatrix}$	$\begin{bmatrix} 1.10 & 1.67 \\ 1.67 & 6.23 \end{bmatrix}$	$\begin{bmatrix} 1.32 & 1.62 \\ 1.62 & 4.49 \end{bmatrix}$	$\begin{bmatrix} 1.20 & 1.40 \\ 1.40 & 3.99 \end{bmatrix}$
$\boldsymbol{\Sigma}_2$	$\begin{bmatrix} 2 & -0.1 \\ -0.1 & 10 \end{bmatrix}$	$\begin{bmatrix} 1.79 & -0.01 \\ -0.01 & 9.46 \end{bmatrix}$	$\begin{bmatrix} 2.47 & -0.02 \\ -0.02 & 9.41 \end{bmatrix}$	$\begin{bmatrix} 2.57 & -0.01 \\ -0.01 & 9.05 \end{bmatrix}$	$\begin{bmatrix} 1.43 & 1.26 \\ 1.26 & 5.34 \end{bmatrix}$	$\begin{bmatrix} 0.60 & 0.52 \\ 0.52 & 2.29 \end{bmatrix}$	$\begin{bmatrix} 0.66 & 0.68 \\ 0.68 & 2.35 \end{bmatrix}$

Table 3.2: The true parameters value , the means and the standard deviations from simulation 2.

Parameters	True values	Means			Standard deviations		
		$n = 100$	$n = 200$	$n = 400$	$n = 100$	$n = 200$	$n = 400$
η_1	0	0.16	0.13	0.12	0.19	0.20	0.26
η_2	-1	-0.52	-0.61	-0.91	0.31	0.19	0.16
τ_1	0.1	0.08	0.15	0.07	0.12	0.11	0.11
τ_2	-0.1	-0.26	-0.06	-0.12	0.13	0.08	0.11
$\boldsymbol{\mu}_1$	$(1, 1)'$	$(1.34, 1.08)'$	$(1.11, 0.52)'$	$(0.78, 0.82)'$	$(3.36, 1.48)'$	$(2.90, 1.35)'$	$(2.45, 0.88)'$
$\boldsymbol{\mu}_g$	$(-4, -4)'$	$(-5.72, -4.11)'$	$(-3.53, -4.41)'$	$(-4.13, -4.69)'$	$(3.29, 1.33)'$	$(3.411, 0.70)'$	$(2.40, 0.94)'$
$\boldsymbol{\alpha}_1$	$(2, -3)'$	$(1.70, -2.72)'$	$(1.79, -2.80)'$	$(1.66, -2.58)'$	$(1.21, 1.95)'$	$(1.16, 1.79)'$	$(0.81, 1.26)'$
$\boldsymbol{\alpha}_2$	$(-2, 3)'$	$(-1.84, 2.71)'$	$(-1.94, 2.07)'$	$(-1.73, 2.52)'$	$(1.24, 1.87)'$	$(0.86, 1.23)'$	$(0.79, 1.22)'$
ν_1	6	5.40	5.61	5.85	1.17	0.04	0.75
ν_g	8	7.62	7.98	7.85	0.98	0.01	0.96
$\boldsymbol{\Sigma}_1$	$\begin{bmatrix} 5 & -4 \\ -4 & 4.5 \end{bmatrix}$	$\begin{bmatrix} 4.95 & -3.02 \\ -3.02 & 3.46 \end{bmatrix}$	$\begin{bmatrix} 4.94 & -3.27 \\ -3.27 & 5.14 \end{bmatrix}$	$\begin{bmatrix} 4.88 & -3.32 \\ -3.32 & 5.38 \end{bmatrix}$	$\begin{bmatrix} 0.52 & 0.05 \\ 0.05 & 0.68 \end{bmatrix}$	$\begin{bmatrix} 0.31 & 0.07 \\ 0.07 & 0.40 \end{bmatrix}$	$\begin{bmatrix} 0.41 & 0.02 \\ 0.02 & 0.51 \end{bmatrix}$
$\boldsymbol{\Sigma}_2$	$\begin{bmatrix} 4.5 & -4 \\ -4 & 5 \end{bmatrix}$	$\begin{bmatrix} 5.27 & -4.93 \\ -4.93 & 5.63 \end{bmatrix}$	$\begin{bmatrix} 5.45 & -3.73 \\ -3.73 & 4.62 \end{bmatrix}$	$\begin{bmatrix} 5.10 & -3.77 \\ -3.77 & 4.62 \end{bmatrix}$	$\begin{bmatrix} 0.51 & 0.04 \\ 0.04 & 0.53 \end{bmatrix}$	$\begin{bmatrix} 0.44 & 0.07 \\ 0.07 & 1.22 \end{bmatrix}$	$\begin{bmatrix} 0.33 & 0.03 \\ 0.03 & 0.45 \end{bmatrix}$

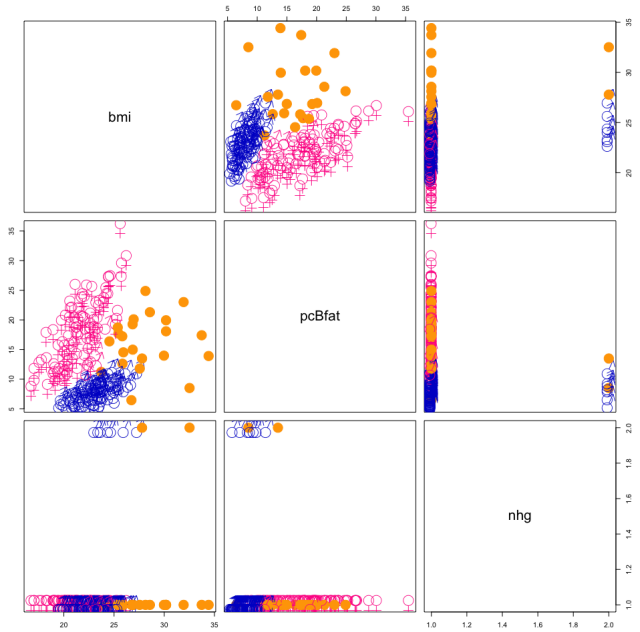
Table 3.3: The number of times that the BIC correctly chose the number of groups G, factors q and the average ARI for simulation 1 and simulation 2 .

n	Simulation 1			Simulation 2		
	G	q	$\overline{\text{ARI}}$	G	q	$\overline{\text{ARI}}$
100	25	25	1.00	23	24	0.81
200	25	25	0.99	24	25	0.86
400	25	25	1.00	24	25	0.90

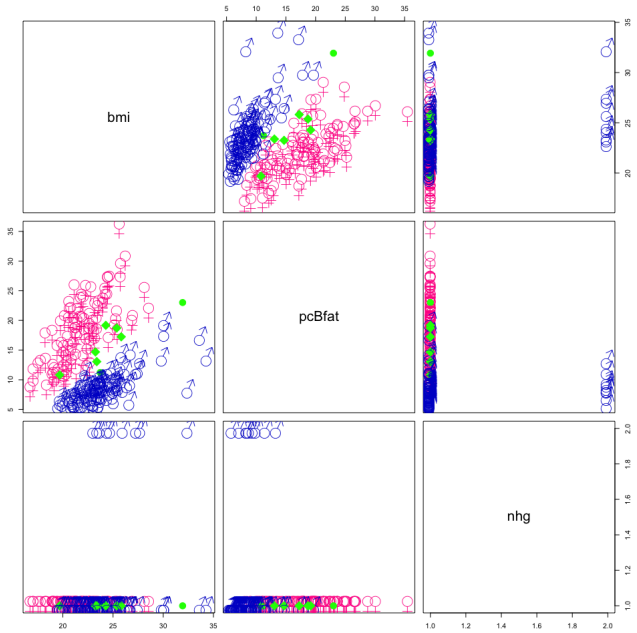
We fit our model (MMSM) to see whether it can distinguish between male and female athletes. We compare this to the model proposed by Browne and McNicholas (2012), fit to the same data. Both model were fit with number of component $G = 1, ..., 5$ and latent factor $q = 1, 2$. Based on the highest BIC (-2333.070), our model chooses the correct number of groups $G = 2$ and has ARI $= 0.847$. The best fit via Browne and McNicholas (2012) model gives three components with BIC $= -2338.973$ and ARI $= 0.701$. Hence, our model performs better on these data than the approach of Browne and McNicholas (2012). The classification performance of each of these models is shown in Table 3.4. Figure 3.2 represents the clustering results for the MMGM (a) and MMSM (b) mixture models on the AIS data; the blue and pink colours highlight the predicted group memberships for the male and female athletes. We can see that there are more mislabelled points in the MMGM that in MMSM.

Table 3.4: Clustering results for the chosen MMSM and MMGM models for the AIS data.

	MMSM		**MMGM**		
	1	2	1	2	3
Female	98	2	1	92	7
Male	6	96	83	5	14



(a)



(b)

Figure 3.2: The AIS data with predicted group memberships by (a) MMGM , (b) MMSM.

3.6 Summary

A mixture approach for mixed-type data has been introduced where the continuous variables are assumed to jointly follow a skewed-t distribution.This model was developed based on the factor analysis model and latent trait model. Two simulation studies were conducted to assess the performance of our model. The mixture components in the first simulation were far whereas in the second they were close and overlapped. Model fitting and parameter estimation was carried via the EM algorithm. The E-step calculation was carried out using numerical integration due to complexity of the model and there is no closed form for all the integrations.

We applied our model to AIS data to distinguish between male and female athletes and compared our model performance with the approach of Browne and McNicholas (2012). Our model performed better than the Browne and McNicholas (2012) model and chose the correct group.

Chapter 4

Mixture for Contaminated Mixed-type Data

4.1 Introduction

Real data is usually considered contaminated data. The contamination can be in the form of outliers, noise, and/or extreme points. It is one of the challenging issues in statistical learning methods, especially for model-based clustering methods. The presence of outliers can effect parameter estimation, lead to overfitting by increasing the number of groups, and/or lead to misclassifications.

Atypical or "outlier" points can be mild or gross atypical observations (Ritter, 2014). Mild outliers are far from, or sampled from a different population than, the assumed model and we can model them by using flexible or heavy-tailed distribution such as the t-distribution. On the other hand, gross outliers are points that are far away from any of the clusters and cannot be modelled by any distribution. In the presence of gross outliers, trimming those observations is one way to handle them.

There has been a great amount of work to handle gross outliers in cluster analysis, e.g.,Cuesta-Albertos *et al.* (1997), Garcia-Escudero and Gordaliza (1999), Gallegos and Ritter (2005), and Ruwet *et al.* (2013).

4.2 Contaminated Gaussian Distribution

The Gaussian scale mixture model is a unimodal, and elliptically symmetric model with heavy tails (Watanabe and Yamaguchi, 2003):

$$\int_0^\infty f_{\text{MN}}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Delta}/w) dH(w), \tag{4.1}$$

where $f_{\text{MN}}(\cdot)$ is density function of multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Delta}/w$ and $H(\cdot)$ is a probability density (or mass) function. We can drive different heavy-tailed distribution by changing the distribution of W. When W follows a gamma distribution with scale and shape parameters $\nu/2$, we obtain t distribution.

Now, let W a dichotomous random variable with probability mass function

$$h(w; \alpha, \eta) = \alpha^{\frac{w-1/\eta}{1-1/\eta}} (1-\alpha)^{\frac{1-w}{1-1/\eta}}, \tag{4.2}$$

where

$$W = \begin{cases} 1 & \text{with probability} \quad \alpha, \\ \frac{1}{\eta} & \text{with probability} \quad (1-\alpha). \end{cases}$$

Then substituting 4.2 in 4.1, we obtain contaminated Gaussian distribution. A m-variate random vector $\mathbf{X}$ from contaminated Gaussian distribution will has a probability density function (pdf) of the form

$$f_{\text{MCN}}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Delta}, \alpha, \eta) = \alpha f_{\text{MN}}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Delta}) + (1-\alpha) f_{\text{MN}}(\mathbf{x}; \boldsymbol{\mu}, \eta\boldsymbol{\Delta}) \tag{4.3}$$

where $f_{\text{MN}}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Delta}) \sim \mathcal{N}_m(\boldsymbol{\mu}, \boldsymbol{\Delta})$, $f_{\text{MN}}(\mathbf{x}; \boldsymbol{\mu}, \eta\boldsymbol{\Delta}) \sim \mathcal{N}_m(\boldsymbol{\mu}, \boldsymbol{\eta\Delta})$, $\alpha \in (0.5, 1)$ represent the proportion of good points and $\eta > 1$ is the degree of contamination (Tukey, 1960). From the pdf of contaminated Gaussian distribution, we can see that this is a Gaussian mixture with two components. Where the first represent the good component with prior probability α and the second is for the "bad" component with $(1-\alpha)$ prior probability. When α and η tend to one, 4.3 becomes the multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and $m \times m$ covariance matrix $\boldsymbol{\Delta}$.

The contaminated Gaussian distribution not only model a data with atypical points but also able to detect them. Once the parameters $\boldsymbol{\mu}, \boldsymbol{\Delta}, \alpha, \eta$ are estimated, we can calculate *a posteriori* probability of a generic observation x_i to determine whether x_i is typical or atypical point. The *a posteriori* probability can be calculated from the following

$$P(x_i \text{ is good}|\hat{\boldsymbol{\vartheta}}) = \frac{\hat{\alpha} f_{\text{MN}}(\mathbf{x}; \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Delta}})}{f_{\text{MCN}}(\mathbf{x}; \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Delta}}, \hat{\alpha}, \hat{\eta})}. \tag{4.4}$$

If the above probability is > 0.5, then x_i is considered typical observation.

4.3 Contaminated Gaussian Factor Analyzers

A t-factor analysis model proposed by McLachlan *et al.* (2007) to model data with atypical observations; however, it can not detect the bad points. A contaminated factor analyzers introduced by Punzo and McNicholas (2014). This model is an extinction of the well known Gaussian factor analysis model (see 2.2.3) that helps to capture and identify outliers in a given data and improve the robustness. Herein, the contaminated Gaussian factor assumes that the joint distribution of a m-dimensional observed variables $X_1, \ldots, X_n$, and q-dimensional latent variables $\mathbf{Y}_i$ is

$$\begin{pmatrix} \mathbf{X}_i \\ \mathbf{Y_i} \end{pmatrix} \sim \mathcal{CN}_{m+q} \left(\begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi} & \boldsymbol{\Lambda} \\ \boldsymbol{\Lambda}' & \mathbf{I}_q \end{bmatrix}, \alpha, \eta \right), \tag{4.5}$$

where $\mathcal{CN}_{m+q}(\cdot)$ denote multivariate contaminated Gaussian distribution, $\boldsymbol{\Lambda}$ is the factor loading with dimension $m \times q$, $\boldsymbol{\Psi}$ is $m \times m$ a diagonal matrix, and $\mathbf{I}_q$ is $q \times q$ identity matrix.

Now, using representation of the contaminated Gaussian distribution that is discussed in Section 4.2, we can write the joint pdf (4.5) given $W_i = w_i$ as

$$\mathbf{X}_i, \mathbf{Y}_i \mid W_i = w_i \sim \mathcal{N}_{m+q}(\mu^\star, \Delta^\star / w_i), \tag{4.6}$$

where

$$\boldsymbol{\mu}^\star = \begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{0} \end{bmatrix},$$

$$\Delta^{\star} = \begin{bmatrix} \mathbf{\Lambda}\mathbf{\Lambda}^{'} + \mathbf{\Psi} & \mathbf{\Lambda} \\ \mathbf{\Lambda}^{'} & \mathbf{I}_q \end{bmatrix},$$

and $W_i \sim \mathcal{C}(\alpha, \eta)$. Then,

$$\boldsymbol{X}_i | w_i \sim \mathcal{N}_m(\boldsymbol{\mu}, (\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi})/w_i),$$

$$\boldsymbol{Y}_i | w_i \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{I}_q/w_i),$$

$$\boldsymbol{\epsilon}_i | w_i \sim \mathcal{N}_m(\mathbf{0}, \mathbf{\Psi}/w_i),$$

and so

$$\boldsymbol{X}_i \sim \mathcal{CN}_m(\boldsymbol{\mu}, (\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}), \alpha, \eta),$$

$$\boldsymbol{Y}_i \sim \mathcal{CN}_q(\mathbf{0}, \boldsymbol{I}_q, \alpha, \eta),$$

$$\boldsymbol{\epsilon}_i \sim \mathcal{CN}_m(\mathbf{0}, \mathbf{\Psi}, \alpha, \eta).$$

Unlike the usual Gaussian factor analysis, the independence between the factor $\mathbf{Y}_i$ and the error $\boldsymbol{\epsilon}_i$ terms no longer holds; however, they remain unconditionally uncorrelated. It can easily show that uncorrelated relation from eq:contfa by conditioning on w.

4.4 Mixture Model for Contaminated Mixed-type Data

4.4.1 Model

In this section, we propose a finite mixture of contaminated mixed-type data as modification to the approach of Browne and McNicholas (2012) to detect the occurrence of the atypical points. We developed our model by following the approach that is used by (Punzo and McNicholas, 2016). Analogously, we used the contaminated normal distribution along with two types of latent models to identify the bad points.

Now, assume we have mixed-type data with i rows and p columns. Without loss of generality, let the first c columns be the categorical variables, and the remaining $p - c$ columns be continuous variables. For p-dimensional observed variables $\mathbf{X}_i = (X_{i1}, ..., X_{ip})'$ there exists q-dimensional latent variables $\mathbf{Y}_i$ with $q < p$. In our model, we assume that the observed variables are independent given the latent variables. Furthermore, let W_i be as defined in (4.2) be another latent variable. Then, $\mathbf{Y}_i \mid w_i \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I}_q/wi)$.

If the manifest variable X_{ij} is categorical with levels $0, 1, 2, \ldots K_j - 1$, then the conditional distribution is Bernoulli with success probability

$$g_1(x_{ij} = k|\mathbf{y}_i, w_i, \boldsymbol{\theta}_j) = \frac{\prod_{k=0}^{K_j-1}[\exp\{\beta_{jk} + \boldsymbol{\tau}'_{jk}\mathbf{y}_i\}]^{x_{ij}(k)}}{1 + \sum_{k=0}^{K_j-1}\exp\{\beta_{jk} + \boldsymbol{\tau}'_{jk}\mathbf{y}_i\}}, \tag{4.7}$$

where $x_{ij}(k) = 1$ if x_{ij} is in category k and 0 otherwise.

If the manifest variables X_{ij} are continuous, we can use a factor analysis representation to write

$$\boldsymbol{X}_i = \mu + \boldsymbol{\Lambda}\boldsymbol{Y}_i + \boldsymbol{\epsilon}_i \tag{4.8}$$

where $\boldsymbol{\Lambda}$ is a matrix of factor loadings, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$, $\mathbf{Y}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_q)$, and $\boldsymbol{\Psi} = \text{diag}(\psi_1, ..., \psi_p)$.

Now, using the assumption that $W_i \sim \mathcal{C}(\alpha, \eta)$, it follows that

$$\boldsymbol{X}_i | w_i \sim \mathcal{N}(\boldsymbol{\mu}, (\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})/w_i)$$

and hence $\boldsymbol{X}_i \sim \mathcal{CN}_p(\boldsymbol{\mu}, (\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}), \alpha, \eta)$. Then the conditional density of X_{ij} given the latent variables can be written as

$$g_1(x_{ij} = k | \mathbf{y}_i, w_i, \boldsymbol{\theta}_j) = \boldsymbol{X}_i | y_i, w_i \sim \mathcal{N}(\boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{y}_i, \boldsymbol{\Psi}/w_i). \tag{4.9}$$

Combining both density functions from (4.7) and (4.9), the density in (2.1) can be extended as follows:

$$f(\mathbf{x}_i) = \int_{\mathbf{y}} \sum_{w} \prod_{j=1}^{p} g_1(x_{ij} | \mathbf{y}_i, w_i, \boldsymbol{\theta}_j) g_2(\mathbf{y}_i | w_i) h(w_i) d\mathbf{y}_i. \tag{4.10}$$

The mixture model for contaminated mixed-type data (MMCM) has density of the form

$$f(\boldsymbol{x} | \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g \left[\int_{\mathbf{y}} \sum_{w} \prod_{j=1}^{p} g_1(x_{ij} | \mathbf{y}_{ig}, w_{ig}, \boldsymbol{\theta}_{jg}) g_2(\mathbf{y}_{ig} | w_{ig}) h(w_{ig}) d\mathbf{y}_{ig} \right], \tag{4.11}$$

where $\boldsymbol{\vartheta}$ denotes all model parameters.

4.4.2 Likelihoods

Using a linear transformation of W, let the indicator variable V be given by

$$V = \frac{w - 1/\eta}{1 - 1/\eta}$$

and, therefore, the density of W in (4.2) can be written as

$$h(v, \boldsymbol{\theta}) = \alpha^v (1 - \alpha)^{(1-v)} \tag{4.12}$$

$$V = \begin{cases} 1 & \text{with probability} \quad \alpha, \\ 0 & \text{with probability} \quad (1 - \alpha). \end{cases}$$

Thus, the observed log-likelihood for the MMCM model is

$$l(\boldsymbol{\vartheta}|\mathbf{x}_1, ..., \mathbf{x}_n) = \sum_{i=1}^{n} \log \left\{ \sum_{g=1}^{G} \pi_g \left[\int_{\mathbf{y}} \sum_{v} \prod_{j=1}^{p} g_1(x_{ij}|\mathbf{y}_i, v_{ig}, \boldsymbol{\theta}_j) g_2(\mathbf{y}_{ig}|v_{ig}) h(v_{ig}) d\mathbf{y}_{ig} \right] \right\}.$$

In the MMCM model, there are three sources of incomplete data: the group memberships, the classification of the type of observation, and the latent variables. We will use $\mathbf{z}_i$ to donate group memberships, and $\mathbf{v}_i$ for the type of each observation so that if $v_{ig} = 1$ if observation i in component g is typical and $v_{ig} = 0$ if observation i in component g is atypical. Therefore, the complete-data log-likelihood for the MMCM model is given by

$$l_c(\boldsymbol{\vartheta}) = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 + \mathcal{L}_4 + \mathcal{L}_5, \tag{4.13}$$

where

$$\mathcal{L}_1 = \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\log\pi_g,$$

$$\mathcal{L}_2 = \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\left[v_{ig}\log\alpha_g + (1-v_{ig})\log(1-\alpha_g)\right],$$

$$\mathcal{L}_3 = \sum_{i=1}^{n}\sum_{g=1}^{G}\sum_{j=1}^{c} z_{ig}\left[\log\left(\prod_{k=0}^{K_j-1}\left[\exp\{\beta_{jkg}+\boldsymbol{\tau}'_{jkg}\mathbf{y}_{ig}\}\right]^{x_{ij}(k)}\right) - \log\left(1+\sum_{k=0}^{K_j-1}\exp\{\beta_{jkg}+\boldsymbol{\tau}'_{jkg}\mathbf{y}_{ig}\}\right)\right],$$

$$\begin{aligned}\mathcal{L}_4 = C &- \frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G} p(1-v_{ig})\log\eta_g - \frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\log|\boldsymbol{\Psi}_g| \\ &- \frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\left(v_{ig}+\frac{1-v_{ig}}{\eta_g}\right)\left[\mathrm{tr}\left\{(\mathbf{x}_i-\boldsymbol{\mu}_g)'\boldsymbol{\Psi}_g^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_g)\right\}\right. \\ &\left.-2\mathrm{tr}\left\{(\mathbf{x}_i-\boldsymbol{\mu}_g)'\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}_g\mathbf{y}_{ig}\right\} + \mathrm{tr}\left\{\boldsymbol{\Lambda}_g\mathbf{y}_{ig}\mathbf{y}'_{ig}\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}'_g\right\}\right],\end{aligned}$$

$$\mathcal{L}_5 = \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\left\{C - \frac{q}{2}(1-v_{ig})\log\eta_g - \frac{\left(v_{ig}+\frac{1-v_{ig}}{\eta_g}\right)}{2}\mathbf{y}_{ig}\mathbf{y}'_{ig}\right\}.$$

4.4.3 Parameter Estimation

Parameter estimation is carried out within the EM algorithm as described below.

1. **Initialization**: Initialize the mode parameters $\boldsymbol{\mu}_g, \alpha_g, \eta_g, \boldsymbol{\Lambda}_g\boldsymbol{\Psi}_g, \beta_g,$ and $\boldsymbol{\tau}_g$.

2. **E Step**:

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

(b) $\mathbb{E}[V_{ig}|\mathbf{x}_i, Z_{ig}=1]$

$$= \frac{\alpha_g \int_{\mathbf{y_{ig}}} \prod_{j=1}^{p} g_1(\mathbf{x}_i|\mathbf{y}_i, v_{ig}=1, \boldsymbol{\theta}_{jg}) g_2(\mathbf{y}_{ig}|v_{ig}=1) d\mathbf{y}_{ig}}{f(\mathbf{x}_i|\boldsymbol{\theta}_g)} =: \hat{v}_{ig}.$$

(c) $\mathbb{E}[Z_{ig}\left(V_{ig} + \frac{1-V_{ig}}{\eta_g}\right)\mathbf{Y}_{ig}|\mathbf{x}_i]$

$$= \int_y z_{ig}\left(v_{ig} + \frac{1-v_{ig}}{\eta_g}\right)\mathbf{y}_{ig}\frac{f(\mathbf{x}_i, \mathbf{y}_{ig})}{f(\mathbf{x}_i)}d\mathbf{y}_{ig} =: \boldsymbol{e}_{1ig}.$$

(d) $\mathbb{E}[Z_{ig}\left(V_{ig} + \frac{1-V_{ig}}{\eta_g}\right)\mathbf{Y}_{ig}\mathbf{Y}'_{ig}|\mathbf{x}_i]$

$$= \int_y z_{ig}\left(v_{ig} + \frac{1-v_{ig}}{\eta_g}\right)\mathbf{y}_{ig}\mathbf{y}'_{ig}\frac{f(\mathbf{x}_i, \mathbf{y}_{ig})}{f(\mathbf{x}_i)}d\mathbf{y}_{ig} =: \boldsymbol{E}_{2ig}.$$

where

$$\begin{aligned} f(\mathbf{x}_i, \mathbf{y}_{ig}) &= \sum_{v}\prod_{j=1}^{p} g_1(x_{ij}|\mathbf{y}_{ig}, v_{ig}, \boldsymbol{\theta}_{jg}) g_2(\mathbf{y}_{ig}|v_{ig}) h(v_{ig}) \\ &= \alpha_g \prod_{j=1}^{p} g_1(\mathbf{x}_i|\mathbf{y}_{ig}, v_{ig}=1, \boldsymbol{\theta}_{jg}) g_2(\mathbf{y}_{ig}|v_{ig}=1) \\ &\quad + (1-\alpha_g)\prod_{j=1}^{p} g_1(\mathbf{x}_i|\mathbf{y}_{ig}, v_{ig}=0, \boldsymbol{\theta}_{jg}) g_2(\mathbf{y}_{ig}|v_{ig}=0), \end{aligned}$$

$$\begin{aligned} f(\mathbf{x}_i|\boldsymbol{\vartheta}) &= \alpha_g \int_{\mathbf{y_{ig}}} \prod_{j=1}^{p} g_1(\mathbf{x}_i|\mathbf{y}_{ig}, v_{ig}=1, \boldsymbol{\theta}_{jg}) g_2(\mathbf{y}_{ig}|v_{ig}=1) d\mathbf{y}_{ig} \\ &\quad + (1-\alpha_g) \int_{\mathbf{y_{ig}}} \prod_{j=1}^{p} g_1(\mathbf{x}_i|\mathbf{y}_{ig}, v_{ig}=0, \boldsymbol{\theta}_{jg}) g_2(\mathbf{y}_{ig}|v_{ig}=0) d\mathbf{y}_{ig}. \end{aligned}$$

All the expectations are calculated and solved numerically since the integrals

do not have a closed form when the data are mixed-type.

3. **M step**: Maximizing the exceptions in step 2.

$$\widehat{\pi}_g = \frac{\sum_{i=1}^{n} \widehat{z}_{ig}}{n},$$

$$\widehat{\alpha}_g = \max\left\{\alpha^*, \frac{\sum_{i=1}^{n} \widehat{z}_{ig}\widehat{v}_{ig}}{\sum_{i=1}^{n} \widehat{z}_{ig}}\right\}, \quad \alpha^* = 0.5,$$

The updated for the parameters $(\beta_g, \boldsymbol{\tau}_g)$ can be calculated by solving the following non-linear system of equation:

$$\sum_{i=1}^{n} \hat{z}_{ig}\left\{\int_{\mathbf{y}_{ig}} (1, \mathbf{y}_{ig})[x_{ij}(k) - g_1(x_{ij} = k|\mathbf{y}_{ig}, \boldsymbol{\theta}_{jg})]h(\mathbf{y}_{ig}|\mathbf{x}_i)d\mathbf{y}_{ig}\right\},$$

$$\widehat{\boldsymbol{\mu}}_g = \frac{\sum_{i=1}^{n}\left(\hat{v}_{ig} + \frac{1-\hat{v}_{ig}}{\hat{\eta}_g}\right)\mathbf{x}_i\hat{z}_{ig}}{\sum_{i=1}^{n}\left(\hat{v}_{ig} + \frac{1-\hat{v}_{ig}}{\hat{\eta}_g}\right)\hat{z}_{ig}},$$

$$\hat{\boldsymbol{\Lambda}}_g = \left\{\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)\boldsymbol{e}_{1ig}^{'}\right\}\left\{\sum_{i=1}^{n}\boldsymbol{E}_{2ig}\right\}^{-1},$$

$$\hat{\boldsymbol{\Psi}}_g = \frac{1}{\sum_{i=1}^{n}\hat{z}_{ig}}\text{dig}\left\{\sum_{i=1}^{n}\left[\hat{z}_{ig}\left(\hat{v}_{ig} + \frac{1-\hat{v}_{ig}}{\hat{\eta}_g}\right)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)^{'}\right.\right.$$
$$\left.\left. -2(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)\boldsymbol{e}_{1ig}^{'}\hat{\boldsymbol{\Lambda}}_g^{'} + \hat{\boldsymbol{\Lambda}}_g\boldsymbol{E}_{2ig}\hat{\boldsymbol{\Lambda}}_g^{'}\right]\right\},$$

$$\hat{\eta}_g = \max\left\{\eta^*, \frac{\sum_{i=1}^{n}\hat{z}_{ig}(1-\hat{v}_{ig})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_g)^{'}(\hat{\boldsymbol{\Psi}}_g)^{-1}(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_g)}{p\sum_{i=1}^{n}\hat{z}_{ig}(1-\hat{v}_{ig})}\right\}, \quad \eta^* = 1.001.$$

4. **Check for convergence**: If converged, stop; otherwise, repeat steps 2–3 until convergence.

4.4.4 Computational Considerations

Numerical Integration and Initialization

As mentioned in the previous section, all the integrals can not be simplified and written in closed form due to the complexity of the model. We used numerical integration methods similar to the one we outlined in Section 3.3.2.

The starting values play an important role when we use the EM algorithm. The choice of initialization can help to speed-up convergence and obtain a closer value to the true parameter. The initial values for the group membership z_{ig} are initialized by generating uniform random numbers over the interval (0,1) and then standardizing to ensure $\sum_{g=1}^{G} z_{ig} = 1$. We then use z_{ig} to calculate the mean component $\boldsymbol{\mu}_g$. The initial values for $\boldsymbol{\Lambda}_g$ and $\boldsymbol{\Psi}_g$ as well as the categorical variable parameters, so that $\beta_g = 0, \boldsymbol{\tau}_g = \mathbf{0}$, are discussed in Section 3.3.2. For the contamination parameters α and η, we follow the initialization recommendation from Punzo and McNicholas (2016). At iteration $t = 0$, $\alpha_g^0 = 0.999$ and $\eta_g^0 = 1.001$, where $g = 1, 2, ..., G$.

Number of Free Parameters

The MMCM has m free parameters that can be calculated from the following

$$(G-1) + Gp_1 + G\left[p_1 q - \frac{1}{2}q(q-1)\right] + Gp_1 + 2G + G\left[p_2 q - \frac{1}{2}q(q-1)\right] + Gp_2,$$

where p_1 is the number of continues variables and p_2 denotes the categorical variables, and G is the number of clusters.

4.5 Simulation

Five simulation cases were considered to assess the performance and the behaviour of the MMCM model. We simulate two components from each case then we fit the MMCM model. The sample size that we used is $n = 50, 100, 200$ per group to assess the effect of the sample size in each simulation study.

1. Each component from contaminated normal distribution with $\alpha_1 = 0.90, \alpha_2 = 0.80$, and $\eta_1 = 20, \eta_2 = 30$.

2. Both clusters are normally distributed.

3. Two t-distributed groups with degree of freedom $\nu_1 = 4, \nu_2 = 60$.

4. Gaussian clusters with 5% of the data are replaced with noise generated from uniform $(-15, 15)$.

5. Gaussian clusters with a perturbed observation. One observation from each group is randomly selected then a constant $c \in \{4, 8, 12, 16\}$ is added to that observation.

The common parameters in all simulation studies cases are:

$$\pi_1 = 0.5,\ \boldsymbol{\mu}_1 = \begin{bmatrix} -5 \\ -5 \end{bmatrix},\ \boldsymbol{\Sigma}_1, = \begin{bmatrix} 0.48 & -0.17 \\ -0.17 & 1.14 \end{bmatrix},\ \beta_1 = 1, \tau_1 = 0$$

$$\pi_2 = 0.5,\ \boldsymbol{\mu}_1 = \begin{bmatrix} 5 \\ 5 \end{bmatrix}\ \boldsymbol{\Sigma}_2, = \begin{bmatrix} 1.39 & 0.27 \\ 0.27 & 0.16 \end{bmatrix},\ \beta_2 = -1,\ \tau_2 = 0$$

Figure 4.1 is pair plots for one of the 25 simulated data sets from each simulation scenario, where the blue and red points represent the typical points and the green with alien shape are atypical points. In all simulation study, we fit MMCM and MMGM model with different latent $q = 1, 2$, and $G = 1, 2, 3, 4$. Then we use BIC and ARI to compare the performance of our model with MMGM model.

4.5.1 Simulation Results

A general pattern we noticed among all simulation scenarios is that sample size indeed has an effect on the accuracy of parameter estimation and the ARI values. There is a positive relationship between the sample size and parameter estimations and with the ARI; as the sample size increases, the value of estimations and ARI improve.

Table 4.1 displays results for scenario 1, where each group has atypical points. Herein, we noticed that the MMCM performed better than MMGM, and the parameter estimates are closer to the true parameters. Not only the values of the parameters are better when we fit MMCM, but also the BIC and ARI values are better, as shown in Table 4.2.

The results under scenario 2, where the first table is a summary of parameter estimates and the second table contains the average of BIC and ARI values under MMCM and MMGM are reported in Tables 4.3 and 4.4. As expected, both models did well in this scenario; however, the MMCM has slightly higher mean ARI than MMGM. This indicates that when the data follow a Gaussian distribution, the MMCM can do same as or better than MMGM.

Table 4.5 reports the results for scenario 3 when both clusters are generated using t-distribution. From Table 4.6, the results are not too surprising that MMGM has the

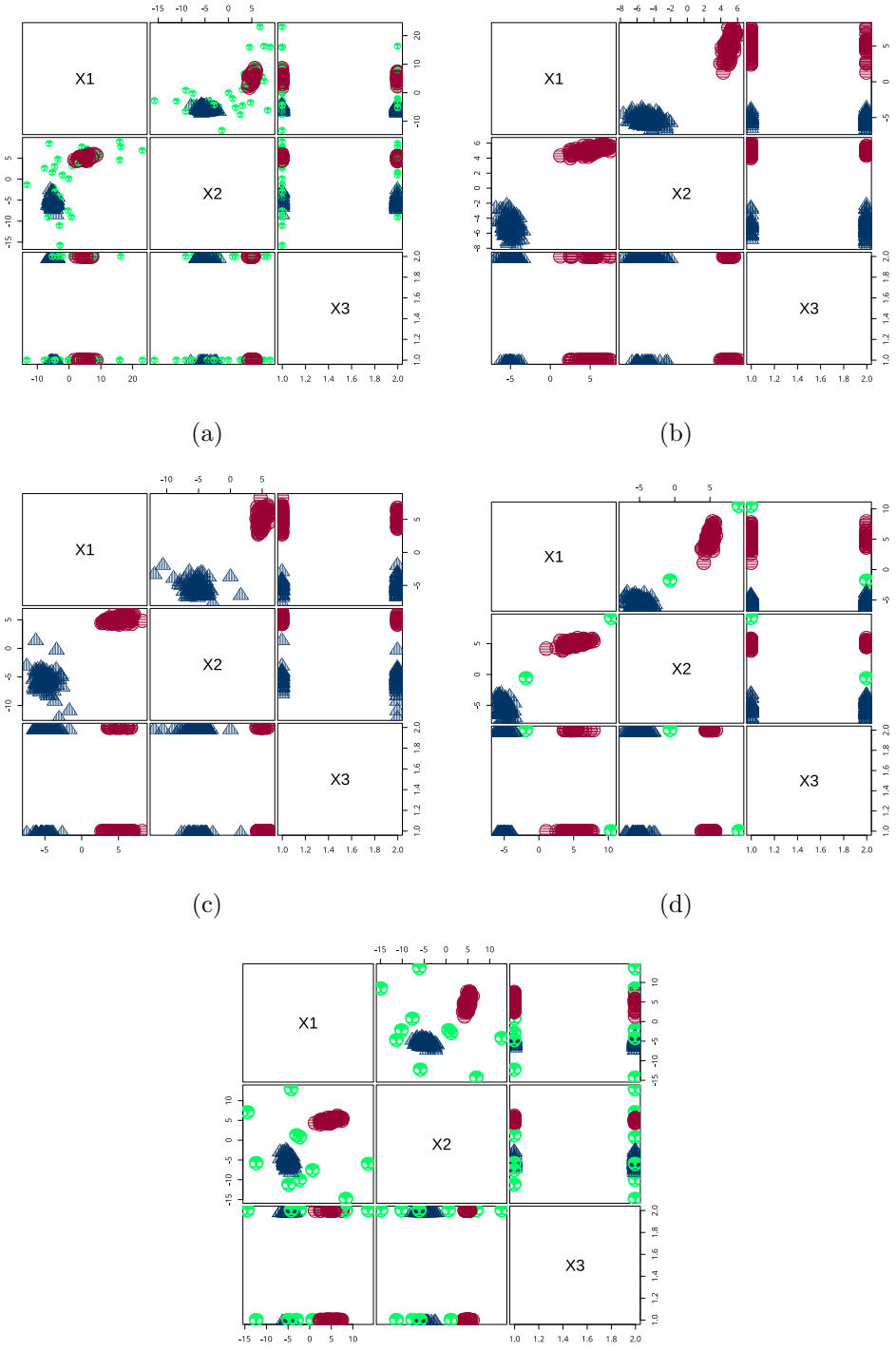


Figure 4.1: Example of one of the simulated data sets from (a) scenario 1, (b) scenario 2, (c) scenario 3, (d) scenario 4, and (e) scenario 5.

Table 4.1: The means and the standard deviations from simulation scenario 1.

	Parameters	MMCM		MMGM	
		Means	Standard deviations	Means	Standard deviations
$n = 50$	$\boldsymbol{\mu}_1$	$(-5.0393, -4.9681)'$	$(0.1173, 0.1741)'$	$(-4.9983, -4.9230)'$	$(0.1988, 0.2372)'$
	$\boldsymbol{\mu}_2$	$(5.0525, 5.0059)'$	$(0.2311, 0.0952)'$	$(5.2107, 5.0349)'$	$(0.3984, 0.1473)'$
	β_1	0.9413	0.3464	0.8651	0.2679
	β_2	-1.1242	0.3753	-1.1357	0.3316
	τ_1	-0.3100	0.4773	-0.3094	0.2914
	τ_1	-0.1610	0.5956	-0.1622	0.4161
$n = 100$	$\boldsymbol{\mu}_1$	$(-4.9780, -4.9719)'$	$(0.0748, 0.1145)'$	$(-4.9880, -4.9679)'$	$(0.1068794, 0.1946200)'$
	$\boldsymbol{\mu}_2$	$(5.0309, 5.0047)'$	$(0.1438, 0.047)'$	$(5.0905, 5.0336)'$	$(0.2912813, 0.0962062)'$
	β_1	1.10343	0.2292	1.0438	0.2028
	β_2	-1.0852	0.5465	-0.9827	0.1810
	τ_1	-0.2182	1.0100	-0.4140	0.2190
	τ_1	-0.0228	0.5574	-0.0376	0.2990
$n = 200$	$\boldsymbol{\mu}_1$	$(-4.9987, -5.0144)'$	$(0.0388, 0.0676)'$	$(-4.9843, -5.0093)'$	$(0.0916, 0.1854)'$
	$\boldsymbol{\mu}_2$	$(4.9682, 4.9854)'$	$(0.1151, 0.0401)'$	$(4.8988, 4.9903)'$	$(0.2042, 0.0726)'$
	β_1	0.9951	0.2241	1.0114	0.1799
	β_2	-1.0079	0.2230	-1.0089	0.1889
	τ_1	-0.1893	0.9389	-0.4649	0.2959
	τ_1	-0.0212	0.4728	-0.0259	0.2468

Table 4.2: The average of ARI, BIC values for the MMCM and MMGM models on contaminated Gaussian clusters.

n	MMCM		MMGM	
	$\overline{\text{ARI}}$	$\overline{\text{BIC}}$	$\overline{\text{ARI}}$	$\overline{\text{BIC}}$
50	0.9615	-1101.8730	0.8215	-1155.6451
100	0.9604	-2009.0872	0.8216	-2074.6815
200	0.9657	-3813.4944	0.7858	-3936.4231

Table 4.3: The means and the standard deviations from simulation scenario 2.

	Parameters	MMCM		MMGM	
		Means	Standard deviations	Means	Standard deviations
$n = 50$	$\boldsymbol{\mu}_1$	$(-4.9856 - 5.0501)'$	$(0.1024, 0.1586)'$	$(-4.9854, -5.0503)'$	$(0.1023, 0.1586)'$
	$\boldsymbol{\mu}_2$	$(5.0107, 5.0005)'$	$(0.2220, 0.0585)'$	$(5.0101, 5.0005)'$	$(0.2215, 0.0585)'$
	β_1	1.0169	0.3313	0.9904	0.3436
	β_2	-1.0253	0.3838	-1.0213	0.3827
	τ_1	-0.4715	0.3244	-0.3108	0.2189
	τ_1	-0.1926	0.2947	-0.1807	0.2767
$n = 100$	$\boldsymbol{\mu}_1$	$(-4.9999, -5.0101)'$	$(0.0665, 0.1224)'$	$(-4.9999, -5.0104)'$	$(0.0665, 0.1225)'$
	$\boldsymbol{\mu}_2$	$(5.04774.9965)'$	$(0.1132, 0.0393)'$	$(5.0479, 4.9966)'$	$(0.1129, 0.0392)'$
	β_1	1.0527	0.2634	1.0479	0.2606
	β_2	-0.9939	0.2066	-0.9864	0.2051
	τ_1	-0.3691	0.1466	-0.3386	0.1334
	τ_1	-0.3725	0.1727	-0.3073	0.1817
$n = 200$	$\boldsymbol{\mu}_1$	$(-4.9931, -5.0074)'$	$(0.0603, 0.066)'$	$(-5.0022, -5.01681)'$	$(0.0484, 0.0547)'$
	$\boldsymbol{\mu}_2$	$(4.9689, 4.9723)'$	$(0.1367, 0.0898)'$	$(4.9869, 4.9899)'$	$(0.0851, 0.0250)'$
	β_1	1.0136	0.2200	1.0110	0.1988
	β_2	-1.0183	0.1907	-1.0321	0.1632
	τ_1	-0.4825	0.2012	-0.3303	0.1387
	τ_1	-0.1680	0.3352	-0.0907	0.2087

Table 4.4: The average of ARI, BIC values for the MMCM and MMGM models on Gaussian clusters.

n	MMCM		MMGM	
	$\overline{\text{ARI}}$	$\overline{\text{BIC}}$	$\overline{\text{ARI}}$	$\overline{\text{BIC}}$
50	0.9396	-932.0885	0.9322	-944.9336
100	0.9410	-1478.2177	0.9365	-1492.3082
200	0.9718	-3232.9921	0.9661	-3318.6522

best mean BIC considering that the number of free parameters is less than MMCM and one of the components has a degree of freedom of 60 (i.e., effectively a Gaussian component). However, the MMCM has the best mean ARI values indicating that it can correctly assign observation to the correct grope when component or the data are non-Gaussian.

Table 4.5: The true parameters, value the means and the standard deviations from simulation scenario 3.

	Parameters	MMCM		MMGM	
		Means	Standard deviations	Means	Standard deviations
$n = 50$	$\boldsymbol{\mu}_1$	$(-4.9718, -5.0396)'$	$(0.1381, 0.2428)'$	$(-4.9661, -5.03995)'$	$(0.1391, 0.2507)'$
	$\boldsymbol{\mu}_2$	$(4.9832, 5.0036)'$	$(0.2147, 0.0637)'$	$(4.9583, 4.995)'$	$(0.2214, 0.0687)'$
	β_1	1.1393	0.4767	1.1241	0.4638
	β_2	-1.3047	0.3390	-1.2834	0.3171
	τ_1	-0.51873	0.3190	-0.4856	0.3506
	τ_1	-0.3581	0.3817	-0.3292	0.3496
$n = 100$	$\boldsymbol{\mu}_1$	$(-4.8539, -4.8284)'$	$(0.8145, 0.8214)'$	$(-5.0233, -4.9910)'$	$(0.0945, 0.1317)'$
	$\boldsymbol{\mu}_2$	$(4.8307, 4.8231)'$	$(0.8656, 0.8457)'$	$(5.0051, 4.9900)'$	$(0.0952, 0.0712)'$
	β_1	1.1659	0.3809	1.1310	0.2226
	β_2	-1.1286	0.2883	-1.1075	0.3168
	τ_1	-0.7778	0.5896	-0.4001	0.1744
	τ_1	-0.6864	0.6699	-0.2774	0.2689
$n = 200$	$\boldsymbol{\mu}_1$	$(-4.9930, -4.9909)'$	$(0.04953, 0.0967)'$	$(-4.9847, -4.9838)'$	$(0.0630, 0.1031)'$
	$\boldsymbol{\mu}_2$	$(4.9978, 5.0041)'$	$(0.0829, 0.0313)'$	$(4.9946, 4.9930)'$	$(0.0859, 0.0474)'$
	β_1	1.0689	0.2099	1.0346	0.1929
	β_2	-0.9060	0.4704	-1.0061	0.1414
	τ_1	-0.4526	0.1789	-0.3934	0.1774
	τ_1	-0.0017	0.1801	-0.0370	0.2439

Under scenario 4, the effect of how far atypical points are from the rest of typical points in each component can be noticed from the simulation results presented in the Tables 4.7, 4.8, 4.9, and 4.10. We can see the further the point from the component, the values for the ARI gets higher. In all cases in this scenario, we can see that the MMCM also fits better than MMGM.

Table 4.6: The of ARI, BIC values for the MMCM and MMGM models on t-distributed clusters.

n	MMCM		MMGM	
	$\overline{\text{ARI}}$	$\overline{\text{BIC}}$	$\overline{\text{ARI}}$	$\overline{\text{BIC}}$
50	0.9960	-1082	0.8624	-1018
100	0.9371	-1910	0.7071	-1754
200	0.9516	-3468	0.9042	-3370

Table 4.7: The means and the standard deviations from simulation scenario 4(a).

	Parameters	MMCM		MMGM	
		Means	Standard deviations	Means	Standard deviations
$n = 50$	$\boldsymbol{\mu}_1$	$(-4.9800, -5.0431)'$	$(0.1180, 0.1624)'$	$(-4.9140, -4.9780)'$	$(0.11500, 0.15840)'$
	$\boldsymbol{\mu}_2$	$(4.9870, 4.9871)'$	$(0.1682, 0.0652)'$	$(5.0500, 5.0500)'$	$(0.1662, 0.0692)'$
	β_1	0.9707	0.3390	1.0060	0.3530
	β_2	-0.9918	0.3460	-1.0360	0.3837
	τ_1	-0.3393	0.2889	-0.2478	0.4676
	τ_1	-0.2205	0.4221	-0.1570	0.4653
$n = 100$	$\boldsymbol{\mu}_1$	$(-4.9944, -5.0101)'$	$(0.0674, 0.1202)'$	$(-4.9670, -4.9800)'$	$(0.0656, 0.1214)'$
	$\boldsymbol{\mu}_2$	$(5.0112, 5.0100)'$	$(0.0961, 0.0312)'$	$(5.0430, 5.0410)'$	$(0.0951, 0.0300)'$
	β_1	1.0700	0.2631	1.0640	0.2557
	β_2	-0.8873	0.4563	-0.9711	0.2004
	τ_1	-0.405	0.2113	-0.3779	0.2134
	τ_1	-0.2651	0.2236	-0.2252	0.2513
$n = 200$	$\boldsymbol{\mu}_1$	$(-4.9810, -4.9960)'$	$(0.1175, 0.1031)'$	$(-4.9830, -4.9980)'$	$(0.0484, 0.0554)'$
	$\boldsymbol{\mu}_2$	$(4.9710, 4.9731)'$	$(0.1626, 0.1333)'$	$(5.0160, 5.0170)'$	$(0.0802, 0.0250)'$
	β_1	1.032 0	0.2630	1.0360	0.2244
	β_2	-0.9175	0.4562	-1.0280	0.1552
	τ_1	-0.4682	0.1927	-0.4144	0.2367
	τ_1	-0.1496	0.1845	-0.0679	0.1559

Table 4.8: The of ARI, BIC values for the MMCM and MMGM models on perturbed Gaussian clusters-a.

n	MMCM		MMGM	
	$\overline{\text{ARI}}$	$\overline{\text{BIC}}$	$\overline{\text{ARI}}$	$\overline{\text{BIC}}$
50	0.9165	-969	0.9063	-1008
100	0.9761	-1721	0.9380	-1759
200	0.9404	-3268	0.9468	-3202

Table 4.9: The means and the standard deviations from simulation scenario 4(d).

	Parameters	MMCM		MMGM	
		Means	Standard deviations	Means	Standard deviations
$n = 50$	$\boldsymbol{\mu}_1$	$(-4.9990, -5.043)'$	$(0.1435, 0.1352)'$	$(-4.5150, -4.5700)'$	$(0.2998, 0.2824)'$
	$\boldsymbol{\mu}_2$	$(5.0840, 5.0250)'$	$(0.3919, 0.1032)'$	$(5.2020, 5.1940)'$	$(0.2364, 0.1654)'$
	β_1	0.8962	3.1180	0.9800	0.3534
	β_2	-1.0203	3.4850	-1.029	0.4004
	τ_1	-0.2165	3.0400	-0.3213	0.2708
	τ_1	-0.2202	2.8460	0.1123	0.5657
$n = 100$	$\boldsymbol{\mu}_1$	$(-4.9770, -4.9940)'$	$(0.0894, 0.1285)'$	$(-4.7470, -4.7590)'$	$(0.1440, 0.1848)'$
	$\boldsymbol{\mu}_2$	$(5.0090, 5.0080)'$	$(0.1021, 0.0307)'$	$(5.1070, 5.1040)'$	$(0.1270, 0.0888)'$
	β_1	1.0860	0.2719	1.088	0.2477
	β_2	-0.9638	0.5462	-0.9549	0.2209
	τ_1	-0.1597	0.5139	-0.3708	0.1978
	τ_1	0.0305	0.7832	0.0315	0.3433
$n = 200$	$\boldsymbol{\mu}_1$	$(-4.9830, -4.9980)'$	$(0.0500, 0.0622)'$	$(-4.8730, -4.8870)'$	$(0.0848, 0.0794)'$
	$\boldsymbol{\mu}_2$	$(4.9988, 4.9980)'$	$(0.0787, 0.0252)'$	$(5.0330, 5.0320)'$	$(0.0863, 0.0460)'$
	β_1	1.0520	0.2174	1.0640	0.2152
	β_2	-0.9399	0.4729	-1.020	0.1509
	τ_1	-0.2012	0.4300	-0.4442	0.1744
	τ_1	-0.0809	0.2825	-0.0926	0.1844

Table 4.10: The of ARI, BIC values for the MMCM and MMGM models on perturbed Gaussian clusters-d.

n	MMCM		MMGM	
	$\overline{\text{ARI}}$	$\overline{\text{BIC}}$	$\overline{\text{ARI}}$	$\overline{\text{BIC}}$
50	0.9497	-987.5	0.9367	-1052
100	0.9594	-1730	0.9440	-1985
200	0.9600	-3159	0.9522	-3590

The final simulation scenario results are reported in Table 4.11 and 4.12. Looking at those tables, we draw a similar conclusion to scenario 1, where the MMCM performs better than MMGM.

Table 4.11: The means and the standard deviations from simulation scenario 5.

	Parameters	MMCM		MMGM	
		Means	Standard deviations	Means	Standard deviations
$n = 50$	$\boldsymbol{\mu}_1$	$(-4.9780, -5.0540)'$	$(0.0998, 0.1490)'$	$(-4.7550, -4.9950)'$	$(0.6251, 0.3487)'$
	$\boldsymbol{\mu}_2$	$(5.0130, 4.9990)'$	$(0.2329, 0.0528)'$	$(4.9730, 4.930)'$	$(0.6560, 0.2904)'$
	β_1	1.0410	0.3511	0.8963	0.3938
	β_2	-0.9943	0.6421	-0.9388	0.4483
	τ_1	-0.3696	0.4780	-0.3704	0.3558
	τ_1	-0.2671	0.6989	-0.2950	0.2723
$n = 100$	$\boldsymbol{\mu}_1$	$(-5.0000, -4.9980)'$	$(0.0690, 0.1194)'$	$(-4.900, -4.9360)'$	$(0.2134, 0.2308)'$
	$\boldsymbol{\mu}_2$	$(5.0480, 4.9980)'$	$(0.1203, 0.0406)'$	$(5.0170, 4.9180)'$	$(0.2375, 0.2010)'$
	β_1	1.0840	0.2823	0.9972	0.2484
	β_2	-0.9999	0.5344	-0.9514	0.2373
	τ_1	-0.3078	0.4347	-0.3130	0.2351
	τ_1	-0.2047	0.2894	-0.3200	0.2418
$n = 200$	$\boldsymbol{\mu}_1$	$(-5.0010, -5.0070)'$	$(0.0497, 0.0628)'$	$(-4.8740, -4.8900)'$	$(0.6462, 0.6775)'$
	$\boldsymbol{\mu}_2$	$(4.9930, 4.9900)'$	$(0.0782, 0.0251)'$	$(4.9010, 4.9300)'$	$(0.2640, 0.2160)'$
	β_1	1.0400	0.2015	0.9619	0.1967
	β_2	-0.9839	0.6188	-0.9803	0.1938
	τ_1	-0.2111	0.5222	-0.3671	0.2118
	τ_1	-0.2035	0.4901	-0.3874	0.3134

Table 4.13 addresses some properties of contamination parameters η and α along with comparing the performance of MMCM in detecting atypical points in each simulation scenario. It can be noticed that in scenario 1, the η and α are getting closer to the true parameters as the sample size increases. In the second scenario, we can see that η has the lowest values, and α has the largest values among all simulations. This could be interpreted as the absence of the atypical points in the data. The values of η are also affected by how far the atypical points are from the rest of the component observations, i.e., the larger value of η, the further is the point. For example, in scenario

Table 4.12: The of ARI, BIC values for the MMCM and MMGM models on Gaussian with noise clusters.

n	MMCM		MMGM	
	$\overline{\text{ARI}}$	$\overline{\text{BIC}}$	$\overline{\text{ARI}}$	$\overline{\text{BIC}}$
50	0.8865	-1120	0.8713	-1186
100	0.8946	-1847	0.8804	-1934
200	0.8953	-3981	0.8952	-3984

4(a) when we add 4 to a randomly chosen point $\eta_1 = 19.850$ whereas $\eta_1 = 179.78$ in 4(d) when we add 16.

Now, to evaluate the performance of MMCM in detecting atypical points, we calculate the true positive rate and the false positive rate. The first rate is to measure the proportion of atypical points that is correctly detected and labelled as typical points, whereas the second rate measures the proportion of falsely labelled typical or "good" points as atypical points. In scenarios that do not have atypical points, such as the second and the third, the MMCM never mistakenly labelled good points as atypical points. In scenarios where are atypical points in the data, in general the true positive rate increase and the false positive rate decrease as the sample size increase.

4.6 Real Data

Possums in Australia and New Guinea

We use the possum dataset from R package DAAG to illustrate the performance of the MMCM model. The data frame with nine morphometric measurements of $n = 104$ Australia possums that can be found in Southern Victoria or other sites. There are

Table 4.13: The mean values of η, α, rate of correctly detected atypical points and the rate of falsely detected atypical points for MMCM across all simulation scenarios.

	Parameters	Scenario 1	Scenario 2	Scenario 3	Scenario 4				Scenario 5
					a	b	c	d	
$n = 50$	η_1	18.560	1.0020	7.4710	19.850	25.600	113.40	179.78	90.300
	η_2	14.510	2.4040	3.5840	39.460	52.830	139.10	125.68	43.750
	α_1	0.8487	0.9941	0.9529	0.9706	0.9713	0.9873	0.9425	0.9325
	α_2	0.7827	0.9946	0.9950	0.9765	0.9799	0.9793	0.8491	0.9469
	CD	0.7286	-	-	0.7800	0.8400	0.8430	0.9200	0.9619
	FD	0.0105	0.0000	0.0000	0.0033	0.0008	0.0000	0.0000	0.0025
$n = 100$	η_1	19.444	1.0020	6.5510	29.960	63.420	88.150	332.30	97.290
	η_2	26.734	1.002	1.8600	1.0010	41.830	176.95	272.00	74.620
	α_1	0.8532	0.9974	0.9389	0.9830	0.9913	0.9869	0.9868	0.9465
	α_2	0.7881	0.9972	0.9963	0.9863	0.9878	0.9941	0.9888	0.9386
	CD	0.7395	-	-	0.8000	0.8600	0.8800	1.0000	0.9788
	FD	0.0140	0.0000	0.0000	0.0004	0.0000	0.0000	0.0000	0.0013
$n = 200$	η_1	20.698	1.1190	8.8000	22.520	62.880	69.950	523.90	78.480
	η_2	29.242	1.0010	5.4550	35.890	98.210	168.35	658.50	191.58
	α_1	0.8554	0.9938	0.9503	0.9890	0.9891	0.9769	0.9911	0.9484
	α_2	0.8388	0.9973	0.9437	0.9941	0.9959	0.9935	0.9949	0.9439
	CD	0.8029	-	-	0.8200	0.8710	0.9000	1.0000	0.9881
	FD	0.0170	0.0000	0.0000	0.0009	0.0003	0.0003	0.0003	0.0101

three missing values that are removed, and so the new sample size for this data is $n = 101$. Now, looking at Figure 4.2, we can see that there is a good separation in tail length (tail), foot length (footlgth) and ear conch length (earconch) among the sites (pop). For illustration, we use three continuous variables, which are tail, footlgtha and earconch and one categorical variable, sex, in our analysis. Then, the MMCM and the MMGM models were fitted to see which model could distinguish between Southern Victoria's possums and other regions' possums. Both models were fitted with $G = 1, \ldots, 5$ components , and $q = 1, 2, 3$ latent factors. Using the BIC and ARI to evaluate the performance of each model, MMCM performs better than MMGM. The best model from MMCM model has BIC $= -1131.3901$ and ARI $= 1$, whereas MMGM has BIC $= -1133.5920$ and ARI$= 0.9604$.

A perturbed version of the data is created to further investigate the effect of the presence of atypical points on the model performance. We add atypical points to that

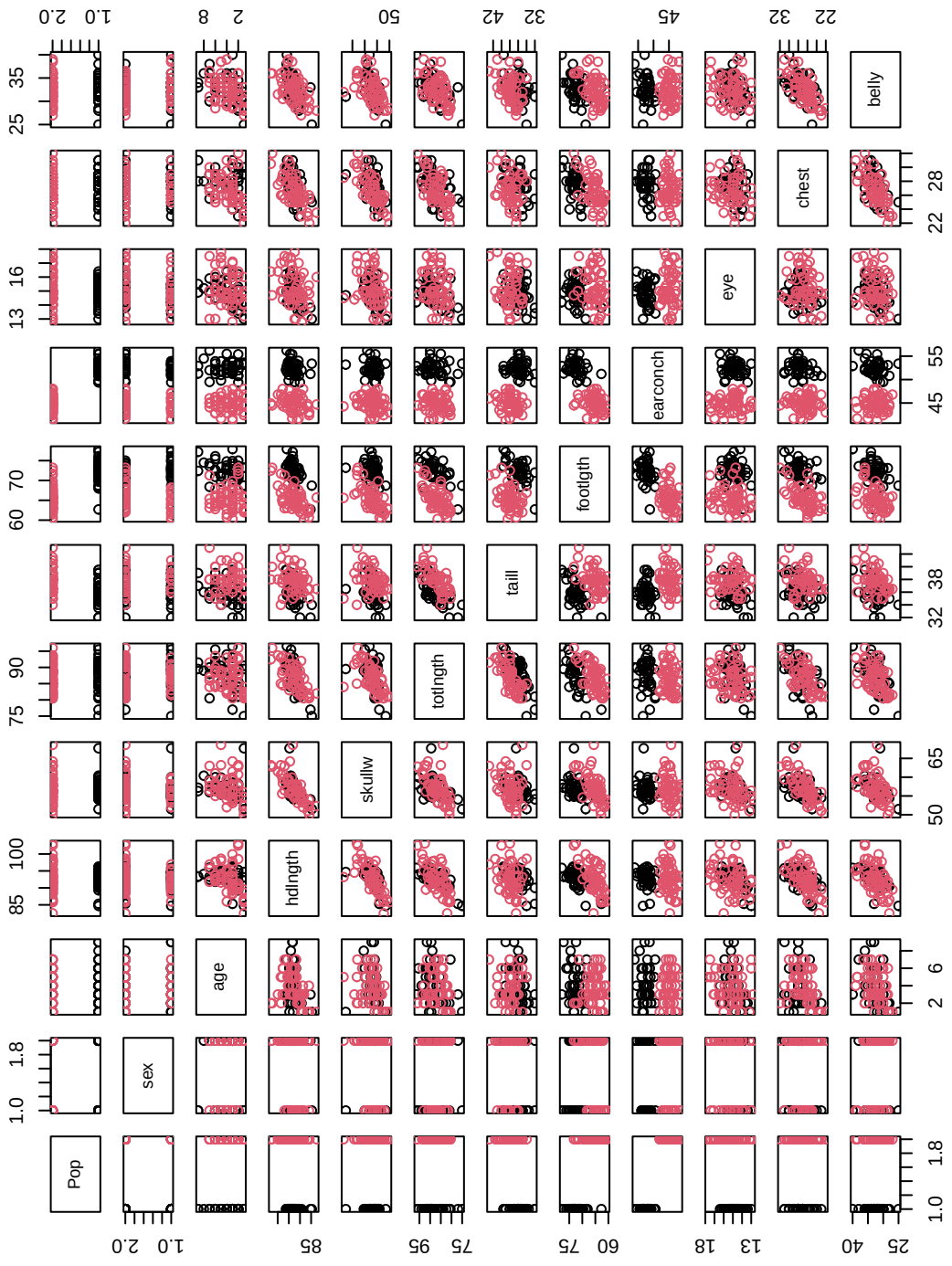


Figure 4.2: Pairs plot of the Possums data.

Table 4.14: Clustering results, BIC, and ARI for the chosen MMCM and MMGM models for the Possums data.

	MMCM		**MMGM**	
	1	2	1	2
Southern Victoria	43	0	43	0
Others	0	58	1	57

data by copying the last three observations from the others component and adding a constant $c = 10$ to them. Then we fit the MMCM and the MMGS models with different G and q and compare their performance. We also check if the MMCM correctly labels those new observations as atypical points or not. The results in Table 4.15 highlight that the MMCM model has lager BIC and higher ARI than the MMGM model. We also notice that our model identifies all three atypical points correctly (see Table 4.16).

Table 4.15: Clustering results for the chosen MMCM and MMGM models for the perturbed Possums data.

	MMCM		**MMGM**	
	1	2	1	2
Southern Victoria	43	0	43	0
Others	1	60	3	58
BIC	-1196.3260		-1207.3961	
ARI	0.9615		0.8868	

Table 4.16: Clustering results for the chosen MMCM models for the perturbed Possums data.

	1	2	Atypical
Southern Victoria	43	0	0
Others	0	58	3

4.7 Summary

A mixture of contaminated Gaussian mixed-type distributions has been introduced in this chapter. This model is based on a mixture of Gaussian mixed-type data. Herein, we extend that mixture by using the contaminated Gaussian distribution along with factor analysis and latent trait models. We discussed in detail the derivation of the parameters estimates, which was carried via the EM algorithm. Five simulation superiors are considered to test our model's performance in different situations. We fit our model and the competing model MMGM and use the ARI and BIC to measure both models' performance. In each case, we find some interesting properties of the model's parameters, namely η and α and how their values could be an indication of the absence or the presence of the atypical points and how far or close those points are. Finally, we fit the MMCM and MMGM to the original possum data and a modified version of possum data. In both versions of the data, our model MMCM fit better and correctly identified all the atypical points.

Chapter 5

Model Averaging for Skewed Data

5.1 Introduction

In this chapter, model averaging methods for skewed data are proposed. Specifically, we present a model averaging approach for a family of parsimonious variance-gamma distributions. Herein, we follow and extend Wei and McNicholas (2015) approach for mixture model averaging. This chapter is organized as follows. In Section 5.2, a brief literature review of existing methods is presented. In Section 5.3, we discuss in detail our methodology. Simulation studies and real data applications to demonstrate our methods in Sections 5.4 and 5.5, respectively. Finally, a summary section where we discuss all the presented work.

5.2 Background

5.2.1 Gaussian Parsimonious Mixture Models

In Section 2.2.2, we discuss one of the well-known mixture model the GMM. This model has in total

$$(G-1)+Gp+\frac{1}{2}Gp(p+1)$$

free parameters, where $G-1$ are the number of free parameters from mixing proportions $\pi_1,...,\pi_G$, Gp from the location parameter, and $Gp(p+1)/2$ from the component covariances $\boldsymbol{\Sigma}_1,...,\boldsymbol{\Sigma}_G$. The component covariances account for a large number of free parameters compared to the other parameters. Thus, Banfield and Raftery (1993) introduced parsimony in the GMM by imposing constraints on the decomposed component covariance matrices. The covariance decomposition uses the eigenvalue decomposition of the form

$$\boldsymbol{\Sigma}_g = \lambda_g \boldsymbol{\Gamma}_g \mathbf{A}_g \boldsymbol{\Gamma}_g^{'},$$

where λ_g is a constant, $\boldsymbol{\Gamma}_g$ is an orthonormal matrix of eigenvectors, and $\mathbf{A}_g$ is a diagonal matrix of eigenvalues with determinant 1.The parameters in above decomposition has a geometric interpretation as follows: λ_g control the volume, $\boldsymbol{\Gamma}_g$ the shape, and $\mathbf{A}_g$ the orientation of the cluster.

Celeux and Govaert (1995) extended the work of Banfield and Raftery (1993) and introduced 14 Gaussian parsimonious clustering models (GPCM) by imposing different combinations of the constraints $\lambda_g = \lambda$, $\boldsymbol{\Gamma}_g = \boldsymbol{\Gamma}$, $\boldsymbol{\Gamma}_g = \boldsymbol{I}_p$, $\mathbf{A}_g = \mathbf{A}$, and $\mathbf{A}_g = \mathbf{I}_p$ (Table 5.1). Those models are categorized into three categories: spherical (EII, VII), diagonal (EEI, EVI, VEI, VVI), and the rest are general. It can be

noticed that from Table 5.1 that the number of free parameters of some component covariance matrices is reduced; however, there are eight models have a number of free parameters that is quadratic in p. An alternative approach is proposed by McNicholas and Murphy (2008, 2010) based extending the mixture of factor analyzers model. For more details, see McNicholas (2016).

Different R packages are available that implement GPCM family and for other parsimonious mixture model families, one of which is mixture (Pocuca *et al.*, 2021). Similar to the regular GMM models, parameter estimation for the GPCM is carried through the EM algorithm (see Section 2.3.1), or a variant of EM, and the BIC is commonly used to select the best model, as discussed in 2.4.1

Table 5.1: Nomenclature, covariance decomposition for G components, and the number of free parameters in the covariance for each member of the GPCM family for p dimensional data.

Model	Volume $\lambda_g = \lambda$	Shape $\boldsymbol{A}_g = \boldsymbol{A}$	Orientation $\boldsymbol{\Gamma}_g = \boldsymbol{\Gamma}$	Covariance Decomposition	Number of Covariance Parameters
EII	Equal	Identity	Identity	$\lambda \boldsymbol{I}$	1
VII	Variable	Identity	Identity	$\lambda_g \boldsymbol{I}$	G
EEI	Equal	Equal	Identity	$\lambda \boldsymbol{A}$	p
VEI	Variable	Equal	Identity	$\lambda_g \boldsymbol{A}$	$G + (p-1)$
EVI	Equal	Variable	Identity	$\lambda \boldsymbol{A}_g$	$Gp - (G+1)$
VVI	Variable	Variable	Identity	$\lambda_g \boldsymbol{A}_g$	Gp
EEE	Equal	Equal	Equal	$\lambda \boldsymbol{\Gamma} \boldsymbol{A} \boldsymbol{\Gamma}'$	$p(p+1)/2$
VEE	Variable	Equal	Equal	$\lambda_g \boldsymbol{\Gamma} \boldsymbol{A} \boldsymbol{\Gamma}'$	$p(p+1)/2 + (G-1)$
EVE	Equal	Variable	Equal	$\lambda \boldsymbol{\Gamma} \boldsymbol{A}_g \boldsymbol{\Gamma}'$	$p(p+1)/2 - (G-1)(p-1)$
VVE	Variable	Variable	Equal	$\lambda_g \boldsymbol{\Gamma} \boldsymbol{A}_g \boldsymbol{\Gamma}'$	$p(p+1)/2 + (G-1)p$
EEV	Equal	Equal	Variable	$\lambda \boldsymbol{\Gamma}_g \boldsymbol{A} \boldsymbol{\Gamma}_g'$	$Gp(p+1)/2 - (G-1)p$
VEV	Variable	Equal	Variable	$\lambda_g \boldsymbol{\Gamma}_g \boldsymbol{A} \boldsymbol{\Gamma}_g'$	$Gp(p+1)/2 - (G-1)(p-1)$
EVV	Equal	Variable	Variable	$\lambda \boldsymbol{\Gamma}_g \boldsymbol{A}_g \boldsymbol{\Gamma}_g'$	$Gp(p+1)/2 - (G-1)$
VVV	Variable	Variable	Variable	$\lambda_g \boldsymbol{\Gamma}_g \boldsymbol{A}_g \boldsymbol{\Gamma}_g'$	$Gp(p+1)/2$

5.2.2 Merging Components

In model-based clustering, specifically in the Gaussian mixture model, the number of clusters is generally taken as equivalent to the number of mixture components for a given dataset. This is correct if the density of the clusters is Gaussian and they are well separated. However, in real applications, clusters are often overlapped and can not be modelled by the Gaussian mixture, especially with skewed or heavily tail data. Fitting a Gaussian mixture to non-Gaussian data will often result in an overestimation issue, i.e., the number of clusters is larger than the number of true groups. Thus, several components can be *a posteriori* merged into one cluster. Different merging approaches have been proposed in model-based clustering. Baudry *et al.* (2010) and Hennig (2010) merge the components hierarchically. The merging criterion used by Baudry *et al.* (2010) is based on the entropy criteria, whereas Hennig (2010) used aggregation criteria such as unimodality or misclassification probabilities.

Wei and McNicholas (2015) developed a merging method that is based on the ARI. Herein, different merging combinations will be considered then the ARI will be calculated between a reference model and each merging combination. The best merging combination will be the one that has the largest ARI value. The authors use Occam's window (see Section 5.2.3) to find a set of models then perform the model averaging approaches. They consider the following two cases to identify the reference model:

- **Case I**: The reference model is the one that has the largest BIC among all models in Occam's window. Models within Occam's window with fewer number of components than the reference model will be discarded.

- **Case II**: The reference model is the one that has the least number of components among all models in Occam's window.

In both cases, merging is performed if any model in Occam's window has more components than the reference model. The two cases will be the same if the model with the largest BIC also has the lowest number of components.

5.2.3 Bayesian Model Averaging

After fitting a set of mixture models, it is common to use some criteria to choose the "best" model. One of the commonly used selection methods is the BIC. When the best model is chosen, all other models will be ignored without taking into account the additional uncertainty that would have resulted from ignoring other models. This approach will result in different issues when the difference between the values of the criteria for two or more models is "small". Model averaging methods such as Bayesian model averaging (BMA; Hoeting *et al.*, 1999) take into account the uncertainty issue by taking the average of parameter estimates of different models.

Borrowing the notation from Hoeting *et al.* (1999), suppose we have a given data D, a set of models $\mathcal{M}_1, ..., \mathcal{M}_K$ and a quantity of interest Δ. Then the posterior distribution of Δ given D is

$$\text{pr}(\Delta \mid D) = \sum_{k=1}^{K} \text{pr}(\Delta \mid \mathcal{M}_k, D)\text{pr}(\mathcal{M}_k \mid D), \tag{5.1}$$

where $\text{pr}(\Delta \mid \mathcal{M}_k, D)$ represent the the posterior density of Δ under model $\mathcal{M}_k$, and

$\text{pr}(\mathcal{M}_k \mid D)$ is the posterior probability for model $\mathcal{M}_k$ that is given by

$$\text{pr}(\mathcal{M}_k \mid D) = \frac{\text{pr}(D \mid \mathcal{M}_k)\text{pr}(\mathcal{M}_k)}{\sum_{i=1}^{K} \text{pr}(D \mid \mathcal{M}_i)\text{pr}(\mathcal{M}_i)}, \tag{5.2}$$

where

$$\text{pr}(D \mid \mathcal{M}_k) = \int \text{pr}(D \mid \boldsymbol{\theta}_k, \mathcal{M}_k)\text{pr}(\boldsymbol{\theta}_k \mid \mathcal{M}_i)d\boldsymbol{\theta}_k. \tag{5.3}$$

where $\boldsymbol{\theta}_k$ parameters of $\mathcal{M}_k$, $\text{pr}(\boldsymbol{\theta}_k \mid \mathcal{M}_i)$ and $\text{pr}(\mathcal{M}_k)$ are the prior density and the prior probability for $\boldsymbol{\theta}_k$ and model $\mathcal{M}_k$, respectively.

It has been shown that BMA can provide better predictions than a single model; however, some computational difficulties are encountered with BMA (Madigan and Raftery, 1994). One issue is that the number of the models to be summed in (5.1) can be very large. Another issue is that the calculation of the posterior probability can be complicated due to the high-dimensional integrals in (5.3). To overcome the first issue, Madigan and Raftery (1994) use Occam's window to select a set of models such that any model that has not included Occam's window

$$\left\{ \mathcal{M}_k : \frac{\max_l\{\text{pr}(\mathcal{M}_l \mid D)\}}{\text{pr}(\mathcal{M}_k \mid D)} \leq c \right\} \tag{5.4}$$

will be discarded, where c is positive constant. Analogous to p-value = 0.05, Madigan and Raftery (1994) proposed using $c = 20$.

The second issue can be addressed by approximating the integral in (5.3). This can be done by using the BIC (Dasgupta and Raftery, 1998), i.e,

$$\text{pr}(D \mid \mathcal{M}_k) \approx \exp\left\{ \frac{1}{2}\text{BIC}_k \right\}, \tag{5.5}$$

where BIC_k is the BIC for model $\mathcal{M}_k$. Therefore, (5.2) can be written as

$$\text{pr}(\mathcal{M}_k \mid D) \approx \frac{\exp\left\{\frac{1}{2}\text{BIC}_k\right\}}{\sum_{i=1}^{K}\exp\left\{\frac{1}{2}\text{BIC}_i\right\}}, \tag{5.6}$$

and Occam's window is

$$\{\mathcal{M}_k : \max_l\{\text{BIC}_l\} - \text{BIC}_k \leq 2\log c\} \tag{5.7}$$

5.2.4 Variance-gamma Distribution

The variance-gamma (VG) distribution is a skewed continuous distribution, which is also known as the generalised Laplace distribution (GAL) or Bessel function (BF) distribution. The VG density can be derived as a special case of the generalized hyperbolic distribution by setting the index parameter $\lambda > 0$, and concentration parameters $\chi \to 1$ Kotz *et al.* (2001). A p-dimensional random random vector $\mathbf{X}$ following the VG distribution has a probability density function of the form:

$$\begin{aligned} f_{\text{VG}}(\mathbf{x}|\boldsymbol{\vartheta}) = &\left(\frac{\delta(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})}{\rho(\boldsymbol{\alpha},\boldsymbol{\Sigma})+2\gamma}\right)^{\frac{\gamma-p/2}{2}} \\ &\times \frac{2\gamma^{\gamma}K_{\left(\gamma-\frac{p}{2}\right)}\left(\sqrt{\left[\rho(\boldsymbol{\alpha},\boldsymbol{\Sigma})+2\gamma\right]\left[\delta(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\right]}\right)}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}\Gamma(\gamma)\exp\left\{(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha})\right\}}, \end{aligned}$$

where $\delta(\mathbf{X};\boldsymbol{\mu},\boldsymbol{\Sigma}) = (\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$, $\rho(\boldsymbol{\alpha},\boldsymbol{\Sigma}) = \boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}$,$\gamma > 0$ is the degrees of freedom, $\boldsymbol{\mu}$ is the location parameter , $\boldsymbol{\Sigma}$ is the scale matrix, $\boldsymbol{\alpha}$ is the skewness parameter, $K(\cdot)$ is the modified Bessel function of the third kind.

We can also obtain the density of VG by using the representation of the variance-mean mixtures (2.17) by letting $W \sim \text{Gamma}(\lambda, \psi/2)$ (McNicholas *et al.*, 2017).

5.2.5 A Mixture of Variance Gamma Distributions

McNicholas *et al.* (2017) introduced a mixture of variance-gamma factor analyzers model and the some properties of the variance-mean mixtures representation. Let $\mathbf{X} \sim \text{VG}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, \lambda, \psi)$ denote that the p-dimensional random vector $\mathbf{X}$ follows a VG distribution. Then the random variable $\mathbf{X}$ can be generated by combining a random variable $W_i \sim \text{G}(\lambda, \psi/2)$, where G denotes a gamma distribution, with the latent $\mathbf{V} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ via

$$\mathbf{X} = \boldsymbol{\mu} + W\boldsymbol{\alpha} + \sqrt{W}\mathbf{V}. \tag{5.8}$$

It is easy to show that $\mathbf{X}_i \mid w_i \sim \mathcal{N}(\boldsymbol{\mu} + w_i\boldsymbol{\alpha}, w_i\boldsymbol{\Sigma})$. Now, let $\gamma = \lambda = \psi/2$. Then, by Bayes' theorem, it follows that $W_i \mid (\mathbf{X}_i = \mathbf{x}) \sim \text{GIG}(2\gamma + \boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}, \delta(\boldsymbol{x}, \boldsymbol{\mu}|\boldsymbol{\Sigma}), \gamma - p/2)$. Then the density of mixture of variance-gamma distributions is

$$f_{\text{MVG}}(\boldsymbol{x}|\boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_{VG}(\boldsymbol{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\alpha}_g, \gamma_g). \tag{5.9}$$

To derive the density of mixture of VG factor analyzers, let $\mathbf{V} = \boldsymbol{\Lambda}\mathbf{Y} + \boldsymbol{\epsilon}_i$. Then, $\mathbf{X}$ in 5.8 can be written

$$\mathbf{X}_i = \boldsymbol{\mu} + W_i\boldsymbol{\alpha} + \sqrt{W_i}(\boldsymbol{\Lambda}\mathbf{Y}_i + \boldsymbol{\epsilon}_i), \tag{5.10}$$

where $\boldsymbol{\Lambda}$ is a matrix of factor loadings, $\mathbf{Y}$ is a q-dimensional latent variables and $\boldsymbol{\epsilon} \mid w_i \sim \mathcal{N}(\mathbf{0}, w_i\boldsymbol{\Psi})$ and $\boldsymbol{\Psi} = \text{dig}(\psi_1, ..., \psi_p)$. Then, it follows that $\mathbf{X}_i \mid w_i \sim \mathcal{N}(\boldsymbol{\mu} + w_i\boldsymbol{\alpha}, w_i(\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}))$.

Now, applying model-based cluster analysis methodology, the density of the mixture of variance-gamma factor analyzers (MVGFA) is given by

$$f_{\text{MVGFA}}(\boldsymbol{x}|\boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_{VG}(\boldsymbol{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g^{'} + \boldsymbol{\Psi}_g, \boldsymbol{\alpha}_g, \gamma_g), \tag{5.11}$$

where $\boldsymbol{\vartheta}$ denotes all model parameters. Parameter estimation is carried out through the alternating expectation-conditional maximization (AECM) algorithm (Meng and Van Dyk, 1997). The AECM is a variant of the EM algorithm that allow different complete-data at each stage (McLachlan and Krishnan, 2007).

5.3 Methodology

5.3.1 Parsimonious VG Family

As disused in Section 5.2.3, we can use the eigenvalue decomposition to write the component scale matrix in the following form

$$\boldsymbol{\Sigma}_g = \lambda_g \boldsymbol{\Gamma}_g \mathbf{A}_g \boldsymbol{\Gamma}_g^{'} \tag{5.12}$$

where λ_g is a constant $\boldsymbol{\Gamma}_g$ is an orthonormal matrix of eigenvectors, and $\mathbf{A}_g$ is a diagonal matrix of eigenvalues with determinant 1. To drive the parsimonious VG (ParVG) family of models, we can adapt the GPCM approach (Celeux and Govaert, 1995) by imposing different combinations of the constraints such as $\lambda_g = \lambda$, $\boldsymbol{\Gamma}_g = \boldsymbol{\Gamma}$, $\boldsymbol{\Gamma}_g = \boldsymbol{I}_p$, $\mathbf{A}_g = \mathbf{A}$, and $\mathbf{A}_g = \mathbf{I}_p$ to the eigen-decomposition of the component scale matrices. Similar to the GPCM family, the ParVG family will also contain 14 models but it is based on scale matrix structure. Accordingly, the destiny of parsimonious

VG mixtures can be written as

$$f_{\text{ParVG}}(\boldsymbol{x}|\boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_{VG}(\boldsymbol{x} \mid \boldsymbol{\mu}_g, \lambda_g \boldsymbol{\Gamma}_g \mathbf{A}_g \boldsymbol{\Gamma}_g^{'}, \boldsymbol{\alpha}_g, \gamma_g), \tag{5.13}$$

Note that the nomenclature, the scale matrix structure and the number of free parameters for each model from the ParVG family are similar to the GPCM models presented in Table 5.1.

5.3.2 Merging Mixture Components

As mentioned before, merging competent helps to tackle the over-estimation issue that results from fitting non-elliptical mixture models into elliptical data. Herein, we will follow Wei and McNicholas (2015) merging approach, where the ARI is used to choose the best merging.

Suppose that we fit a VG mixture model with G-components, and we want to merge them into H components, where $H < G$. Then, the density of the mixture after merging is equivalent to the original model and considered as another representation of it, i.e.,

$$f(\mathbf{x}) = \sum_{i=1}^{S} \pi_j^{\star} f_i^{\star}(\mathbf{x}) = \sum_{g=1}^{G} \pi_g f_{\text{VG}}(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\alpha}_g, \gamma_g), \tag{5.14}$$

where $\pi_i^{\star}$ is one of or the sum of two or more of the mixing proportions proportion, $f_i^{\star}(\cdot)$ is one of or a mixture of different component densities

$$f_{\text{VG}}(\mathbf{x} \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\alpha}_1, \gamma_1), \ldots, f_{\text{VG}}(\mathbf{x} \mid \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_G, \boldsymbol{\alpha}_G, \gamma_G).$$

To illustrate the above idea of merging, suppose $G = 4$ and $H = 3$; then $f_1^{\star}(\mathbf{x})$

could be a linear combination of the first two VG components, $f_2^\star(\mathbf{x})$ and $f_3^\star(\mathbf{x})$ are the third and the forth VG components or $f_1^\star(\mathbf{x})$ is a linear combination of the second and the fourth VG components and $f_2^\star(\mathbf{x})$ and $f_3^\star(\mathbf{x})$ are the first and the third components or any other combinations. Finding all the combinations is useful for calculating the possibility of merging but it never provides a way to choose the best combination. Following the approach that was proposed by Wei and McNicholas (2015), we can find the best merging.

Suppose that the reference model has H components and a model that has been chosen using Occam's window has G components. Here, we want to merge G components into H components, where $H < G$. As in Wei and McNicholas (2015), the merging procedure will be explained below through an example.

1. Generate a combination matrix $\mathbf{A}$ that has dimension $\binom{G}{H} \times G$, where each row is a subset of G.

   For example, let $G = \{1, 2, ..., 5\}$ denote the components in a model from Occam's window, and let $H = \{a, b, c\}$ denote the components in the reference model. Herein, we want to merge 5 components into 3 components. The dimension of $\mathbf{A}$ is $\binom{5}{3} \times 5 = 10 \times 3$. Suppose the fifth row from $\mathbf{A}$ is $\mathbf{a}_5 = (1, 4, 5)$, then it means that competent 1 assigned to new component a, competent 4 assigned to new component b and competent 5 assigned to new component c. The reaming components $\{2, 3\}$ will be merged with the same or different component from $\{1, 4, 5\}$.

2. Generate a permutation matrix $\mathbf{B}$, where $\mathbf{B}$ is $H^{(G-H)} \times (G - H)$ and each row is a subset from H. Herein, we calculate all merging possibilities to assign the reaming components from $\mathbf{A}$.

Continuing with our example, $\mathbf{B}$ is of size 9×2. The reaming components from step 1 example is $\{2, 3\}$ can be merged to according to one of the $\mathbf{B}$ row. Suppose $\mathbf{b}_3 = (a, c)$, then $\{2, 3\}$ will be merged with components 1 and 5 respectively such that the Occam's window model's components after merging become $\{a, b, c\} = \{1 \cup 2, 4, 5 \cup 3\}$.

3. Calculate the ARI's matrix $\mathbf{C}$ of size $\binom{G}{H} \times H^{(G-H)}$. Each entry in C is an ARI value calculated between the label from reference model and label from the model after merging.

   For example, from the first two steps we get merging that is based on the fifth row in matrix $\mathbf{A}$ and the third row in matrix $\mathbf{B}$. If we calculate the ARI value for this merging, then it will be sorted in the fifth row and the third column of $\mathbf{C}$.

4. Choose the best merging combination based on the largest ARI value in $\mathbf{C}$.

5.3.3 Averaging *A Posteriori* Probabilities

In model-based clustering paradigm, the component membership of observation x_i is denoted by z_{ig} where $z_{ig} = 1$ if observation x_i belongs to component g and $z_{ig} = 0$ otherwise. It is common to use *a posteriori* probabilities to get the predicted classifications, which can be obtained after estimation of the model parameters from the following equation:

$$\hat{z}_{ig} := \frac{\hat{\pi}_g f_{\mathrm{VG}}(\mathbf{x}_i \mid \hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Sigma}}_g, \hat{\boldsymbol{\alpha}}_g, \hat{\gamma}_g)}{\sum_{h=1}^{G} \hat{\pi}_h f_{\mathrm{VG}}(\mathbf{x}_i \mid \hat{\boldsymbol{\mu}}_h, \hat{\boldsymbol{\Sigma}}_h, \hat{\boldsymbol{\alpha}}_h, \hat{\gamma}_h)}. \tag{5.15}$$

Usually, we harden $\hat{z}_{ig}$ to obtain the predicted classifications via maximum *a posteriori* (MAP) probabilities, where $\text{MAP}(\hat{z}_{ig}) = 1$ if $\max_h\{\hat{z}_{ig}\}$ is occurring in group $h = g$ and $\text{MAP}(\hat{z}_{ig}) = 0$ otherwise.

In the averaging approach framework used by Wei and McNicholas (2015), if we merge components in any model in Occam's window, the *a posteriori* probabilities after merging are denoted by $\hat{z}^{\star}_{ij}$, for $j = 1, \ldots, H$, will be the sum of some of $\hat{z}_{ig}$. For example, if we merge the first three components, then $\hat{z}^{\star}_{i1} = \hat{z}_{i1} + \hat{z}_{i2} + \hat{z}_{i3}$.

Now to perform the average a posteriori probabilities (AAP); first, we need to identify which model to consider from Occam's window based on the reference model from Case I or Case II and do any necessary merging. Then calculate the weighted average of the *a posteriori* probabilities for each observation i and harden them to get the predicted classifications, where the weight for each model from $\mathcal{M}_1, \ldots, \mathcal{M}_K$ can be calculated from (5.6).

5.3.4 Model Averaging

This approach is a direct averaging method where we average parameters of models chosen within Occam's window. Unlike the AAP method, the merging criterion in the model averaging (MA) method cannot be applied. Thus, we cannot use Case I or Case II discussed in 5.2.2 to identify the reference model and which models to consider. In the MA method, the reference model is the model that has the largest BIC value and we only consider models that have the same number of components as the reference model to perform MA.

Suppose there are k models in Occam's window that have same number of components in the reference model. Then, to preform the MA method, we will calculate

the a weighted average of the parameter estimates for each parameter as follows.

$$\begin{aligned}
\bar{\pi}_g &= \sum_{k=1}^{K} \text{pr}(\mathcal{M}_k \mid D)\hat{\pi}_{kg}, \quad \hat{\pi}_{kg} = \frac{\sum_{i=1}^{n} \hat{z}_{ig}}{\sum_{g=1}^{G}\sum_{i=1}^{n} \hat{z}_{ig}}, \\
\bar{\boldsymbol{\mu}}_g &= \sum_{k=1}^{K} \text{pr}(\mathcal{M}_k \mid D)\hat{\boldsymbol{\mu}}_{kg}, \\
\bar{\boldsymbol{\Sigma}}_g &= \sum_{k=1}^{K} \text{pr}(\mathcal{M}_k \mid D)\hat{\boldsymbol{\Sigma}}_{kg}, \\
\bar{\boldsymbol{\alpha}}_g &= \sum_{k=1}^{K} \text{pr}(\mathcal{M}_k \mid D)\hat{\boldsymbol{\alpha}}_{kg}, \\
\bar{\gamma}_g &= \sum_{k=1}^{K} \text{pr}(\mathcal{M}_k \mid D)\hat{\gamma}_{kg},
\end{aligned}$$

where $\text{pr}(\mathcal{M}_k \mid D)$ is the model weights that can be calculated from (5.6). Then, we need to compute the averaged *a posteriori* probabilities $\bar{z}_{ig}$ that is based on the averaged parameter estimates such that

$$\bar{z}_{ig} = \frac{\bar{\pi}_g f_{\text{VG}}(\mathbf{x}_i \mid \bar{\boldsymbol{\mu}}_g, \bar{\boldsymbol{\Sigma}}_g, \bar{\boldsymbol{\alpha}}_g, \bar{\gamma}_g)}{\sum_{h=1}^{G} \bar{\pi}_h f_{\text{VG}}(\mathbf{x}_i \mid \bar{\boldsymbol{\mu}}_h, \bar{\boldsymbol{\Sigma}}_h, \bar{\boldsymbol{\alpha}}_h, \bar{\gamma}_h)}. \tag{5.16}$$

Once the $\bar{z}_{ig}$'s are calculated, we harden them via maximum *a posteriori* probabilities to get the predicted classifications.

5.3.5 Matching Components

One common issue in mixture models analysis is label switching, and there have been different solutions proposed to solve this issue (Stephens, 2000). For both averaging approaches that we introduce here, it is important to correctly match components

among models according to the reference model. Herein, we use the ARI and misclassification rate as criteria to match components. Components are matched based on the highest ARI and lowest misclassification rate.

Suppose we choose a certain number of models from Occam's window and do the necessary merging. We aim to match the components across those models before using any averaging methods. At this stage, all the models have the same number of components as the reference model. For simplicity, let the reference model have four components, and the corresponding label of the components are a, b, c, d. Suppose there is a model that also has four components with the following labels b, a, d, c. We notice that component $1, 2, 3, 4$ has a different label than the reference model. The following steps are used to match and relabel the components in a given model.

1. Calculate all the possibilities for re-labelling the components by generating a permutation matrix $\mathbf{A}$ of size 24×4. Each row in this matrix represents one possible match. For, example $\mathbf{a}_3 = (c, a, b, d)$, then component $1, 2, 3$ and 4 will be relabelled as c, a, b, d, respectively.

2. Calculate the ARI and the misclassification rate between the label from the reference model and the new label that we get from each row in matrix $\mathbf{A}$ and recorded in the 24×2 matrix $\mathbf{B}$.

3. Choose the best match that is based on the one that has the largest ARI value and the lowest misclassification rate.

5.4 Simulation

Different simulation scenarios are considered to assess the performance of the averaging methods. In all simulations, we use the gpcm() and vgpcm() functions from the R package mixture to fit the GPCM and ParVG models over $G = 1, \ldots, 10$. The ARI is used to compare the classification performance between GPCM and ParVG models before and after applying the averaging approaches.

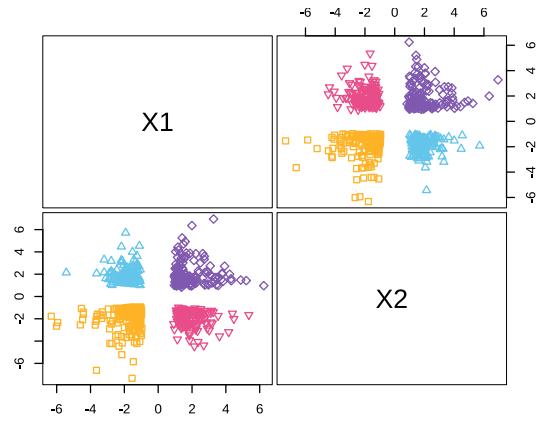
- Scenario I: 25 datasets were simulated from VG with 4 components that are well separated. All components have equal sample size $n = 150$ and number of variables $p = 2$. Figure 4.1-(a) is a pairs plot for one of the simulated datasets from this scenario.

  We can notice that, from Table 5.2, the classification performances of ParVG models after applying both averaging approaches are increased more than GPCM models. In the ParVG models, the average ARI of MA is slightly higher than both APP cases, whereas, in the GPCM, APP Case II has the highest ARI value.

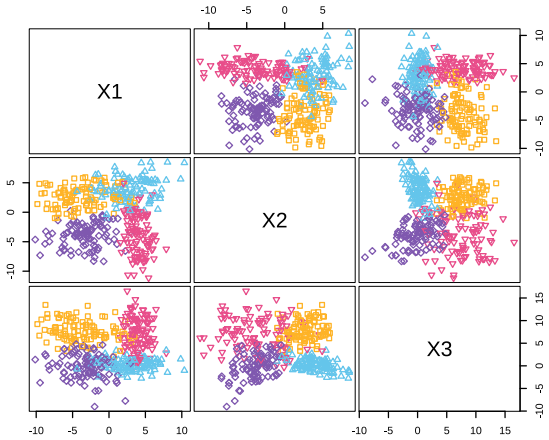
Table 5.2: The means and the standard deviations of ARI values for the best model, averaging *a posteriori* probabilities (AAP), and model averaging (MA) from simulation scenario 1.

ARI	ParVG				GPCM			
	Best	APP		MA	Best	APP		MA
		Case I	Case II			Case I	Case II	
Mean	0.9739	0.9833	0.9939	0.9948	0.6805	0.7270	0.7293	0.7177
Std. Deviation	0.0438	0.0295	0.0143	0.0090	0.0528	0.0581	0.0701	0.0885

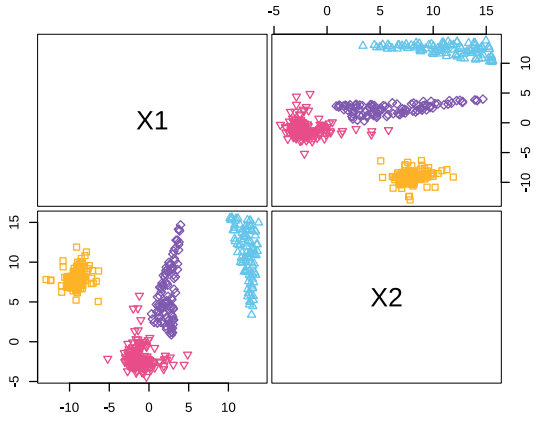
- Scenario II: We generate 25 data sets using genRandomClust() function from the



(a)



(b)



(c)

Figure 5.1: Example of one of the simulated data sets from (a) scenario 1, (b) scenario 2, (c) scenario 3.

R package clusterGeneration (Qiu and Joe, 2020). We generate 4 components that are overlapped (see Figure 5.1-(b)) with the following setting: the number of variables numNonNoisy=3, the separation between clusters sepVal=0.03 , and sample size $n = 400$.

As expected, both GPCM and ParVG have similar average ARI for the best model due to the high degree of overlap between the components. After applying the averaging approaches on both families of models, the APP method Case II leads to improvement in classification performance. However, averaging the ParVG models shows more improvement from 0.8469 to 0.9732, compared to the GPCM from 0.8862 to 0.8883.

Table 5.3: The means and the standard deviations of ARI values for the best model, averaging *a posteriori* probabilities (AAP), and model averaging (MA) from simulation scenario 2.

ARI	ParVG				GPCM			
	Best	APP		MA	Best	APP		MA
		Case I	Case II			Case I	Case II	
Mean	0.8469	0.8708	0.9723	0.8699	0.8862	0.8634	0.8883	0.8442
Std. Deviation	0.0372	0.0325	0.0301	0.0421	0.0312	0.0716	0.0311	0.0798

- Scenario III: We simulate 25 data sets with four components with a total number of observations 500 and the number of variables $p = 2$. One component with 150 observations from Gaussian distribution, one from VG with $n = 150$ and two triangles shape components from a uniform distribution, with each having 100 observations.

  From Table 5.4, as expected, the ParVG models not only fit better but also show more improvement after averaging compared to the GPCMs. In the ParVG

models, the AAP Case I method increases the classification performance from 0.9136 to 0.9708. However, in GPCM, the MA method helped to improve the average ARI from 0.7667 to 0.8651.

Table 5.4: The means and the standard deviations of ARI values for the best model, averaging *a posteriori* probabilities (AAP), and model averaging (MA) from simulation scenario 3.

ARI	ParVG				GPCM			
	Best	APP		MA	Best	APP		MA
		Case I	Case II			Case I	Case II	
Mean	0.9136	0.9708	0.9291	0.9202	0.7667	0.7839	0.8202	0.8651
Std. Deviation	0.0453	0.01494	0.0423	0.0206	0.0630	0.0348	0.0579	0.0624

5.5 Real Data Examples

5.5.1 Coffee data

The coffee data can be found in pgmm R package. It has 43 observations, and 12 variables represent the chemical constituents of two types of coffee (Arabica and Robusta) from different countries. For this analysis, we considered the following variables: Free Acid, Fat, Caffeine, Trigonelline, Chlorogenic Acid, Neochlorogenic Acid, and Isochlorogenic Acid. The ParVg and GPCM models are fitted to this data data for $G = 1, ..., 10$ component.

For the ParVG, the VEI model with $G = 2$ is the best model that is chosen based on the BIC (-688.3675) and has ARI (0.3601), whereas for the GPCM the best mode is VEI with BIC $= -682.8923$ and ARI $= 0.3732$ (Table 5.5). After applying the averaging methods, we notice that both families have perfect classification with

ARI = 1.

Table 5.5: Models that are chosen by Occam's window, along with the number of components, the weight for each model, BIC, and ARI values for the best model, from AAP and MA for the coffee data set.

ParVG						
Occam's window			$\Pr(\mathcal{M}_i \mid D)$	ARI values		
Model	BIC	G		Best	APP	MA
VEI	-688.3675	2	0.7442	0.3601	1	1
VII	-690.5036	2	0.2558			
GPCM						
Occam's window			$\Pr(\mathcal{M}_i \mid D)$	ARI values		
Model	BIC	G		Best	APP	MA
VEI	-682.8923	3	0.7762	0.3732	1	-
VEI	-685.3796	2	0.2238			

5.5.2 Hormone data

Disclaimer: This data is used to demonstrate the averaging methodologies ONLY. We are not trying to make any assumption or conclusion about any group.

The hormone data is available in faraway R package. It contains the sexual orientation of 26 males and the concentration of two hormones (androgen and estrogen). Using the mixture package, we fit the ParVG and GPCM models for $G = 1, ..., 10$ and then apply the averaging approaches on a set of models within Occam's window.

In the ParVG family, the APP Case I significantly improved the classification performance from -0.0372 to 0.6977. However, in the GPCM, none of the averaging methods help to improve the ARI values and remain the same. If we ignore all models

with only one component, we get a slight increase in the ARI value to 0.0587 from the MA method and 0.0152 from APP Case I.

5.5.3 Cathedral data

The Cathedral data is freely available in the faraway package, and it comprises the height and width that are measured in feet of two cathedral nave styles (Romanesque and Gothic) for 25 churches in England. Similar to the previous examples, both families of models are fitted for $G = 1, ..., 10$.

The VEV and the EEE are the best models selected by the BIC for the ParVG and the GPCM, respectively. In ParVG, two models fall in Occam's window with the same number of components but, in GPCM, twelve models are chosen with different components (Table 5.7). Both averaging approaches result in some improvement in the classification performance compared to the best model. However, the MA method gives higher ARI values than the APP approach. The ARI in ParVG increased from 0.0726 to 0.3214, and the GPCM from -0.0037 to -0.0333.

5.5.4 AIS data

The AIS data that has been used in Section 3.5 is used here to illustrate the model averaging techniques. The following variables are used: haemoglobin concentration (hg), body mass index (bmi), and percentage body fat (pcBfat). The ParVG and GPCM are fitted to the AIS data for $G = 1, \ldots, 10$.

The ParVG VEI model has been chosen using the BIC as the best model (BIC = -2784.1750), and it has three components with ARI = 0.7755, and GPCM the best model is VVE with BIC = -2811.0190 and ARI lower than the ParVG (0.6963).

Table 5.6: Models that are chosen by Occam's window, along with the number of components, the weight for each model, BIC, and ARI values for the best model, from AAP and MA for the hormone data set.

ParVG								
Occam's window			$\Pr(\mathcal{M}_i \mid D)$		ARI values			
Model	BIC	G	Case I/ MA	Case II	Best	APP		MA
						Case I	Case II	
VEV	-154.6879	3	0.6566	0.5879	-0.0372	0.6977	0.3535	0.1133
VII	-155.1414	2		0.3074				
VII	-156.4340	3	0.3434	0.1047				
GPCM								
Occam's window			$\Pr(\mathcal{M}_i \mid D)$		ARI values			
Model	BIC	G	MA	CaseI/Case II	Best	APP		MA
						Case I	Case II	
EII	-163.3608	1	0.2237	0.2175	0.0000	0.0000	0.0000	0.0000
VII	-163.3608	1	0.2237	0.2175				
EEI	-165.6811	1	0.0701	0.0682				
EVI	-165.6811	1	0.0701	0.0682				
VEI	-165.6811	1	0.0701	0.0682				
VVI	-165.6811	1	0.0701	0.0682				
EEE	-167.1293	1	0.0340	0.0330				
VEE	-167.1293	1	0.0340	0.0330				
EEV	-167.1293	1	0.0340	0.0330				
EVV	-167.1293	1	0.0340	0.0330				
VEV	-167.1293	1	0.0340	0.0330				
VVV	-167.1293	1	0.0340	0.0330				
VVE	-167.1293	1	0.0340	0.0330				
EVE	-167.1293	1	0.0340	0.0330				
EEI	-168.7945	2		0.0144				
EII	-168.9209	2		0.0135				

Table 5.7: Models that are chosen by Occam's window, along with the number of components, the weight for each model, BIC, and ARI values for the best model, from AAP and MA for the cathedral data set.

ParVG							
Occam's window			$\Pr(\mathcal{M}_i \mid D)$	ARI values			
Model	BIC	G	MA/ Case I/ Case II	Best	APP		MA
					Case I	Case II	
VEV	-136.9548	2	0.6242	0.0726	0.2226		0.3214
EEV	-137.9694	2	0.3758				

GPCM							
Occam's window			$\Pr(\mathcal{M}_i \mid D)$		ARI values		
Model	BIC	G	Case I/ MA	Case II	Best	APP	MA
						Case I Case II	
EEE	-142.6224	2	0.7097	0.1156	-0.0037	-0.0586 0.0000	-0.0333
EEE	-142.822	1		0.1046			
VEE	-142.8220	1		0.1046			
EVE	-142.822	1		0.1046			
EEV	-142.8220	1		0.1046			
VVE	-142.8220	1		0.1046			
EVV	-142.8220	1		0.1046			
VEV	-142.8220	1		0.1046			
VVV	-142.8220	1		0.1046			
VEE	-145.3386	2	0.1825	0.0297			
VEV	-147.5532	2	0.0603	0.0098			
VVE	-148.0303	2	0.0475	0.0077			

Occam's window chose two models in ParVG with a different number of components; one model has three components, and the other one has two. Hence, only APP Case II is used, and we merge the best model's components into two components. This leads to an increase in the ARI value to 0.8471. In the GPCM, two models fall in Occam's window with the same number of components. The MA and APP method is used without merging since all models have the same number of components and it slightly improve the classification performance (Table 5.8).

Table 5.8: Models that are chosen by Occam's window, along with the number of components, the weight for each model, BIC, and ARI values for the best model, from AAP and MA for the AIS data set.

ParVG						
Occam's window			$\Pr(\mathcal{M}_i \mid D)$	ARI values		
Model	BIC	G		Best	APP	MA
VEI	-2784.1750	3	0.8281	0.7755	0.8471	-
EEI	-2787.3190	2	0.1719			
GPCM						
Occam's window			$\Pr(\mathcal{M}_i \mid D)$	ARI values		
Model	BIC	G		Best	APP	MA
VVE	-2811.0190	3	0.9012	0.6963	0.7033	0.6867
VVV	-2815.4400	3	0.0988			

5.5.5 Pottery data

The Pottry2 data, which is available from heplots package, contains a chemical composition of 48 Romano-British potteries from three different regions. We considers five variables in this analysis: amount of iron oxide (Fe), amount of magnesium oxide (Mg), amount of calcium oxide (Ca), amount of sodium oxide (Na), and amount

of potassium oxide (K). We fit the data using the ParVG and GPCM families for $G = 1, \ldots, 10$.

From the ParVG family, three models fall in Occam's window: the first is VEV with three components, and the second and the third is EVI with two and three components, respectively. When we performed the APP Case I and MA methods, no merging was required because all models had $G = 3$ components whereas, in the APP Case II, we merged components of the models that have more than two components. All the averaging methods result in a great improvement in the classification performance compared to the best model (Table 5.9). However, APP Case II has the highest ARI (0.9512). Since only two models from the GPCM family fall in Occam's window with a different number of components, only APP Case II is performed. From Table 5.9, we can see a slight increase in the ARI value from 0.8853 to 0.8881.

Table 5.9: Models that are chosen by Occam's window, along with the number of components, the weight for each model, BIC, and ARI values for the best model, from AAP and MA for the Pottery data set.

ParVG								
Occam's window			$\Pr(\mathcal{M}_i \mid D)$		ARI values			
Model	BIC	G	Case I/ MA	Case II	Best	APP		MA
						Case I	Case II	
VEI	-363.0794	3	0.9009	0.9511	0.6352	0.9512	0.7151	0.8280
EVI	-368.7517	2		0.0528				
EVI	-369.0153	3	0.0463	0.0489				
GPCM								
Occam's window			$\Pr(\mathcal{M}_i \mid D)$		ARI values			
Model	BIC	G	Case I/ MA	Case II	Best	APP		MA
						Case I	Case II	
VVI	-309.4076	4	-	0.5643	0.8853	-	0.8881	-
VVI	-309.9249	3	-	0.4357				

5.6 Summary

In this chapter, we introduced a family of parsimonious VG mixture models. The eigenvalue decomposition of the component scale matrix was used to develop the ParVG family. We used the ParVG family to introduce two model averaging approaches that are analogues of the work of Wei and McNicholas (2015). The first approach is based on averaging the model parameters, and the second on averaging the *a posteriori* probabilities. In both approaches, we follow Wei and McNicholas (2015) in choosing the number of models to be averaged based on Occam's window and use the ARI as a criterion to merge components. A method to tackle the label switching issue, which is a common issue in cluster analysis, has been introduced. Herein, we use the ARI and the misclassification rate to match components across all models with the reference model.

Simulated and real datasets are used here to illustrate our methods. Based on our simulation results and the real data sets, we noticed that in most cases, the AAP averaging method tends to perform better than the MA and the best models. In general, there is an improvement in the classification performance after using some or all averaging methods, but not all the improvements are significant.

Chapter 6

Conclusions

6.1 Discussion

In this thesis, we proposed different unsupervised machine learning methods to model skewed and mixed-type data. A mixture for skewed-t mixed-type data, a mixture for contaminated mixed-type data, and a model averaging method for skewed data.

In chapter 3, we discussed the derivation of the mixture model for skewed-t mixed-type data. Herein, we use latent variable models to jointly model the mixed-type data by assuming that the observed variables are independent given the latent variables. Furthermore, the continuous variables are assumed to jointly follow a skew-t distribution.

In chapter 4, a mixture of contaminated mixed-type data is introduced to handle atypical "bad" points in clustering mixed-type data. We fit our model to different simulation scenarios along with real data to demonstrate the performance of our model. The result indicates that, in the presence of atypical observations, the model performs better than the normal mixed-type model.

In chapter 5, we introduced a family of parsimonious VG mixtures. Two model averaging methods (APP and MA) for skewed data are developed based on the BMA framework, Occam's razor and BIC. In both approaches, we match the components based on the misclassification rate and ARI. When merging components is required, we use the ARI to choose the best merging.

6.2 Future Work

Due to the paucity of work that has been done on clustering mixed-type data, there are endless possibilities to expand such methods. Our methods can be extended to other distributions. For instance, one can use another skewed distribution for the continuous such as the variance-gamma or one could use the Poisson distribution for the discrete variable.

Modern data gets more complex due to the "big data" phenomenon, and traditional methods developed for data mining are hard to apply to high-dimensional longitudinal data. Thus, expanding the methods that we developed in this thesis to high-dimensional or longitudinal data. This can be done by using a mixture of matrix variate distributions model and expand the proposed methodologies in the literature, e.g., Bouveyron *et al.* (2007), Anderlucci and Viroli (2015), Gallaugher and McNicholas (2018).

Furthermore, not only methodological developments are needed but also computational developments such as developing packages for different language programs such as R or Julia. The complexity and the numerical calculation come up with a price that needs high-performance computing. This problem can be tackled through implementing the method in different languages such as Julia.

Appendix A

Parameters updates for MMSM model

If the manifest variable X_{ij} is categorical with levels $0, 1, 2, \ldots K_j - 1$, then the complete log-likelihood is

$$\mathcal{L}_2 = \sum_{i=1}^{n} \hat{z}_{ig} \log g_1(x_{ij} = k | w_{ig}, \mathbf{y}_{ig}, \boldsymbol{\theta}_{jg})$$

where

$$g_1(x_{ij} = k | \mathbf{y}_i, w_i, \boldsymbol{\theta}_j) = \frac{\prod_{k=0}^{K_j-1} [\exp\{\eta_{jk} + \boldsymbol{\tau}'_{jk}\mathbf{y}_i\}]^{x_{ij}(k)}}{1 + \sum_{k=0}^{K_j-1} \exp\{\eta_{jk} + \boldsymbol{\tau}'_{jk}\mathbf{y}_i\}},$$

where $x_{ij}(k) = 1$ if x_{ij} is in category k and 0 otherwise.

Now, differentiating the conditional expected value of $\mathcal{L}_2$ with respect to η

$$\begin{aligned}\frac{\partial \mathcal{L}_2}{\partial \eta_{jkg}} E\left[\sum_{i=1}^{n} \hat{z}_{ig} \log g_1(x_{ij}=k|w_{ig},\mathbf{y}_{ig},\boldsymbol{\theta}_{jg})\right] &= E\left[\sum_{i=1}^{n} \frac{\partial \mathcal{L}_2}{\partial \eta_{jkg}} \hat{z}_{ig} \log g_1(x_{ij}=k|w_{ig},\mathbf{y}_{ig},\boldsymbol{\theta}_{jg})\right] \\ &= E\left[\sum_{i=1}^{n} \hat{z}_{ig}\left[x_{ij}(k) - g_1(x_{ij}=k|\mathbf{y}_{ig},w_{ig},\boldsymbol{\theta}_{jg})\right]\right] \\ &= \sum_{i=1}^{n} \hat{z}_{ig} \int_y \left[x_{ij}(k) - g_1(x_{ij}=k|\mathbf{y}_{ig},w_{ig},\boldsymbol{\theta}_{jg})\right] \\ &\quad \times h(\mathbf{y}_{ig}|\mathbf{x}_i) d\mathbf{y}_{ig}\end{aligned}$$

Similarly, differentiating the conditional expected value of $\mathcal{L}_2$ with respect to $\boldsymbol{\tau}$

$$\begin{aligned}\frac{\partial \mathcal{L}_2}{\partial \boldsymbol{\tau}_{jkg}} E\left[\sum_{i=1}^{n} \hat{z}_{ig} \log g_1(x_{ij}=k|w_{ig},\mathbf{y}_{ig},\boldsymbol{\theta}_{jg})\right] &= E\left[\sum_{i=1}^{n} \frac{\partial \mathcal{L}_2}{\partial \boldsymbol{\tau}_{jkg}} \hat{z}_{ig} \log g_1(x_{ij}=k|w_{ig},\mathbf{y}_{ig},\boldsymbol{\theta}_{jg})\right] \\ &= \sum_{i=1}^{n} \hat{z}_{ig} \int_y \mathbf{y}_{ig}\left[x_{ij}(k) - g_1(x_{ij}=k|\mathbf{y}_{ig},w_{ig},\boldsymbol{\theta}_{jg})\right] \\ &\quad \times h(\mathbf{y}_{ig}|\mathbf{x}_i) d\mathbf{y}_{ig}\end{aligned}$$

Thus, the updated for the parameters $(\eta_g, \boldsymbol{\tau}_g)$ can be calculated by solving the following equation:

$$\sum_{i=1}^{n} \hat{z}_{ig} \int_y (1, \mathbf{y}_{ig})\left[x_{ij}(k) - g_1(x_{ij}=k|\mathbf{y}_{ig},w_{ig},\boldsymbol{\theta}_{jg})\right] h(\mathbf{y}_i|\mathbf{x}_i) d\mathbf{y}_{ig} = \mathbf{0}$$

If the manifest variable X_{ij} is continuous, then the conditional expected value of the

complete log-likelihood is

$$\begin{aligned}\mathcal{L}_3^* = &\frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G}\hat{z}_{ig}\log(2\pi) + \frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G}\hat{z}_{ig}\log\left(\frac{1}{w_{ig}}\right) - \frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G}\hat{z}_{ig}\log|\boldsymbol{\Psi}_g| \\ &- \frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G}\hat{z}_{ig}\Big[b_{ig}\text{tr}\left\{(x_{ij}-\boldsymbol{\mu}_g)(x_{ij}-\boldsymbol{\mu}_g)^{'}\boldsymbol{\Psi}_g^{-1}\right\} - 2\text{tr}\left\{(x_{ij}-\boldsymbol{\mu}_g)\boldsymbol{\alpha}_g^{'}\boldsymbol{\Psi}_g^{-1}\right\} \\ &+a_{ig}\text{tr}\left\{\boldsymbol{\alpha}_g\boldsymbol{\alpha}_g^{'}\boldsymbol{\Psi}_g^{-1}\right\} - 2\text{tr}\left\{(x_{ij}-\boldsymbol{\mu}_g)^{'}\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}_g\boldsymbol{e}_{2ig}\right\} + 2\text{tr}\left\{\boldsymbol{\alpha}_g^{'}\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}_g\boldsymbol{e}_{1ig}\right\} \\ &+2\text{tr}\left\{\boldsymbol{\Lambda}_g\boldsymbol{E}_{3ig}\boldsymbol{\Lambda}_g^{'}\boldsymbol{\Psi}_g^{-1}\right\}\Big].\end{aligned}$$

Differentiating $\mathcal{L}_3^*$ with respect to $\boldsymbol{\Lambda}_g$ then with respect to $\boldsymbol{\Psi}_g^{-1}$ give $S_1(\boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g)$ and $S_2(\boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g)$ respectively

$$S_1(\boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g) = \frac{\partial\mathcal{L}_3^*}{\partial\boldsymbol{\Lambda}_g} = -\frac{1}{2}\sum_{i=1}^{n}\left[-2\boldsymbol{\Psi}_g^{-1}(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_g)\boldsymbol{e}_{1ig}^{'} + \boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}_g(\boldsymbol{E}_{3ig}^{'}+\boldsymbol{E}_{3ig})\right]$$

$$\begin{aligned}S_2(\boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g) = \frac{\partial\mathcal{L}_3^*}{\partial\boldsymbol{\Psi}_g} = &\frac{1}{2}\sum_{i=1}^{n}\hat{z}_{ig}\boldsymbol{\Psi}_g - \frac{1}{2}\left\{\sum_{i=1}^{n}\hat{z}_{ig}\left[b_{ig}(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_g)(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_g)^{'} - 2\hat{\boldsymbol{\alpha}}_g(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_g)^{'}\right.\right. \\ &\left.\left.+a_{ig}\hat{\boldsymbol{\alpha}}_g\hat{\boldsymbol{\alpha}}_g^{'} - 2(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_g)\boldsymbol{e}_{2ig}^{'}\boldsymbol{\Lambda}_g^{'} + 2\hat{\boldsymbol{\alpha}}_g\boldsymbol{e}_{1ig}^{'}\boldsymbol{\Lambda}_g^{'} + \boldsymbol{\Lambda}_g\boldsymbol{E}_{3ig}\boldsymbol{\Lambda}_g^{'}\right]\right\}.\end{aligned}$$

Now, solving $S_1(\hat{\boldsymbol{\Lambda}}_g, \hat{\boldsymbol{\Psi}}_g) = \mathbf{0}$ and $S_2(\hat{\boldsymbol{\Lambda}}_g, \hat{\boldsymbol{\Psi}}_g) = \mathbf{0}$ gives

$$\hat{\boldsymbol{\Lambda}}_g = \left\{\sum_{i=1}^{n}\hat{z}_{ig}\left[(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_g)\boldsymbol{e}_{2ig}^{'} - \hat{\boldsymbol{\alpha}}_g\boldsymbol{e}_{1ig}^{'}\right]\right\}\left\{\sum_{i=1}^{n}\hat{z}_{ig}\boldsymbol{E}_{3ig}^{'}\right\}^{-1},$$

$$\begin{aligned}\hat{\boldsymbol{\Psi}}_g = &\frac{1}{\sum_{i=1}^{n}\hat{z}_{ig}}\text{dig}\left\{\sum_{i=1}^{n}\hat{z}_{ig}\left[b_{ig}(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_g)(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_g)^{'} - 2\hat{\boldsymbol{\alpha}}_g(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_g)^{'} + a_{ig}\hat{\boldsymbol{\alpha}}_g\hat{\boldsymbol{\alpha}}_g^{'}\right.\right. \\ &\left.\left.-2(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_g)\boldsymbol{e}_{2ig}^{'}\hat{\boldsymbol{\Lambda}}_g^{'} + 2\hat{\boldsymbol{\alpha}}_g\boldsymbol{e}_{1ig}^{'}\hat{\boldsymbol{\Lambda}}_g^{'} + \hat{\boldsymbol{\Lambda}}_g\boldsymbol{E}_{3ig}\hat{\boldsymbol{\Lambda}}_g^{'}\right]\right\}.\end{aligned}$$

Appendix B

Parameters updates for MMCM model

The updates for categorical variables are similar to the updates in MMSM model (see Appendix A.

The updates for continuous variables are derived by differentiating the conditional expected value of $\mathcal{L}_4$ with respect of parameter of interest. The conditional expected value of $\mathcal{L}_4$ is

$$\mathcal{Q}_1 = C - \frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G}\hat{z}_{ig}\log|\boldsymbol{\Psi}_g| - \frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G}\left[\hat{z}_{ig}\left(\hat{v}_{ig} + \frac{1-\hat{v}_{ig}}{\eta_g}\right)\text{tr}\left\{(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)^{'}\boldsymbol{\Psi}_g^{-1}\right.\right.$$
$$\left.\left.(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)\right\} - 2\text{tr}\left\{(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)^{'}\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}_g\boldsymbol{e}_{1ig}\right\} + \text{tr}\left\{\boldsymbol{\Lambda}_g\boldsymbol{E}_{2ig}\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}_g^{'}\right\}\right]$$

Differentiating $\mathcal{Q}_1$ with respect to $\boldsymbol{\Lambda}_g$ then with respect to $\boldsymbol{\Psi}_g^{-1}$ gives

$$S_1(\boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g) = \frac{\partial\mathcal{Q}_1}{\partial\boldsymbol{\Lambda}_g} = -\frac{1}{2}\sum_{i=1}^{n}[-2\boldsymbol{\Psi}_g^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)\boldsymbol{e}_{1ig}^{'} + \boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}_g(\boldsymbol{E}_{2ig}^{'} + \boldsymbol{E}_{2ig})]$$

$$S_2(\boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g) = \frac{\partial \mathcal{Q}_1}{\partial \boldsymbol{\Psi}_g} = \frac{1}{2}\sum_{i=1}^{n} \hat{z}_{ig}\boldsymbol{\Psi}_g - \frac{1}{2}\sum_{i=1}^{n}\left[\hat{z}_{ig}\left(\hat{v}_{ig} + \frac{1-\hat{v}_{ig}}{\eta_g}\right)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)^{'} \right.$$
$$\left. -2(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)\boldsymbol{e}_{1ig}^{'}\boldsymbol{\Lambda}_g^{'} + \boldsymbol{\Lambda}_g\boldsymbol{E}_{2ig}\boldsymbol{\Lambda}_g^{'}\right]$$

Solving $S_1(\hat{\boldsymbol{\Lambda}}_g, \hat{\boldsymbol{\Psi}}_g) = \mathbf{0}$ and $S_2(\hat{\boldsymbol{\Lambda}}_g, \hat{\boldsymbol{\Psi}}_g) = \mathbf{0}$ gives

$$\hat{\boldsymbol{\Lambda}}_g = \left\{\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)\boldsymbol{e}_{1ig}^{'}\right\}\left\{\sum_{i=1}^{n}\boldsymbol{E}_{2ig}\right\}^{-1}$$

$$\hat{\boldsymbol{\Psi}}_g = \frac{1}{\sum_{i=1}^{n}\hat{z}_{ig}}\text{dig}\left\{\sum_{i=1}^{n}\left[\hat{z}_{ig}\left(\hat{v}_{ig} + \frac{1-\hat{v}_{ig}}{\hat{\eta}_g}\right)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)^{'}\right.\right.$$
$$\left.\left. -2(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)\boldsymbol{e}_{1ig}^{'}\hat{\boldsymbol{\Lambda}}_g^{'} + \hat{\boldsymbol{\Lambda}}_g\boldsymbol{E}_{2ig}\hat{\boldsymbol{\Lambda}}_g^{'}\right]\right\}$$

Appendix C

Computational timing for simulation

The following tables show the time to run each simulation per dataset. Note that we run the simulation in parallel with the following machine specifications Cpu: 80, memory: 190000M in Linux x86_64.

Table C.1: Average run-times in second per dataset for simulation in section 3.4.

n	Simulation 1	Simulation 2
100	214946.22	268515.01
200	403612.79	535370.32
400	773007.43	835940.57

Table C.2: Average run-times in second per dataset for simulation in section 4.5.

Simulation	n	MMCM	MMGM
Scenario 1	50	213733.59	2545.20
	100	309559.38	3448.08
	200	560028.39	6896.16
Scenario 2	50	63249.36	2233.80
	100	71996.64	5525.49
	200	302947.64	14141.30
Scenario 3	50	68749.71	1220.40
	100	121952.41	1314.00
	200	265654.36	2302.20
Scenario 4-a	50	58887.64	17946.00
	100	219895.32	32646.60
	200	275771.52	55511.73
Scenario 4-b	50	4332.00	20847.57
	100	196260.00	31186.80
	200	216286.57	57231.59
Scenario 4-c	50	3319.20	10567.54
	100	119308.10	39112.20
	200	173731.30	105938.54
Scenario 4-d	50	3670.20	19571.38
	100	165066.70	23310.00
	200	239419.88	40754.20
Scenario 5	50	3720.60	774.00
	100	4818.60	6809.40
	200	8868.60	7666.20

Bibliography

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.

Amiri, L., Khazaei, M., and Ganjali, M. (2018). A mixture latent variable model for modeling mixed data in heterogeneous populations and its applications. *AStA Advances in Statistical Analysis*, **102**(1), 95–115.

Anderlucci, L. and Viroli, C. (2015). Covariance pattern mixture models for the analysis of multivariate heterogeneous longitudinal data. *The Annals of Applied Statistics*, **9**(2), 777–800.

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

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

Bartholomew, D. J. and Knott, M. (1999). *Latent variable models and Factor analysis*. Arnold.

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. Taylor & Francis.

Baudry, J.-P., Raftery, A. E., Celeux, G., Lo, K., and Gottardo, R. (2010). Combining mixture components for clustering. *Journal of Computational and Graphical Statistics*, **19**(2), 332–353.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., and Lindsay, B. G. (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**(2), 373–388.

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

Browne, R. P. and McNicholas, P. D. (2012). Model-based clustering, classification, and discriminant analysis of data with mixed type. *Journal of Statistical Planning and Inference*, **142**(11), 2976–2984.

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

Clogg, C. C. and Goodman, L. A. (1984). Latent structure analysis of a set of multidimensional contingency tables. *Journal of the American Statistical Association*, **79**(388), 762–771.

Collins, L. M. and Lanza, S. T. (2009). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*, volume 718. John Wiley & Sons.

Cuesta-Albertos, J. A., Gordaliza, A., and Matrán, C. (1997). Trimmed k-means: An attempt to robustify quantizers. *The Annals of Statistics*, **25**(2), 553–576.

Dasgupta, A. and Raftery, A. E. (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association*, **93**(441), 294–302.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, **39**(1), 1–22.

Fowlkes, E. B. and Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, **78**(383), 553–569.

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.

Gallaugher, M. P. and McNicholas, P. D. (2018). Finite mixtures of skewed matrix variate distributions. *Pattern Recognition*, **80**, 83–93.

Gallegos, M. T. and Ritter, G. (2005). A robust method for cluster analysis. *The Annals of Statistics*, **33**(1), 347–380.

Garcia-Escudero, L. A. and Gordaliza, A. (1999). Robustness properties of k means and trimmed k means. *Journal of the American Statistical Association*, **94**(447), 956–969.

Ghahramani, Z., Hinton, G. E., *et al.* (1996). The EM algorithm for mixtures of

factor analyzers. Technical report, Technical Report CRG-TR-96-1, University of Toronto.

Gollini, I. and Murphy, T. B. (2014). Mixture of latent trait analyzers for model-based clustering of categorical data. *Statistics and Computing*, **24**(4), 569–588.

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

Guidotti, E. (2020). calculus: High dimensional numerical and symbolic calculus in R. *arXiv preprint arXiv:2101.00086*.

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

Hasselman, B. and Hasselman, M. B. (2018). Package nleqslv.

Hennig, C. (2010). Methods for merging gaussian mixture components. *Advances in Data Analysis and Classification*, **4**(1), 3–34.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial (with comments by M. Clyde, David Draper and EI George, and a rejoinder by the authors. *Statistical Science*, **14**(4), 382–417.

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

Jason, L. and Glenwick, D. (2016). *Handbook of methodological approaches to community-based research: Qualitative, quantitative, and mixed methods*. Oxford University Press.

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

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

Kotz, S., Kozubowski, T., and Podgórski, K. (2001). *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*. Number 183. Springer Science & Business Media.

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.

Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin Company.

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

Lindsay, B. G. (1995). Mixture models: theory, geometry and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages i–163. JSTOR.

MacQueen, J. *et al.* (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model

uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, **89**(428), 1535–1546.

McCutcheon, A. L. (1987). *Latent class analysis*. Number 64. Sage.

McLachlan, G. J. and Krishnan, T. (2007). *The EM algorithm and extensions*. John Wiley & Sons.

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

McNeil, A. J., Frey, R., and Embrechts, P. (2015). *Quantitative risk management: concepts, techniques and tools-revised edition*. Princeton University Press.

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

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

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

McNicholas, S. M., McNicholas, P. D., and Browne, R. P. (2017). A mixture of variance-gamma factor analyzers. In *Big and Complex Data Analysis*, pages 369–385. Springer.

McParland, D. and Gormley, I. C. (2016). Model based clustering for mixed data: clustmd. *Advances in Data Analysis and Classification*, **10**(2), 155–169.

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.

Muthen, B. and Asparouhov, T. (2006). Item response mixture modeling: Application to tobacco dependence criteria. *Addictive Behaviors*, **31**(6), 1050–1066.

Narasimhan, B., Koller, M., Johnson, S. G., Hahn, T., Bouvier, A., Kiêu, K., Gaure, S., and Narasimhan, M. B. (2021). Package "cubature".

Pocuca, N., Browne, R. P., and McNicholas, P. D. (2021). *Mixture: mixture models for clustering and classification*. R package version 2.0.4.

Punzo, A. and McNicholas, P. D. (2014). Robust high-dimensional modeling with the contaminated Gaussian distribution. *arXiv preprint arXiv:1408.2128*.

Punzo, A. and McNicholas, P. D. (2016). Parsimonious mixtures of multivariate contaminated normal distributions. *Biometrical Journal*, **58**(6), 1506–1537.

Qiu, W. and Joe, H. (2020). *Random Cluster Generation (with Specified Degree of Separation)*. R package version 1.3.7.

Raftery, A. E. and Dean, N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association*, **101**(473), 168–178.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**(336), 846–850.

Ritter, G. (2014). *Robust cluster analysis and variable selection*. CRC Press.

Robertson, J. and Kaptein, M. (2016). *Modern statistical methods for HCI*. Springer.

Ruwet, C., Garcia-Escudero, L. A., Gordaliza, A., and Mayo-Iscar, A. (2013). On the breakdown behavior of the tclust clustering procedure. *Test*, **22**(3), 466–487.

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

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

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **62**(4), 795–809.

Tang, Y., Browne, R. P., and McNicholas, P. D. (2015). Model based clustering of high-dimensional binary data. *Computational Statistics & Data Analysis*, **87**, 84–101.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **61**(3), 611–622.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. *Contributions to Probability and Statistics*, pages 448–485.

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology*, **33**(1), 213–239.

Vermunt, J. K. (2007). Multilevel mixture item response theory models: an application in education testing. *Proceedings of the 56th session of the International Statistical Institute. Lisbon, Portugal*, **2228**.

Vermunt, J. K. and Magidson, J. (2002). *Latent Class Cluster Analysis*, pages 89–106. Cambridge University Press.

Watanabe, M. and Yamaguchi, K. (2003). *The EM Algorithm and Related Statistical Models*. CRC Press.

Wei, Y. and McNicholas, P. D. (2015). Mixture model averaging for clustering. *Advances in Data Analysis and Classification*, **9**(2), 197–217.

Wolfe, J. H. (1963). *Object cluster analysis of social areas*. Ph.D. thesis, University of California.

Yoshida, R., Higuchi, T., and Imoto, S. (2004). A mixed factors model for dimension reduction and extraction of a group structure in gene expression data. In *Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference*, pages 161–172.