Parsimonious Contaminated Shifted Asymmetric Laplace Mixtures: Unsupervised Learning with Outlier Identification for Asymmetric Clusters in High Dimensions

by

Paul McLaughlin

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Statistics

Department of Mathematical and Statistical Sciences
University of Alberta

Abstract

A family of parsimonious contaminated shifted asymmetric Laplace mixtures is developed for asymmetric clusters in the presence of outliers and noise (referred to as bad points herein). A series of constraints are applied to a modified factor analyzer structure of the scale matrix parameters, yielding the twelve models comprising the family. Application of the modified factor analyzer structure and this series of parsimonious constraints makes this model effective at analyzing high-dimensional data by reducing the quantity of free parameters that need to be estimated in the model. Notably, these models are developed for an unsupervised setting and do not rely on any previous information about identified outliers or the underlying group structure of the data. A variant of the EM algorithm is developed for parameter estimation. Various implementation issues are discussed, and a series of analyses and comparisons to well-established clustering methods is conducted on real and simulated data.

Acknowledgements

Firstly, I would like to thank my supervisors Dr. Brian Franczak and Dr. Adam Kashlak for their support and guidance throughout my studies. Your commitment to my education and my success has had an immeasurable impact on who I have become as a researcher and as a person.

I would also like to thank Dr. Bei Jiang and Dr. Cristina Tortora for taking their time to review my thesis.

I must thank my colleagues in the department Katie Burak, Liam Welsh and Garnet Liam Peet-Pare. Without our study sessions, coffee breaks and all of your kindness and humor, I don't know if I would have survived the trials and stresses of my first year. I felt at home studying in our little cubby office , in a way I was never sure that I would when I first began this degree, and it is all thanks to you three.

But I can't neglect my other friends who have also done so much to encourage me on my path. Thank you to Tamer Harb, Donovan Eckstrom, Samantha Fraughton, Andrew Smith, and Tyler and Holly Wilde for being there for me when I needed it most in these past few years. Whether it was giving me a chance to decompress or lending me a comforting voice during difficult times, I will always remember what you have all done for me.

Finally, my sincerest gratitude to my mother and father, who have supported me and cared for me through all of my antics over the years. The wisdom you have conveyed to me has not entirely fallen on deaf ears, and it has been invaluable as I have pursued this work and beyond.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Cluster analysis is the process of identifying and assigning a meaningful group structure to a data set. Observations are classified into sub-populations of the data known as components or groups with the intention of maximizing within group similarity and between group dissimilarity. More specifically, cluster analysis attempts this process in an unsupervised setting where no previous information of any component membership properties for the observations is available. Many methods and techniques for clustering exist, including non-parametric methods such as distance-based methods and agglomerative hierarchical clustering. However, this research is focused on the parametric method of model-based clustering, where the parameter estimates of a finite mixture model are fitted by maximum likelihood estimation.

Model-based clustering provides several advantages over non-parametric methods. In particular, model-based clustering provides a more rigorous definition of a component and established methods of comparing proposed models. As noted by Marriott (1974) this approach "is about the only clustering technique that is entirely satisfactory from the mathematical point of view. It assumes a well-defined mathematical model, investigates it by well-established statistical techniques, and provides a test of significance for the results." The strong mathematical framework and a clear method of model comparisons justifies our use of model-based clustering as the foundation for the models developed in this thesis.

1.1 Motivation

Despite these advantages, many of the model-based clustering algorithms possess their own set of limitations. If the data features outliers or noise, models will often assimilate them into the group structure. This results in unwanted influence on the parameter estimates, compromising the clustering solution. By introducing the distinction between "good" points and "bad' points, with the former referring to observations within the components and the latter referring to the aforementioned outliers and noise, contamination can be introduced to the distributions in the mixture model (Punzo and McNicholas 2016; Tukey 1960). This allows for clustering algorithms to identify outliers and omit these points when calculating the parameter estimates.

Additionally, it is typical that the finite mixture model assumes a Gaussian distribution for each component in the data. Consequentially, the model will assume that components in the data are symmetric. This can result in the model over-fitting the data, an issue in which multiple symmetric components are assigned to describe a single asymmetric component. While the introduction of contamination can help to alleviate the effect of outliers, in the presence of asymmetric components it is prone to incorrectly classifying observations within the fringes of the skew as bad points. Many distributions have been introduced that allow for the parametrization of skewness to accommodate asymmetry, such as skew-normal (Lin 2009) and skew-t (Lin 2010) distributions, but the shifted asymmetric Laplace distribution (Franczak et al. 2014) offers the unique of advantage of compatibility with the contamination protocol.

Finally, when dealing with high-dimensional data algorithms can become incredibly computationally expensive making them an impractical method of analyzing data. This is primarily the result of the difficulty of estimating a $p \times p$ covariance or scale matrix parameter for large values of p. Factor analysis is a well established method of dimension reduction, which assumes the variability in the data can be described by a small number of latent factors. Applying a factor analyzer decomposition to the covariance or scale matrix parameters in a mixture model (Ghahramani and Hinton 1997) eases the computational demands for parameter estimates. This can be further alleviated through the application of parsimonious constraints to the factor analyzer decomposition (McNicholas and Murphy 2008).

Within the existing literature of model-based clustering, clustering algorithms exist that utilize the solutions described to address each possible pair of these limitations. Mixtures of contaminated shifted asymmetric Laplace distribution (Morris et al. 2019) account for outliers and asymmetric components; mixtures of contaminated Gaussian factor analyzers (Punzo et al. 2020) account for outliers and high dimensions; parsimonious shifted asymmetric Laplace mixtures (Franczak et al. 2013) account for asymmetric components and high dimensions. None currently exist that simultaneously account for all three.

1.2 Thesis Objectives

In this thesis we propose a family of parsimonious contaminated shifted asymmetric Laplace mixtures to address all three of the limitations we have detailed in the motivation. An alternating expectation conditional-maximization algorithm is used to implement this family of models. Once these models are developed, they are applied to a set of real data on athletes collected by the Australian Institute of Sports (Telford and Cunningham 1991) as well a series of simulated data sets with the intent of showing improved performance in comparison to existing model-based clustering algorithms.

1.3 Thesis Outline

1.3.1 Chapter 2

The second chapter discusses the concepts and work in mixture modelling literature that this thesis builds upon. This review includes, but is not limited to: finite mixture models, mixtures of shifted asymmetric Laplace distributions, mixtures of contaminated distributions, families of parsimonious mixture models, and variants of the expectation maximization algorithm.

1.3.2 Chapter 3

The third chapter provides the methodology for the development of the algorithm used to implement our proposed family of models. We discuss how the factor analyzer decomposition and the family of parsimonious constraints are applied to a mixture of contaminated shifted asymmetric Laplace distributions. Next, a derivation of the expected values and parameter estimates used in the implementation of these models is given. The chapter concludes with the discussion of computational considerations that were required to ensure the algorithm used to implement our family of models remained practical.

1.3.3 Chapter 4

Chapter 4 provides an exploratory analysis conducted on simulated data sets using our family of models to observe classification performance, component recovery, and the reliability of model selection criteria for identifying an optimal model.

1.3.4 Chapter 5

Chapter 5 provides an analysis conducted on a real data set provided by the Australian Institute of Sport (AIS) using our family of models to compare the classification performance and component recovery to results obtained using competing mixture models in Punzo et al. (2020).

1.3.5 Chapter 6

Chapter 6 concludes this thesis with a discussion of our results and recommendations for future work.

Chapter 2

Model-based Clustering

2.1 Mixture Models

2.1.1 Finite mixture models

A finite mixture models assumes that the data is sampled from a population comprised of a finite number of sub-populations, such that each sub-population can be modelled by a probability distribution. It is typical that the distribution type is also assumed to be constant across all sub-populations, but not necessary.

More formally, a p-dimensional random vector $\mathbf{X}$ arises from a finite mixture model if for all $\mathbf{x} \subset \mathbf{X}$,

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

such that

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

where, for a group g, π_g is the mixing proportions, $\boldsymbol{\theta}_g$ is a vector of parameters and $f_g(\mathbf{x}|\boldsymbol{\theta}_g)$ is the component density. See McLachlan and Peel (2000a), McNicholas (2016), and Titterington et al. (1985) for a detailed review of finite mixture models.

2.1.2 Gaussian mixture models

At the time of the review paper by Fraley and Raftery (2002), the most frequently used component density in mixture modelling to date is the multivariate Gaussian.

This is primarily due to their mathematical tractability and prominence in the history of statistical research. The density of a Gaussian mixture model (GMM) is expressed as

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

where $\boldsymbol{\vartheta}$ is the vector of parameters, π_g is the probability of membership in component g, and the component densities given by

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

are the density of a multivariate Gaussian distribution with a mean vector $\boldsymbol{\mu}_g$ and a covariance matrix $\boldsymbol{\Sigma}_g$.

2.1.3 Contaminated Gaussian mixture models

The density for the contaminated Gaussian distribution (Tukey 1960) is given by

$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \rho\phi(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) + (1-\rho)\phi(\mathbf{x} \mid \boldsymbol{\mu}, \eta\boldsymbol{\Sigma}) \tag{2.4}$$

where $\rho \in (0,1)$, $\eta > 1$, $\boldsymbol{\vartheta} = \{\rho, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \eta\}$ and $\phi(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is as defined in Equation (2.3). These two embedded components are distinguished as a “good” component which features lower dispersion to describe the “true” density, and a “bad” component which accommodates the outliers or contamination. It follows then that ρ denotes the proportion of “good” observations in the component while η denotes the degree of the contamination. Since $\eta > 1$ it can be interpreted as an inflation parameter for the increased variability due to contamination (Punzo and McNicholas 2016).

Therefore, the density for each of the G groups in a mixture of contaminated Gaussian distributions is itself a mixture of two groups centered at the same point with proportional covariance matrices. The density for mixtures of contaminated Gaussian distributions is then given by

$$f(\mathbf{x}|\ \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g \left[\rho_g \phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) + (1-\rho_g)\phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \eta_g\boldsymbol{\Sigma}_g)\right] \tag{2.5}$$

where all terms are as previously defined (Punzo and McNicholas 2016).

2.2 Mixtures of Shifted Asymmetric Laplace Distributions

2.2.1 Generalized inverse Gaussian distribution

The density of a random variable W following a generalized inverse Gaussian (GIG) distribution, notated as $W \sim GIG(a, b, \upsilon)$, is given by

$$g(w) = \frac{(a/b)^{\upsilon/2} w^{\upsilon-1}}{2K_\upsilon(\sqrt{ab})} \exp\left\{-\frac{aw + b/w}{2}\right\} \tag{2.6}$$

for $w > 0$, where $a, b \in \mathbb{R}^+,\ \upsilon \in \mathbb{R}$ and K_υ is the modified Bessel function of the third kind with index υ (Barndorff-Nielsen et al. 1982). The GIG distribution possesses the property of tractability for the following expected values:

$$\mathbb{E}[W] = \sqrt{\frac{b}{a}} R_\upsilon(\sqrt{ab}) \tag{2.7}$$

$$\mathbb{E}[1/W] = \sqrt{\frac{a}{b}} R_\upsilon(\sqrt{ab}) - \frac{2\upsilon}{b} \tag{2.8}$$

$$\mathbb{E}[\log W] = \log \sqrt{\frac{a}{b}} + \frac{\partial}{\partial \upsilon} \log K_\upsilon(\sqrt{ab}) \tag{2.9}$$

where $R_\upsilon(z) := K_{\upsilon+1}(z)/K_\upsilon(z)$.

2.2.2 Centralized asymmetric Laplace distributions

Let $\mathbf{V}$ be a p-dimensional random vector from a centralized asymmetric Laplace (CAL) distribution (Kotz et al. 2001). The density of $\mathbf{V}$ is given by

$$f(\mathbf{v} \mid \boldsymbol{\alpha}, \boldsymbol{\Sigma}) = \frac{2\exp\{\mathbf{v}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}\}}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \times \left(\frac{\mathbf{v}'\boldsymbol{\Sigma}\mathbf{v}}{2 + \boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}\right)^{\upsilon/2} K_\upsilon(u) \tag{2.10}$$

where $\upsilon = (2-p)/2,\ u = \sqrt{(2 + \boldsymbol{\alpha}\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha})(\mathbf{v}'\boldsymbol{\Sigma}^{-1}\mathbf{v})}$, $\boldsymbol{\Sigma}$ is a scale matrix, and $\boldsymbol{\alpha} \in \mathbb{R}^p$ is a skewness parameter which allows the distribution to account for asymmetry in the data. The notation $\mathbf{V} \sim \mathcal{AL}_p(\boldsymbol{\alpha}, \boldsymbol{\Sigma})$ indicates the random variable $\mathbf{V}$ follows a p-dimensional CAL distribution.

2.2.3 Shifted asymmetric Laplace distributions

The CAL density is not effective for model-based clustering and classification since each component density will be centered at the same origin (Franczak et al. 2014). This is addressed for a random vector $\mathbf{V} \sim \mathcal{AL}_p(\boldsymbol{\alpha}, \boldsymbol{\Sigma})$ through the introduction of a shift parameter $\boldsymbol{\mu}$ by considering the random vector $\mathbf{X} = (\mathbf{V} + \boldsymbol{\mu})$. This random vector $\mathbf{X}$ follows a p-dimensional shifted asymmetric Laplace (SAL; Franczak et al. 2014) distribution, notated $\mathbf{X} \sim \mathcal{SAL}_p(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Sigma})$, with density given by

$$\xi(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) = \frac{2\ \exp\{(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}\}}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \times \left(\frac{\delta(\mathbf{x}, \boldsymbol{\mu}|\boldsymbol{\Sigma})}{2 + \boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}\right)^{\upsilon/2} K_{\upsilon}(u) \tag{2.11}$$

where $u = \sqrt{(2 + \boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha})\delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma})}$, $\delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma}) = (\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ and υ, $\boldsymbol{\alpha}$, and $\boldsymbol{\Sigma}$ are defined as before. It follows that the density for mixtures of SAL (MSAL; Franczak et al. 2014) distributions is given by

$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g\ \xi(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\alpha}_g, \boldsymbol{\Sigma}_g) \tag{2.12}$$

where all terms are as previously defined and $\xi(\mathbf{x})$ is given in Equation (2.11).

2.2.4 Relationship to the normal distribution

Kotz et al. (2001) show that a random vector $\mathbf{V} \sim \mathcal{AL}_p(\boldsymbol{\alpha}, \boldsymbol{\Sigma})$ can be generated through the relationship

$$\mathbf{V} = W\boldsymbol{\alpha} + \sqrt{W}\mathbf{N}$$

where $W \sim \text{Exp}(1)$ and $\mathbf{N} \sim \mathcal{N}_p(0, \boldsymbol{\Sigma})$ are independent of one another. Consequentially, the random vector $\mathbf{X} \sim \mathcal{SAL}_p(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Sigma})$ can be generated through the relationship

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

It follows that

$$\mathbf{X} \mid w \sim \mathcal{N}_p(\boldsymbol{\mu} + w\boldsymbol{\alpha}, w\boldsymbol{\Sigma})$$

and therefore from Bayes' theorem,

$$W \mid \mathbf{x} \sim GIG(a, b, v)$$

where $a = 2 + \boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha},\ b = \delta(\mathbf{x},\ \boldsymbol{\mu} \mid \boldsymbol{\Sigma})$ and $v = (2 - p)/2$.

2.2.5 Mixtures of contaminated SAL distributions

Morris et al. (2019) applies contamination to MSALs in a manner analogous to the approach used in Punzo and McNicholas (2016). For a random vector $\mathbf{X} \sim \mathcal{SAL}_p(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Sigma})$ the covariance is given by

$$\text{Cov}(\mathbf{X}) = \boldsymbol{\Sigma} + \boldsymbol{\alpha}\boldsymbol{\alpha}^T.$$

When a contamination scheme is applied to a SAL distribution, the covariance of the bad observations is inflated by a factor of η relative to the covariance of the good observations. Hence, if $\{\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}\}$ are the parameters of the good component the covariance for bad observations is given by

$$\text{Cov}(\mathbf{X}) = \eta(\boldsymbol{\Sigma} + \boldsymbol{\alpha}\boldsymbol{\alpha}^T) = \eta\boldsymbol{\Sigma} + \sqrt{\eta}\boldsymbol{\alpha}\sqrt{\eta}\boldsymbol{\alpha}^T$$

leading to the contaminated SAL (CSAL) distribution, the density of which is given by

$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \rho\xi(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) + (1 - \rho)\xi(\mathbf{x} \mid \boldsymbol{\mu}, \sqrt{\eta}\boldsymbol{\alpha}, \eta\boldsymbol{\Sigma}) \tag{2.13}$$

where $\boldsymbol{\vartheta}$ is the vector of parameters, ρ and η are as defined in Section 2.1.3, and $\xi(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Sigma})$ is as defined in Equation (2.11). It follow that the density for mixtures of CSAL distributions is given by

$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g \left[\rho_g\xi(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\alpha}_g, \boldsymbol{\Sigma}_g) + (1 - \rho_g)\xi(\mathbf{x}_g \mid \boldsymbol{\mu}_g, \sqrt{\eta_g}\boldsymbol{\alpha}_g, \eta_g\boldsymbol{\Sigma}_g)\right] \tag{2.14}$$

where all terms are as previously defined.

2.3 Parsimonious Mixture Models

2.3.1 Introduction

When estimating the covariance matrices $\mathbf{\Sigma}_g$ in a Gaussian mixture model, each matrix will have $p(p-1)/2$ free parameters. Since the number of free parameters scales quadratically with respect to the number of variables in a data set, estimation of the covariance matrices can become incredibly computationally intensive and tedious for higher dimensional data sets. McNicholas and Murphy (2008) impose a factor analyzer decomposition to the covariance matrix and introduce parsimonious constraints to that decomposition, causing the number of free parameters in $\mathbf{\Sigma}_g$ to scale linearly with respect the number of variables in a data set.

2.3.2 Factor analyzers

Factor analysis (Spearman 1904) is a data reduction technique that attempts to explain variability within a data set by replacing the observed variables with a reduced number of unobserved, but underlying, random quantities known as factors. The model assumes a p-dimensional random vector $\mathbf{X}$ is modelled using a q-dimensional random vector $\mathbf{U}$ with $q \ll p$ such that

$$\mathbf{X} = \boldsymbol{\mu} + \mathbf{\Lambda}\mathbf{U} + \boldsymbol{\varepsilon} \tag{2.15}$$

where $\mathbf{\Lambda}$ is a $p \times q$ matrix of factor loadings, $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \mathbf{I_q})$ is the vector of factors and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$ with $\boldsymbol{\Psi} = diag(\psi_1, \ldots, \psi_p)$. It follows that the marginal distribution of $\mathbf{X}$ is multivariate Gaussian with mean $\boldsymbol{\mu}$ and covariance $\mathbf{\Lambda}\mathbf{\Lambda}' + \boldsymbol{\Psi}$. The probabilistic principal component analysis (PPCA; Tipping and Bishop 1999b) model is a special case of the factor analysis model that assumes the distribution of the errors are isotropic, that is that $\boldsymbol{\Psi} = \psi\mathbf{I}_p$.

2.3.3 Mixtures of factor analyzers

The mixtures of factor analyzers (MFA) model introduced by Ghahramani and Hinton (1997) assumes mixtures of Gaussian distributions with a factor analysis covariance structure, the density of which is given by

$$f(\mathbf{x}|\ \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g \phi(\mathbf{x}|\boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g). \tag{2.16}$$

Tipping and Bishop (1999a) introduced a mixtures of PPCA models by assuming that the distribution of errors is isotropic with a covariance structure $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \psi_g\mathbf{I}_p$. McLachlan and Peel (2000b) further generalize the MFA by introducing the fully unconstrained covariance structure where $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g$.

2.3.4 Parsimonious Gaussian mixture models

The family of parsimonious Gaussian mixture models (PGMM) introduced by McNicholas and Murphy (2008) further extends the MFA model by allowing for constraints across groups on the $\boldsymbol{\Lambda}_g$ and $\boldsymbol{\Psi}_g$ matrices, as well as the isotropic constraint $\boldsymbol{\Psi}_g = \psi_g\mathbf{I}_p$. The possible combinations of these constraints provides eight distinct parsimonious models.

2.3.5 The expanded PGMM family

In McNicholas and Murphy (2010) the family of PGMMs is expanded upon by expressing the $\boldsymbol{\Psi}_g$ matrices as

$$\boldsymbol{\Psi}_g = \omega_g\boldsymbol{\Delta}_g$$

where $\omega_g \in \mathbb{R}$ and $\boldsymbol{\Delta}_g = diag(\delta_1, \ldots, \delta_p)$ such that $|\boldsymbol{\Delta}_g| = 1$ for $g = 1, 2 \ldots, G$. This results in the modified factor analysis covariance structure $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \omega_g\boldsymbol{\Delta}_g$. Constraints across groups can now be applied to the parameters ω_g and $\boldsymbol{\Delta}_g$ separately, resulting in four new models and extending the total number of models to twelve. The estimates for the eight pre-existing models are obtained from the PGMM estimates by setting $\omega_g = |\boldsymbol{\Psi}_g|^{1/p}$ and $\boldsymbol{\Delta}_g = \boldsymbol{\Psi}_g/|\boldsymbol{\Psi}_g|^{1/p}$. The nomenclature and constraint combinations for the twelve models in the family of extended parsimonious Gaussian mixture models (EPGMM), as well as there PGMM equivalents when applicable, is provided in Table 2.1. The covariance structure and the number of free parameters in this structure for all models in the EPGMM family is provided in Table 2.2.

2.3.6 Mixtures of contaminated Gaussian factor analyzers

In Punzo et al. (2020), the factor analysis covariance structure and the parameter constraints given in McNicholas and Murphy (2008) are applied to the mixtures of contaminated Gaussian distributions developed in Punzo and McNicholas (2016). This this resulted in a family of eight mixtures of contaminated Gaussian factor analyzers that are analogous to the models given in Table 2.1.

Table 2.1: Nomenclature for each member of the EPGMM family and its PGMM equivalent. (C = constrained, U = unconstrained)

	EPGMM Nomenclature				**PGMM Equivalent**	
Model ID	$\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$	$\boldsymbol{\Delta}_g = \boldsymbol{\Delta}$	$\omega_g = \omega$	$\boldsymbol{\Delta}_g = \mathbf{I}_p$	$\boldsymbol{\Psi}_g = \boldsymbol{\Psi}$	PGMM ID
CCCC	C	C	C	C	C	CCC
CCUC	C	C	U	C	U	CUC
CCCU	C	C	C	U	C	CCU
CCUU	C	C	U	U	-	-
CUCU	C	U	C	U	-	-
CUUU	C	U	U	U	U	CUU
UCCC	U	C	C	C	C	UCC
UCUC	U	C	U	C	C	UUC
UCCU	U	C	C	U	C	UCU
UCUU	U	C	U	U	-	-
UUCU	U	U	C	U	-	-
UUUU	U	U	U	U	U	UUU

Table 2.2: Covariance structure and number of covariance parameters for each member of the EPGMM family.

Model ID	Covariance Structure	Number of Covariance Parameters
CCCC	$\mathbf{\Sigma}_g = \mathbf{\Lambda}\mathbf{\Lambda}' + \omega \mathbf{I}_p$	$[pq - q(q-1)/2] + 1$
CCUC	$\mathbf{\Sigma}_g = \mathbf{\Lambda}\mathbf{\Lambda}' + \omega_g \mathbf{I}_p$	$[pq - q(q-1)/2] + G$
CCCU	$\mathbf{\Sigma}_g = \mathbf{\Lambda}\mathbf{\Lambda}' + \omega \mathbf{\Delta}$	$[pq - q(q-1)/2] + p$
CCUU	$\mathbf{\Sigma}_g = \mathbf{\Lambda}\mathbf{\Lambda}' + \omega_g \mathbf{\Delta}$	$[pq - q(q-1)/2] + [G + (p-1)]$
CUCU	$\mathbf{\Sigma}_g = \mathbf{\Lambda}\mathbf{\Lambda}' + \omega \mathbf{\Delta}_g$	$[pq - q(q-1)/2] + [1 + G(p-1)]$
CUUU	$\mathbf{\Sigma}_g = \mathbf{\Lambda}\mathbf{\Lambda}' + \omega_g \mathbf{\Delta}_g$	$[pq - q(q-1)/2] + Gp$
UCCC	$\mathbf{\Sigma}_g = \mathbf{\Lambda}_g\mathbf{\Lambda}_g' + \omega \mathbf{I}_p$	$G[pq - q(q-1)/2] + 1$
UCUC	$\mathbf{\Sigma}_g = \mathbf{\Lambda}_g\mathbf{\Lambda}_g' + \omega_g \mathbf{I}_p$	$G[pq - q(q-1)/2] + G$
UCCU	$\mathbf{\Sigma}_g = \mathbf{\Lambda}_g\mathbf{\Lambda}_g' + \omega \mathbf{\Delta}$	$G[pq - q(q-1)/2] + p$
UCUU	$\mathbf{\Sigma}_g = \mathbf{\Lambda}_g\mathbf{\Lambda}_g' + \omega_g \mathbf{\Delta}$	$G[pq - q(q-1)/2] + [G + (p-1)]$
UUCU	$\mathbf{\Sigma}_g = \mathbf{\Lambda}_g\mathbf{\Lambda}_g' + \omega \mathbf{\Delta}_g$	$G[pq - q(q-1)/2] + [1 + G(p-1)]$
UUUU	$\mathbf{\Sigma}_g = \mathbf{\Lambda}_g\mathbf{\Lambda}_g' + \omega_g \mathbf{\Delta}_g$	$G[pq - q(q-1)/2] + Gp$

2.3.7 Woodbury identity

The Woodbury identity (Woodbury 1950) states that given a $p \times p$ matrix $\mathbf{A}$, a $p \times q$ matrix $\mathbf{U}$, a $q \times q$ matrix $\mathbf{B}$, and a $q \times p$ matrix $\mathbf{V}$ such that the matrix $(\mathbf{A} + \mathbf{UBV})$ is invertible, then that inverse can be expressed as:

$$(\mathbf{A} + \mathbf{UBV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{B}^{-1} + \mathbf{V}\mathbf{A}^{-1}\mathbf{U})^{-1}\mathbf{V}\mathbf{A}^{-1}$$

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

$$(\mathbf{\Lambda}_g\mathbf{\Lambda}_g' + \mathbf{\Psi}_g)^{-1} = \mathbf{\Psi}_g^{-1} - \mathbf{\Psi}_g^{-1}\mathbf{\Lambda}_g(\mathbf{I}_q^{-1} + \mathbf{\Lambda}_g'\mathbf{\Psi}_g^{-1}\mathbf{\Lambda}_g)^{-1}\mathbf{\Lambda}_g'\mathbf{\Psi}_g^{-1}. \qquad (2.17)$$

When modelling high-dimensional data, calculating the inverse of the $p \times p$ covariance matrices $\mathbf{\Sigma}_g$ can be very computationally expensive and impractical. By re-expressing the covariance matrix via the Woodbury identity, the calculation now only requires the inversion of the diagonal $p \times p$ matrix $\mathbf{\Psi}_g$ and the $q \times q$ matrix $(\mathbf{I}_q^{-1} + \mathbf{\Lambda}_g' \mathbf{\Psi}_g^{-1} \mathbf{\Lambda}_g)$. Additionally, Woodbury (1950) also provides the identity for calculating the determinant

$$|\mathbf{\Lambda}_g \mathbf{\Lambda}_g' + \mathbf{\Psi}_g| = |\mathbf{\Psi}_g| \;/\; |\mathbf{I}_q - (\mathbf{\Lambda}_g'(\mathbf{\Lambda}_g \mathbf{\Lambda}_g' + \mathbf{\Psi}_g)^{-1} \mathbf{\Lambda}_g)|. \tag{2.18}$$

2.4 The Expectation Maximization Algorithm

An expectation-maximization (EM) algorithm (Dempster et al. 1977) is an iterative procedure that is used to find maximum likelihood estimates (MLE) when data is either incomplete or assumed to be incomplete. The complete-data is considered to consist of the observed and missing data. The algorithm operates by alternating between two steps, an expectation step (E-step) and a maximization-step (M-step). In the E-step, the expected value of the complete-data log-likelihood is calculated using the parameter estimates from the preceding M-step. In the M-step, the expected value of complete-data log-likelihood is maximized with respect to the model parameters. This process is iterated until convergence has been reached.

2.4.1 Expectation conditional-maximization

The expectation-conditional maximization (ECM) algorithm (Meng and Rubin 1993) is a modification of the EM algorithm, where the M-step is replaced by multiple conditional-maximization-steps (CM-step). The set of parameters is partitioned $\boldsymbol{\vartheta} = \{\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2, \ldots \boldsymbol{\vartheta}_m\}$ such that each subset of the partition correlates to a CM-step. In each CM-step, the complete-data log-likelihood is maximized with respect to the model parameters in its respective partition with the other parameters fixed at their most recent estimates.

For illustrative purposes we consider the partition $\boldsymbol{\vartheta} = \{\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2\}$ at a given iteration $(k+1)$. In the first CM-step, the estimates $\hat{\boldsymbol{\vartheta}}_1^{(k+1)}$ are calculated to maximize the complete-data log-likelihood with $\boldsymbol{\vartheta}_2$ fixed at $\hat{\boldsymbol{\vartheta}}_2^{(k)}$. In the second CM-step, the estimates $\hat{\boldsymbol{\vartheta}}_2^{(k+1)}$ are calculated to maximize the complete-data log-likelihood with $\boldsymbol{\vartheta}_1$ fixed at $\hat{\boldsymbol{\vartheta}}_1^{(k+1)}$.

2.4.2 Alternating expectation conditional-maximization

The alternating expectation-conditional maximization (AECM) algorithm (Meng and Van Dyk 1997) is an extension of the ECM algorithm that allows for the specification of the complete-data to change between CM-steps. The AECM algorithm is used in this thesis to implement the parsimonious contaminated SAL mixtures developed in Chapter 3. McLachlan and Krishnan (2008) provides an extensive overview of the AECM and its application to fitting mixtures of factor analyzers models.

2.4.3 Aitken's acceleration stopping criterion

Convergence of an EM algorithm is typically evaluated by comparing successive estimates until a point where improvement is deemed sufficiently small or negligible. However, when an EM algorithm is applied for mixture model fitting, the log-likelihood can often experience "speed bumps" in the rate at which the log-likelihood increases resulting in a false sense of stability. Therefore, Aitken's acceleration (Aitken 1926) is used to estimate the asymptotic maximum of the log-likelihood at each iteration. Let $l^{(k)}$ be the log-likelihood at iteration k, then the Aitken acceleration at iteration k is given by

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

The asymptotic estimate of the log-likelihood at iteration $(k+1)$ is then given by

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

in Böhning et al. (1994). This thesis uses the convergence criterion proposed by McNicholas et al. (2010), in which the EM algorithm is considered to have converged on an iteration $(k+1)$ if

$$l_{\infty}^{(k+1)} - l^{(k)} < \varepsilon.$$

where $\varepsilon \in \mathbb{R}$ is some small constant. More specifically $\varepsilon = 10^{-2}$ is used.

2.5 Model Selection and Performance

2.5.1 Bayesian information criterion

The Bayesian information criteriom (BIC; Schwarz 1978) is one of the most widely used methods for model selection. For a model with parameter $\boldsymbol{\vartheta}$, the BIC is calculated as

$$\text{BIC} = 2l(x, \hat{\boldsymbol{\vartheta}}) - p\text{log}(n) \tag{2.19}$$

where $l(x, \hat{\boldsymbol{\vartheta}})$ is the maximized log-likelihood, $\hat{\boldsymbol{\vartheta}}$ is the MLEs of $\boldsymbol{\vartheta}$, p is the number of free parameters, and n is the number of observations. Fraley and Raftery (1998, 2002) provide practical evidence that BIC is an effective model selection criterion for mixture models.

2.5.2 Integrated complete likelihood

The integrated completed likelihood (ICL; Biernacki et al. 2000) is essentially an extension of the BIC, designed specifically for clustering and classification applications. First, we introduce a component membership label

$$Z_{ig} = \begin{cases} 1 & \text{if observation } i \text{ is in group } g \\ 0 & \text{otherwise} \end{cases}$$

then, the ICL of a model is calculated as

$$\text{ICL} = \text{BIC} + \sum_{i=1}^{n} \sum_{g=1}^{G} \text{MAP}(z_{ig})\text{log}(z_{ig})$$

where z_{ig} is the expected value of Z_{ig}, $\text{MAP}(z_{ig})$ is the maximum *a posteriori* classification given by z_{ig}, and BIC is defined as in Equation (2.19). The ICL penalizes the BIC by subtracting a measure of the estimated entropy, or the uncertainty in the classification of observations into components.

2.5.3 Rand and adjusted Rand indices

The Rand index (Rand 1971) is used to compare partitions and is given by the proportion of pairwise agreements out of the total number of pairs. In classification applications, this manifests as the proportion of correctly classified observations. The Rand index takes on a value belonging to $[0, 1]$ with 1 indicating perfect classification and 0 indicating that no observations were classified correctly. It follows that expected value of the Rand index under random assignment will be given by $1/G$. The adjusted Rand index (ARI; Hubert and Arabie 1985) corrects for this issue so that the expected value under random classification is 0, but still takes a value of 1 under perfect classification. If classification performance is worse than would be expected under random assignment, the ARI will take on a negative value. Consequentially, the ARI will penalize incorrect classifications more harshly than the Rand index.

Chapter 3

Methodology

3.1 Introduction

This thesis introduces a family of parsimonious contaminated SAL mixtures (PC-SALM) by imposing the constraints contained in McNicholas and Murphy (2008, 2010) to the modified factor analysis decomposition of the scale matrix parameter $\mathbf{\Sigma}_g$ in the MCSAL model given in Equation (2.14). This leads to a family of twelve models with scale matrix structures analogous to the twelve covariance structures provided in Table 2.2.

3.2 Parsimonious Contaminated SAL Mixtures

The CSAL density given in Equation (2.13) can be alternatively expressed using the relationships in Section 2.2.4 as

$$\begin{aligned} f_{CSAL}(\mathbf{x} \mid \boldsymbol{\vartheta}) = \rho \int_0^\infty & \phi(\mathbf{x} \mid \boldsymbol{\mu} + w\boldsymbol{\alpha}, w\mathbf{\Sigma})h(w)dw \\ & + (1-\rho)\int_0^\infty \phi(\mathbf{x} \mid \boldsymbol{\mu} + w\sqrt{\eta}\boldsymbol{\alpha}, w\eta\mathbf{\Sigma})h(w)dw \qquad (3.1) \end{aligned}$$

where w is an exponential random variable, $h(w)$ is the density of an exponential random variables with rate 1, the parameters $\boldsymbol{\vartheta} = \{\rho, \eta, \boldsymbol{\mu}, \boldsymbol{\alpha}, \mathbf{\Sigma}\}$ are defined as in Equation (2.13), and $\phi(\mathbf{x} \mid\boldsymbol{\mu}, \mathbf{\Sigma})$ is defined as in Equation (2.3). It follows that the density for mixtures of CSAL distribution given in Equation (2.14), can also be reexpressed as

$$
\begin{aligned}
f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \ \pi_g \Bigg[\rho_g \int_0^\infty \phi(\mathbf{x} \mid \boldsymbol{\mu}_g + w_g\boldsymbol{\alpha}_g, w_g\boldsymbol{\Sigma}_g)h(w_g)dw_g \\
+ \ (1-\rho_g)\int_0^\infty \phi(\mathbf{x} \mid \boldsymbol{\mu}_g + w_g\sqrt{\eta_g}\boldsymbol{\alpha}_g, w_g\eta_g\boldsymbol{\Sigma}_g)h(w_g)dw_g\Bigg]
\end{aligned}
\qquad (3.2)
$$

where π_g is as defined in Equation (2.13) and all other parameters are as previously defined.

We can now apply the modified factor analyzer decomposition $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \omega_g\boldsymbol{\Delta}_g$ described in McNicholas and Murphy (2010) to our scale matrices. This gives us the density for a mixture of unconstrained parsimonious CSAL distributions

$$
\begin{aligned}
f(\mathbf{x}|\boldsymbol{\vartheta}) = \sum_{g=1}^{G} \ \pi_g \Bigg[\rho_g \int_0^\infty \phi(\mathbf{x}|\boldsymbol{\mu}_g + w_g\boldsymbol{\alpha}_g, w_g(\boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \omega_g\boldsymbol{\Delta}_g))h(w_g)dw_g \\
+ \ (1-\rho_g)\int_0^\infty \phi(\mathbf{x}|\boldsymbol{\mu}_g + w_g\sqrt{\eta_g}\boldsymbol{\alpha}_g, w_g\eta_g(\boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \omega_g\boldsymbol{\Delta}_g))h(w_g)dw_g\Bigg]
\end{aligned}
\qquad (3.3)
$$

Applying the constraints given in Table 2.1 to the parameters $\boldsymbol{\Lambda}_g, \omega_g, \boldsymbol{\Delta}_g$ produces the family of parsimonious CSAL mixtures (PCSALM). Reformatting our density in this way also allows us to more easily obtain the maximum likelihood estimates from the complete-data log-likelihood.

To properly express the complete data log- likelihood, we must first account for the missing data. In our case, there are four sources of incompleteness. For each observation vector $\mathbf{x}_i$ we do not know its component membership or its status as a good or bad observation. We introduce two indicator variables: Z_{ig} and V_{ig} defined as

$$
Z_{ig} = \begin{cases} 1 & \text{if observation } \mathbf{x}_i \text{ belongs to group } g, \\ 0 & \text{otherwise,} \end{cases}
$$

$$
V_{ig} = \begin{cases} 1 & \text{if observation } \mathbf{x}_i \text{ in group } g \text{ is good,} \\ 0 & \text{if observation } \mathbf{x}_i \text{ in group } g \text{ is bad.} \end{cases}
$$

The other two sources of missing data in each PCSALM model are the latent weight variables w_{ig} and the latent factors $\mathbf{u}_{ig}$. The complete-data log-likelihood for Equation (3.3) is given by

$$l_C(\boldsymbol{\vartheta}) = l_{C1}(\boldsymbol{\pi}) + l_{C2}(\boldsymbol{\rho}) + l_{C3}^{\text{good}}(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Lambda}, \omega, \boldsymbol{\Delta}) + l_{C3}^{\text{bad}}(\boldsymbol{\eta}, \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Lambda}, \omega, \boldsymbol{\Delta}) \tag{3.4}$$

where

$$l_{C1}(\boldsymbol{\pi}) = \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\log(\pi_g),$$

$$l_{C2}(\boldsymbol{\rho}) = \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\left[v_{ig}\log(\rho_g) + (1-v_{ig})\log(1-\rho_g)\right],$$

$$l_{C3}^{\text{good}}(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Lambda}, \omega, \boldsymbol{\Delta}) = \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}v_{ig}\log\left[\phi(\mathbf{x}|\boldsymbol{\mu}_g + w_{ig}\boldsymbol{\alpha}_g, w_{ig}(\boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \omega_g\boldsymbol{\Delta}_g))h(w_{ig})\right],$$

$$\begin{aligned} l_{C3}^{\text{bad}}(\boldsymbol{\eta}, \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Lambda}, \omega, \boldsymbol{\Delta}) = &\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}(1-v_{ig}) \\ &\times \log\left[\phi(\mathbf{x}|\boldsymbol{\mu}_g + w_{ig}\sqrt{\eta_g}\boldsymbol{\alpha}_g, w_{ig}\eta_g(\boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \omega_g\boldsymbol{\Delta}_g))h(w_{ig})\right]. \end{aligned}$$

3.3 The AECM algorithm for PCSALM

The parameters $\boldsymbol{\vartheta}$ for all models in the PCSALM family are partitioned into the sets $\boldsymbol{\vartheta} = \{\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2\}$ where $\boldsymbol{\vartheta}_1 = \{\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\eta}\}$ and $\boldsymbol{\vartheta}_2 = \{\boldsymbol{\Lambda}, \boldsymbol{\omega}, \boldsymbol{\Delta}\}$. The parameter set $\boldsymbol{\vartheta}_1$ is further partitioned such that $\boldsymbol{\vartheta}_{11} = \{\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\alpha}\}$ and $\boldsymbol{\vartheta}_{12} = \{\boldsymbol{\eta}\}$. The parameter set $\boldsymbol{\vartheta}_1$ corresponds to the first alternation in the AECM and the parameter set $\boldsymbol{\vartheta}_2$ corresponds to the second alternation. The parameter subsets $\boldsymbol{\vartheta}_{11}$ and $\boldsymbol{\vartheta}_{12}$ correspond to the first and second CM-steps respectively within the first alternation. For a model with G groups $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_G)$, $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_G)$, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G)$, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_G)$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_G)$, $\boldsymbol{\Lambda} = (\boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_G)$, $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_G)$, and $\boldsymbol{\Delta} = (\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_G)$ where the appropriate constraints for each model are applied to the parameter sets $\boldsymbol{\Lambda}, \boldsymbol{\omega}, \boldsymbol{\Delta}$.

Calculations for the expected values and parameter updates in the first alternation follow from Franczak et al. (2014) and Morris et al. (2019). Calculations for the expected values and parameter updates in the second alternation follow from McNicholas and Murphy (2008, 2010) and Punzo et al. (2020).

3.3.1 Alternation 1: E-step

In the E-step of the first alteration of our AECM algorithm, the complete-data consists of the observed data $\mathbf{x}_i$, the component membership labels Z_{ig}, the good observation labels V_{ig} and the latent weight variables W_{ig} for $i = 1, \ldots, n$ and $g = 1, \ldots, G$. The expected values for the missing data on iteration k are given by

$$z_{ig}^{(k)} := \mathbb{E}[Z_{ig} \mid \mathbf{x}_i] = \frac{\pi_g^{(k)} f_{\text{CSAL}}(\mathbf{x}_i \mid \rho_g^{(k)}, \eta_g^{(k)}, \boldsymbol{\mu}_g^{(k)}, \boldsymbol{\alpha}_g^{(k)}, \boldsymbol{\Sigma}_g^{(k)})}{\sum_{h=1}^{G} \pi_h^{(k)} f_{\text{CSAL}}(\mathbf{x}_i \mid \rho_h^{(k)}, \eta_h^{(k)}, \boldsymbol{\mu}_h^{(k)}, \boldsymbol{\alpha}_h^{(k)}, \boldsymbol{\Sigma}_h^{(k)})}$$

$$v_{ig}^{(k)} := \mathbb{E}[V_{ig} \mid \mathbf{x}_i] = \frac{\rho_g^{(k)} \xi(\mathbf{x}_i \mid \boldsymbol{\mu}_g^{(k)}, \boldsymbol{\alpha}_g^{(k)}, \boldsymbol{\Sigma}_g^{(k)})}{f_{\text{CSAL}}(\mathbf{x}_i \mid \rho_g^{(k)}, \eta_g^{(k)}, \boldsymbol{\mu}_g^{(k)}, \boldsymbol{\alpha}_g^{(k)}, \boldsymbol{\Sigma}_g^{(k)})}$$

$$E_{1ig}^{(k)} := \mathbb{E}[W_{ig} \mid \mathbf{x}_i, Z_{ig} = 1] = \sqrt{\frac{b_{ig}^{(k)}}{a_g^{(k)}}} R_\upsilon \left(\sqrt{a_g^{(k)} b_{ig}^{(k)}} \right)$$

$$E_{2ig}^{(k)} := \mathbb{E}[1/W_{ig} \mid \mathbf{x}_i, Z_{ig} = 1] = \sqrt{\frac{a_g^{(k)}}{b_{ig}^{(k)}}} R_\upsilon \left(\sqrt{a_g^{(k)} b^{(k)} ig} \right) - \frac{2\upsilon}{b_{ig}^{(k)}}$$

$$E_{3ig}^{(k)} := \mathbb{E}[\log W_{ig} \mid \mathbf{x}_i, Z_{ig} = 1] = \log \sqrt{\frac{a_g^{(k)}}{b_{ig}^{(k)}}} + \frac{\partial}{\partial \upsilon} \log K_\upsilon \left(\sqrt{a_g^{(k)} b_{ig}^{(k)}} \right)$$

$$\widetilde{E}_{1ig}^{(k)} := \mathbb{E}\left[\widetilde{W}_{ig} \mid \mathbf{x}_i, Z_{ig} = 1\right] = \sqrt{\frac{\widetilde{b}_{ig}^{(k)}}{a_g^{(k)}}} R_\upsilon \left(\sqrt{a_g^{(k)} \widetilde{b}_{ig}^{(k)}} \right)$$

$$\widetilde{E}_{2ig}^{(k)} := \mathbb{E}\left[1/\widetilde{W}_{ig} \mid \mathbf{x}_i, Z_{ig} = 1\right] = \sqrt{\frac{a_g^{(k)}}{\widetilde{b}_{ig}^{(k)}}} R_\upsilon \left(\sqrt{a_g^{(k)} \widetilde{b}_{ig}^{(k)}} \right) - \frac{2\upsilon}{\widetilde{b}_{ig}^{(k)}}$$

$$\widetilde{E}^{(k)}_{3ig} := \mathbb{E}[\log \widetilde{W}_{ig} \mid \mathbf{x}_i, Z_{ig} = 1] = \log\sqrt{\frac{a_g^{(k)}}{\widetilde{b}_{ig}^{(k)}}} + \frac{\partial}{\partial \upsilon}\log\, K_\upsilon\left(\sqrt{a_g^{(k)}\widetilde{b}_{ig}^{(k)}}\right)$$

where $a_g^{(k)} = 2 + \boldsymbol{\alpha}_g^{(k)\prime}(\boldsymbol{\Sigma}_g^{(k)})^{-1}\boldsymbol{\alpha}_g^{(k)},\ b_{ig}^{(k)} = \delta(\mathbf{x}_i, \boldsymbol{\mu}_g^{(k)}|\boldsymbol{\Sigma}_g^{(k)}),\ \widetilde{b}_{ig}^{(k)} = \delta(\mathbf{x}_i, \boldsymbol{\mu}_g^{(k)}|\eta_g^{(k)}\boldsymbol{\Sigma}_g^{(k)}),$ $R_\upsilon(z) := K_{\upsilon+1}(z)/K_\upsilon(z), \upsilon = (2-p)/2$ and $K_\upsilon(z)$ is a modified Bessel function of the third kind. The closed form updates for $E_{1ig},\ E_{2ig},\ E_{3ig},\ \widetilde{E}_{1ig},\ \widetilde{E}_{2ig},\ \widetilde{E}_{3ig}$ follow from $W_{ig}|\mathbf{x}_i, z_{ig} = 1 \sim GIG(a_g, b_{ig}, \upsilon)$ as shown in Section 2.2.1.

Using these expected values for the missing data and the logarithmic expression of the Gaussian density, we can compute the complete-data log-likelihood for the first alternation as

$$\mathcal{Q}(\boldsymbol{\vartheta}) = \mathcal{Q}_1(\boldsymbol{\pi}) + \mathcal{Q}_2(\boldsymbol{\rho}) + \mathcal{Q}_3^{\text{good}}(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) + \mathcal{Q}_3^{\text{bad}}(\boldsymbol{\eta}, \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) + \mathcal{Q}_4$$

where

$$\mathcal{Q}_1(\boldsymbol{\pi}) = \sum_{g=1}^{G} n_g \log(\pi_g) \tag{3.5}$$

$$\mathcal{Q}_2(\boldsymbol{\rho}) = \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\left[v_{ig}\log(\rho_g) + (1 - v_{ig})\log(1 - \rho_g)\right] \tag{3.6}$$

$$\begin{aligned}\mathcal{Q}_3^{\text{good}}(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) = &-\frac{np}{2}\log(2\pi) - \frac{1}{2}\sum_{g=1}^{G} n_{g,\text{good}}\log|\boldsymbol{\Sigma}_g| - \frac{p}{2}\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}v_{ig}E_{3ig} \\ &- \frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}v_{ig}E_{2ig}(\mathbf{x}_i - \boldsymbol{\mu}_g)'\boldsymbol{\Sigma}_g^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_g) \\ &+ \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}v_{ig}(\mathbf{x}_i - \boldsymbol{\mu}_g)'\boldsymbol{\Sigma}_g^{-1}\boldsymbol{\alpha}_g \\ &- \frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}v_{ig}E_{1ig}\boldsymbol{\alpha}_g'\boldsymbol{\Sigma}_g^{-1}\boldsymbol{\alpha}_g\end{aligned} \tag{3.7}$$

$$\begin{aligned}\mathcal{Q}_3^{\text{bad}}(\boldsymbol{\eta}, \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) = &- \frac{np}{2}\log(2\pi) - \frac{1}{2}\sum_{g=1}^{G} n_{g,\text{bad}}\log|\boldsymbol{\Sigma}_g| \\ &- \frac{p}{2}\sum_{g=1}^{G} n_{g,\text{bad}}\log(\eta_g) - \frac{p}{2}\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}(1-v_{ig})\widetilde{E}_{3ig} \\ &- \frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}(1-v_{ig})\widetilde{E}_{2ig}\frac{1}{\eta_g}(\mathbf{x}_i - \boldsymbol{\mu}_g)'\boldsymbol{\Sigma}_g^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_g) \\ &+ \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}(1-v_{ig})\frac{1}{\sqrt{\eta_g}}(\mathbf{x}_i - \boldsymbol{\mu}_g)'\boldsymbol{\Sigma}_g^{-1}\boldsymbol{\alpha}_g \\ &- \frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}(1-v_{ig})\widetilde{E}_{1ig}\boldsymbol{\alpha}_g'\boldsymbol{\Sigma}_g^{-1}\boldsymbol{\alpha}_g \end{aligned} \tag{3.8}$$

and

$$\mathcal{Q}_4 = -\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\left[v_{ig}E_{1ig} + (1-v_{ig})\widetilde{E}_{1ig}\right]$$

where all parameters and expected values are assumed to be their respective estimates for iteration k, $n_g = \sum_{i=1}^{n} z_{ig}$ is the expected number of observations in group g, $n_{g,\text{good}} = \sum_{i=1}^{n} z_{ig}v_{ig}$ is the expected number of good observations in group g, and $n_{g,\text{bad}} = \sum_{i=1}^{n} z_{ig}(1-v_{ig})$ is the expected number of bad observations in group g.

3.3.2 Alternation 1: CM-step 1

In the first CM step of alternation one, the updates for the parameters in the subset $\boldsymbol{\vartheta}_{11} = \{\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\alpha}\}$ on iteration $(k+1)$ are given by

$$\pi_g^{(k+1)} = \frac{n_g^{(k)}}{n},$$

$$\rho_g^{(k+1)} = \frac{n_{g,\text{good}}^{(k)}}{n_g^{(k)}},$$

$$\boldsymbol{\mu}_g^{(k+1)} = \frac{B^{(k)}\left[\sum_{i=1}^{n} a_{ig}^{(k)}\mathbf{x}_i\right] - D^{(k)}\left[\sum_{i=1}^{n} z_{ig}^{(k)}\left(v_{ig}^{(k)} + \frac{1-v_{ig}^{(k)}}{\eta_g^{(k)}}\right)\mathbf{x}_i\right]}{B^{(k)}A^{(k)} - (D^{(k)})^2},$$

$$\boldsymbol{\alpha}_g^{(k+1)} = \frac{A^{(k)}\left[\sum_{i=1}^n z_{ig}^{(k)}\left(v_{ig}^{(k)} + \frac{1-v_{ig}^{(k)}}{\sqrt{\eta_g^{(k)}}}\right)\mathbf{x}_i\right] - D^{(k)}\left[\sum_{i=1}^n a_{ig}^{(k)}\mathbf{x}_i\right]}{B^{(k)}A^{(k)} - (D^{(k)})^2}$$

where

$$a_{ig}^{(k)} = z_{ig}^{(k)}\left(v_{ig}^{(k)}E_{2ig}^{(k)} + \frac{1-v_{ig}^{(k)}}{\eta_g^{(k)}}\widetilde{E}_{2ig}^{(k)}\right), \quad A^{(k)} = \sum_{i=1}^n z_{ig}^{(k)}\left(v_{ig}^{(k)}E_{2ig}^{(k)} + \frac{1-v_{ig}^{(k)}}{\eta_g^{(k)}}\widetilde{E}_{2ig}^{(k)}\right),$$

$$B^{(k)} = \sum_{i=1}^n z_{ig}^{(k)}\left(v_{ig}^{(k)}E_{1ig}^{(k)} + \frac{1-v_{ig}^{(k)}}{\eta_g^{(k)}}\widetilde{E}_{1ig}^{(k)}\right) \quad \text{and} \quad D^{(k)} = \sum_{i=1}^n z_{ig}^{(k)}\left(v_{ig}^{(k)} + \frac{1-v_{ig}^{(k)}}{\sqrt{\eta_g^{(k)}}}\right).$$

3.3.3 Alternation 1: CM-step 2

In the second CM step of alternation one, closed form updates for the parameters in the subset $\boldsymbol{\vartheta}_{12} = \{\boldsymbol{\eta}\}$ on iteration $(k+1)$ are obtained by differentiating $\mathcal{Q}(\boldsymbol{\vartheta})$ with respect to η_g for $g = (1, \ldots, G)$:

$$\begin{aligned}\frac{\partial}{\partial \eta_g}\mathcal{Q}(\boldsymbol{\vartheta}) = &-\frac{p}{2\eta_g}n_{g,\text{good}} + \frac{1}{2\eta_g^2}\sum_{i=1}^n z_{ig}(1-v_{ig})\widetilde{E}_{2ig}(\mathbf{x}_i - \boldsymbol{\mu}_g)'\boldsymbol{\Sigma}_g^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_g)\\ &-\frac{1}{2\eta_g^{3/2}}\sum_{i=1}^n z_{ig}(1-v_{ig})(\mathbf{x}_i - \boldsymbol{\mu}_g)'\boldsymbol{\Sigma}_g^{-1}\boldsymbol{\alpha}_g.\end{aligned}$$

Setting the partial derivative to zero and multiplying by $-2\eta_g^2$ allows us to express this equation as

$$0 = a_g\eta_g + b_g\sqrt{\eta_g} + c_g$$

where, for iteration $(k+1)$

$$a_g^{(k+1)} = p(n_{g,\text{good}}^{(k)}), \quad b_g^{(k+1)} = \sum_{i=1}^n z_{ig}^{(k)}(1-v_{ig}^{(k)})(\mathbf{x}_i - \boldsymbol{\mu}_g^{(k+1)})'(\boldsymbol{\Sigma}_g^{(k)})^{-1}\boldsymbol{\alpha}_g^{(k+1)}$$

$$\text{and} \quad c_g^{(k+1)} = -\sum_{i=1}^n z_{ig}^{(k)}(1-v_{ig}^{(k)})\widetilde{E}_{2ig}^{(k)}(\mathbf{x}_i - \boldsymbol{\mu}_g^{(k+1)})'(\boldsymbol{\Sigma}_g^{(k)})^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_g^{(k+1)}).$$

By using the quadratic formula to find the real positive root to this equation, a closed form solution for η_g can then be obtained by

$$\eta_g^* = \left(\frac{-b_g^{(k+1)} + \sqrt{\left(b_g^{(k+1)}\right)^2 - 4a_g^{(k+1)}c_g^{(k+1)}}}{2a_g^{(k+1)}}\right)^2.$$

A lower limit of 1 is set to ensure the integrity of the parameter, hence the update for η_g is then given by

$$\eta_g^{(k+1)} = \max\{1, \eta_g^*\}.$$

3.3.4 Alternation 2: E-step

In the E-step of the second alternation, the missing data consists of the three previous sources of incompleteness and the missing latent factors $\mathbf{U}_{ig}$. The expected values z_{ig}, v_{ig} and the series of expected values related to W_{ig} are updated in a similar manner to Section 3.3.1, but they are calculated with the parameter set $\{\boldsymbol{\vartheta}_1^{(k+1)}, \boldsymbol{\vartheta}_2^{(k)}\}$ for iteration $(k+1)$. The expected values calculated on this step for iteration $(k+1)$ are denoted by the superscript $(k+1/2)$. We can express the complete log likelihood in terms of the latent factors $\mathbf{U}_{ig}$ as

$$\begin{aligned} l_{2C}(\boldsymbol{\vartheta}_2) =& C + \sum_{g=1}^{G} \Bigg\{ -\frac{n_g}{2}\log|\boldsymbol{\Psi}_g| - \frac{n_g}{2}\mathrm{tr}(\boldsymbol{\Psi}_g^{-1}\mathbf{S}_g^{(k+1)}) \\ &+ \sum_{i=1}^{n} z_{ig}\left(v_{ig}E_{2ig} + \frac{1-v_{ig}}{\eta_g^{(k+1)}}\widetilde{E}_{2ig}\right)(\mathbf{x}_i - \boldsymbol{\mu}_g^{(k+1)})'\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}_g\mathbf{u}_{ig} \\ &- \sum_{i=1}^{n} z_{ig}\left(v_{ig} + \frac{1-v_{ig}}{\sqrt{\eta_g^{(k+1)}}}\right)(\boldsymbol{\alpha}_g^{(k+1)})'\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}_g\mathbf{u}_{ig} \\ &- \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Lambda}_g'\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}_g\sum_{i=1}^{n} z_{ig}\left(v_{ig}E_{2ig} + \frac{1-v_{ig}}{\eta_g^{(k+1)}}\widetilde{E}_{2ig}\right)\mathbf{u}_{ig}\mathbf{u}_{ig}'\right]\Bigg\}. \end{aligned} \tag{3.9}$$

where C is a constant with respect to $\boldsymbol{\vartheta}_2$, $\boldsymbol{\Psi}_g = \omega_g\boldsymbol{\Delta}_g$ and the matrices $\mathbf{S}_g^{(k+1)}$ are given by

$$\begin{aligned} \mathbf{S}_g^{(k+1)} =& \frac{1}{n_g}\sum_{i=1}^{n} z_{ig}\left(v_{ig}E_{2ig} + \frac{1-v_{ig}}{\eta_g^{(k+1)}}\widetilde{E}_{2ig}\right)(\mathbf{x}_i - \boldsymbol{\mu}_g^{(k+1)})(\mathbf{x}_i - \boldsymbol{\mu}_g^{(k+1)})' \\ &- \frac{2}{n_g}\sum_{i=1}^{n} z_{ig}\left(v_{ig} + \frac{1-v_{ig}}{\sqrt{\eta_g^{(k+1)}}}\right)(\mathbf{x}_i - \boldsymbol{\mu}_g^{(k+1)})(\boldsymbol{\alpha}_g^{(k+1)})' \\ &+ \frac{1}{n_g}\sum_{i=1}^{n} z_{ig}\left[v_{ig}E_{1ig} + (1-v_{ig})\widetilde{E}_{1ig}\right]\boldsymbol{\alpha}_g^{(k+1)}(\boldsymbol{\alpha}_g^{(k+1)})'. \end{aligned} \tag{3.10}$$

Recalling the latent factor model defined in Equation (2.15), if the observation vector $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$, then the expectations for the latent factor $\mathbf{U}$ and the outer product $\mathbf{U}\mathbf{U}'$ are given by

$$\mathbb{E}\big[\mathbf{U} \mid \mathbf{X}\big] = \boldsymbol{\beta}(\mathbf{X} - \boldsymbol{\mu}^*) \tag{3.11}$$

$$\mathbb{E}\big[\mathbf{U}\mathbf{U}' \mid \mathbf{X}\big] = \mathbf{I}_q + \boldsymbol{\beta}\boldsymbol{\Lambda} + \boldsymbol{\beta}\big[(\mathbf{X} - \boldsymbol{\mu}^*)(\mathbf{X} - \boldsymbol{\mu}^*)\big]\boldsymbol{\beta}' \tag{3.12}$$

where $\boldsymbol{\beta} = \boldsymbol{\Lambda}'(\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})^{-1}$. Using Equations (3.11) and (3.12) we can calculate the expected values shown in Appendix A.1 that enable us to express the complete data log-likelihood Equation (3.9) as

$$\begin{aligned}\mathcal{Q}_{2C}(\boldsymbol{\vartheta}_2) =& C + \sum_{g=1}^{G} -\frac{n_g}{2}\log|\boldsymbol{\Psi}_g| - \frac{n_g}{2}\mathrm{tr}(\boldsymbol{\Psi}_g^{-1}\mathbf{S}_g^{(k+1)}) \\ &+ n_g\mathrm{tr}\left(\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}_g\boldsymbol{\beta}_g^{(k)}\mathbf{S}_g^{(k+1)}\right) - \frac{n_g}{2}\mathrm{tr}\left(\boldsymbol{\Lambda}_g'\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}_g\boldsymbol{\Theta}_g^{(k+1/2)}\right)\end{aligned} \tag{3.13}$$

where the matrices $\boldsymbol{\beta}_g^{(k)}$ and $\boldsymbol{\Theta}_g^{(k+1/2)}$ are given by

$$\boldsymbol{\beta}_g^{(k)} = \boldsymbol{\Lambda}_g^{(k)'}(\boldsymbol{\Lambda}_g^{(k)'}\boldsymbol{\Lambda}_g^{(k)'} + \omega_g^{(k)}\boldsymbol{\Delta}_g^{(k)})^{-1} \tag{3.14}$$

$$\boldsymbol{\Theta}_g^{(k+1/2)} = \mathbf{I}_q - \boldsymbol{\beta}_g^{(k)}\boldsymbol{\Lambda}_g^{(k)} + \boldsymbol{\beta}_g^{(k)}\mathbf{S}_g^{(k+1)}\boldsymbol{\beta}_g^{(k)'} \tag{3.15}$$

with the appropriate constraints for for each model applied to the parameters $\boldsymbol{\Lambda}_g, \omega_g$, and $\boldsymbol{\Delta}_g$.

3.3.5 Alternation 2: CM-step

In the CM-step of the second alternation on iteration $(k+1)$ we maximize $Q_{2C}(\boldsymbol{\vartheta}_2)$ with respect to $\boldsymbol{\vartheta}_2$ with $\boldsymbol{\vartheta}_1 = \boldsymbol{\vartheta}_1^{(k+1)}$. It is important to note that for models belonging to the original PGMM family, we obtain estimates of the parameter $\boldsymbol{\Psi}$ and derive ω and $\boldsymbol{\Delta}$ from this estimate. The updates for the parameter set $\boldsymbol{\vartheta}_2^{(k+1)}$ for each set of constraint conditions in Table 2.1 are as follows:

- First, let us define the terms

$$\widetilde{\mathbf{S}}^{(k+1)} = \sum_{g=1}^{g} \pi_g^{(k+1)} \mathbf{S}_g^{(k+1)}$$
$$\mathbf{\Theta}^{(k+1/2)} = \mathbf{I}_q - \boldsymbol{\beta}^{(k)}\mathbf{\Lambda}^{(k)} + \boldsymbol{\beta}^{(k)}\mathbf{S}^{(k+1)}\boldsymbol{\beta}^{(k)\prime}$$

- Model CCCC: $\mathbf{\Sigma}_g = \mathbf{\Lambda}\mathbf{\Lambda}' + \omega\mathbf{I}_p$

$$\boldsymbol{\beta}^{(k)} = \mathbf{\Lambda}^{(k)\prime}(\mathbf{\Lambda}^{(k)}\mathbf{\Lambda}^{(k)\prime} + \omega^{(k)}\mathbf{I}_p)^{-1}$$
$$\mathbf{\Lambda}^{(k+1)} = \widetilde{\mathbf{S}}^{(k+1)}\boldsymbol{\beta}^{(k)}(\mathbf{\Theta}^{(k+1/2)})^{-1}$$
$$\omega^{(k+1)} = \frac{1}{p}\mathrm{tr}\left\{\widetilde{\mathbf{S}}^{(k+1)} + \mathbf{\Lambda}^{(k+1)}\boldsymbol{\beta}^{(k)}\widetilde{\mathbf{S}}^{(k+1)}\right\}$$

- Model CCUC: $\mathbf{\Sigma}_g = \mathbf{\Lambda}\mathbf{\Lambda}' + \omega_g\mathbf{I}_p$

$$\boldsymbol{\beta}_g^{(k)} = \mathbf{\Lambda}^{(k)\prime}(\mathbf{\Lambda}^{(k)}\mathbf{\Lambda}^{(k)\prime} + \omega_g^{(k)}\mathbf{I}_p)^{-1}$$
$$\mathbf{\Lambda}^{(k+1)} = \left[\sum_{g=1}^{G} \frac{n_g^{(k)}}{\omega_g^{(k)}}\mathbf{S}_g^{(k+1)}\boldsymbol{\beta}_g^{(k)}\right]\left[\sum_{g=1}^{G} \frac{n_g^{(k)}}{\omega_g^{(k)}}\mathbf{\Theta}_g^{(k+1/2)}\right]^{-1}$$
$$\omega_g^{(k+1)} = \frac{1}{p}\mathrm{tr}\left\{\mathbf{S}_g^{(k+1)} - 2\mathbf{\Lambda}^{(k+1)}\boldsymbol{\beta}_g^{(k)}\mathbf{S}_g^{(k+1)} + \mathbf{\Lambda}^{(k+1)}\mathbf{\Theta}_g^{(k+1/2)}\mathbf{\Lambda}^{(k+1)\prime}\right\}$$

- Model CCCU: $\mathbf{\Sigma}_g = \mathbf{\Lambda}\mathbf{\Lambda}' + \omega\mathbf{\Delta}$

$$\boldsymbol{\beta}^{(k)} = \mathbf{\Lambda}^{(k)\prime}(\mathbf{\Lambda}^{(k)}\mathbf{\Lambda}^{(k)\prime} + (\omega\mathbf{\Delta})^{(k)})^{-1}$$
$$\mathbf{\Lambda}^{(k+1)} = \widetilde{\mathbf{S}}^{(k+1)}\boldsymbol{\beta}^{(k)}(\mathbf{\Theta}^{(k+1/2)})^{-1}$$
$$(\omega\mathbf{\Delta})^{(k+1)} = \mathbf{\Psi}^{(k+1)} = \mathrm{diag}\left\{\widetilde{\mathbf{S}}^{(k+1)} + \mathbf{\Lambda}^{(k+1)}\boldsymbol{\beta}^{(k)}\widetilde{\mathbf{S}}^{(k+1)}\right\}$$

- Model CCUU: $\quad \boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \omega_g \boldsymbol{\Delta}$

$$\boldsymbol{\beta}_g^{(k)} = \boldsymbol{\Lambda}^{(k)'}(\boldsymbol{\Lambda}^{(k)}\boldsymbol{\Lambda}^{(k)'} + \omega_g^{(k)}\boldsymbol{\Delta}^{(k)})^{-1}$$

$$\boldsymbol{\Lambda}^{(k+1)} = \left[\sum_{g=1}^{G} \frac{n_g^{(k)}}{\omega_g^{(k)}} \mathbf{S}_g^{(k+1)} \boldsymbol{\beta}_g^{(k)}\right] \left[\sum_{g=1}^{G} \frac{n_g^{(k)}}{\omega_g^{(k)}} \boldsymbol{\Theta}_g^{(k+1/2)}\right]^{-1}$$

$$\omega_g^{(k+1)} = \frac{1}{p}\mathrm{tr}\left\{(\boldsymbol{\Delta}^{(k)})^{-1}\left[\mathbf{S}_g^{(k+1)} - 2\boldsymbol{\Lambda}^{(k+1)}\boldsymbol{\beta}_g^{(k)}\mathbf{S}_g^{(k+1)} + \boldsymbol{\Lambda}^{(k+1)}\boldsymbol{\Theta}_g^{(k+1/2)}\boldsymbol{\Lambda}^{(k+1)}\right]\right\}$$

$$\boldsymbol{\Delta}^{(k+1)} = \frac{1}{\kappa}\mathrm{diag}\left\{\boldsymbol{\Xi}^{(k+1/2)}\right\},$$

where the matrix $\boldsymbol{\Xi}^{(k+1/2)}$ and the coefficient $n + 2\kappa$ are given by

$$\boldsymbol{\Xi}^{(k+1/2)} = \sum_{g=1}^{G} \frac{n_g}{\omega_g^{(k+1)}}\left[\mathbf{S}_g^{(k+1)} - 2\boldsymbol{\Lambda}^{(k+1)}\boldsymbol{\beta}_g^{(k)}\mathbf{S}_g^{(k+1)} + \boldsymbol{\Lambda}^{(k+1)}\boldsymbol{\Theta}_g^{(k+1/2)}\boldsymbol{\Lambda}^{(k+1)'}\right]$$

$$\kappa = \left(\prod_{j=1}^{p} \boldsymbol{\Xi}_{jj}^{(k+1/2)}\right)^{1/p}$$

- Model CUCU: $\quad \boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \omega \boldsymbol{\Delta}_g$

$$\boldsymbol{\beta}_g^{(k)} = \boldsymbol{\Lambda}^{(k)'}(\boldsymbol{\Lambda}^{(k)}\boldsymbol{\Lambda}^{(k)'} + \omega^{(k)}\boldsymbol{\Delta}_g^{(k)})^{-1}$$

$$\lambda_j^{(k+1)} = \mathbf{r}_j^{(k+1/2)}\left[\sum_{g=1}^{G} \frac{n_g^{(k)}}{\boldsymbol{\Delta}_{g(jj)}^{(k)}} \boldsymbol{\Theta}_g^{(k+1/2)}\right]^{-1}$$

$$\omega^{(k+1)} = \frac{1}{p}\sum_{g=1}^{G} \pi_g^{(k+1)}\mathrm{tr}\left\{(\boldsymbol{\Delta}_g^{(k)})^{-1}\boldsymbol{\Xi}_g^{(k+1/2)}\right\}$$

$$\boldsymbol{\Delta}^{(k+1)} = \frac{1}{\omega^{(k+1)}\kappa_g}\mathrm{diag}\left\{\boldsymbol{\Xi}^{(k+1/2)}\right\},$$

where $\mathbf{r}_j^{(k+1/2)}, \boldsymbol{\Xi}_g^{(k+1/2)}$ and the coefficient κ_g are given by

$$\mathbf{r}_j^{(k+1/2)} = \left[\sum_{g=1}^{G} \frac{n_g^{(k)}}{\boldsymbol{\Delta}_{g(jj)}^{(k)}} \mathbf{S}_g^{(k+1)}\boldsymbol{\beta}_g^{(k)'}\right]_j$$

$$\boldsymbol{\Xi}_g^{(k+1/2)} = \left[\mathbf{S}_g^{(k+1)} - 2\boldsymbol{\Lambda}^{(k+1)}\boldsymbol{\beta}_g^{(k)}\mathbf{S}_g^{(k+1)} + \boldsymbol{\Lambda}^{(k+1)}\boldsymbol{\Theta}_g^{(k+1/2)}\boldsymbol{\Lambda}^{(k+1)'}\right]$$

$$\kappa_g = \left(\prod_{j=1}^{p} \boldsymbol{\Xi}_{g(jj)}^{(k+1/2)}\right)^{1/p}$$

- Model CUUU: $\mathbf{\Sigma}_g = \mathbf{\Lambda}\mathbf{\Lambda}' + \omega_g\mathbf{\Delta}_g$

$$\boldsymbol{\beta}^{(k)} = \mathbf{\Lambda}^{(k)'}(\mathbf{\Lambda}^{(k)}\mathbf{\Lambda}^{(k)'} + (\omega_g\mathbf{\Delta}_g)^{(k)})^{-1}$$
$$\mathbf{R}^{(k+1/2)} = \sum_{g}^{G} n_g^{(k)}(\mathbf{\Psi}_g^{(k)})^{-1}\mathbf{S}_g^{(k+1)}\boldsymbol{\beta}_g^{(k)'}$$
$$\lambda_j^{(k+1)} = \mathbf{r}_j^{(k+1/2)}\left[\sum_{g=1}^{G}\frac{n_g^{(k)}}{\omega_g^{(k)}}\mathbf{\Theta}_g^{(k+1/2)}\right]^{-1}$$
$$(\omega_g\mathbf{\Delta}_g)^{(k+1)} = \text{diag}\left\{\mathbf{S}^{(k+1)} - 2\mathbf{\Lambda}^{(k+1)}\boldsymbol{\beta}_g^{(k)}\mathbf{S}_g^{(k+1)} + \mathbf{\Lambda}^{(k+1)}\mathbf{\Theta}_g^{(k+1/2)}\mathbf{\Lambda}^{(k+1)}\right\},$$

where $\lambda_j^{(k+1)}$ is the jth row of $\mathbf{\Lambda}^{(k+1)}$ and $\mathbf{r}_j^{(k+1/2)}$ is the jth row of $\mathbf{R}^{(k+/2)}$.

- Model UCCC: $\mathbf{\Sigma}_g = \mathbf{\Lambda}_g\mathbf{\Lambda}_g' + \omega\mathbf{I}_p$

$$\boldsymbol{\beta}_g^{(k)} = \mathbf{\Lambda}_g^{(k)'}(\mathbf{\Lambda}_g^{(k)}\mathbf{\Lambda}_g^{(k)'} + \omega^{(k)}\mathbf{I}_p)^{-1}$$
$$\mathbf{\Lambda}_g^{(k+1)} = \mathbf{S}_g^{(k+1)}\boldsymbol{\beta}_g^{(k)}(\mathbf{\Theta}_g^{(k+1/2)})^{-1}$$
$$\omega^{(k+1)} = \frac{1}{p}\sum_{g=1}^{G}\pi_g^{(k+1)}\text{tr}\left\{\mathbf{S}_g^{(k+1)} + \mathbf{\Lambda}_g^{(k+1)}\boldsymbol{\beta}_g^{(k)}\mathbf{S}_g^{(k+1)}\right\}$$

- Model UCUC: $\mathbf{\Sigma}_g = \mathbf{\Lambda}_g\mathbf{\Lambda}_g' + \omega_g\mathbf{I}_p$

$$\boldsymbol{\beta}_g^{(k)} = \mathbf{\Lambda}_g^{(k)'}(\mathbf{\Lambda}_g^{(k)}\mathbf{\Lambda}_g^{(k)'} + \omega_g^{(k)}\mathbf{I}_p)^{-1}$$
$$\mathbf{\Lambda}_g^{(k+1)} = \mathbf{S}_g^{(k+1)}\boldsymbol{\beta}_g^{(k)}(\mathbf{\Theta}_g^{(k+1/2)})^{-1}$$
$$\omega_g^{(k+1)} = \frac{1}{p}\text{tr}\left\{\mathbf{S}_g^{(k+1)} + \mathbf{\Lambda}_g^{(k+1)}\boldsymbol{\beta}_g^{(k)}\mathbf{S}_g^{(k+1)}\right\}$$

- Model UCCU: $\mathbf{\Sigma}_g = \mathbf{\Lambda}_g\mathbf{\Lambda}_g' + \omega\mathbf{\Delta}$

$$\boldsymbol{\beta}_g^{(k)} = \mathbf{\Lambda}_g^{(k)'}(\mathbf{\Lambda}_g^{(k)}\mathbf{\Lambda}_g^{(k)'} + (\omega\mathbf{\Delta})^{(k)})^{-1}$$
$$\mathbf{\Lambda}_g^{(k+1)} = \mathbf{S}_g^{(k+1)}\boldsymbol{\beta}_g^{(k)}(\mathbf{\Theta}_g^{(k+1/2)})^{-1}$$
$$(\omega\mathbf{\Delta})^{(k+1)} = \mathbf{\Psi}^{(k+1)} = \sum_{g=1}^{G}\pi_g^{(k+1)}\text{diag}\left\{\mathbf{S}^{(k+1)} + \mathbf{\Lambda}^{(k+1)}\boldsymbol{\beta}^{(k)}\mathbf{S}^{(k+1)}\right\}$$

- Model UCUU: $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}'_g + \omega_g\boldsymbol{\Delta}$

$$\boldsymbol{\beta}_g^{(k)} = \boldsymbol{\Lambda}_g^{(k)\prime}(\boldsymbol{\Lambda}_g^{(k)}\boldsymbol{\Lambda}_g^{(k)\prime} + \omega_g^{(k)}\boldsymbol{\Delta}^{(k)})^{-1}$$
$$\boldsymbol{\Lambda}_g^{(k+1)} = \mathbf{S}_g^{(k+1)}\boldsymbol{\beta}_g^{(k)}(\boldsymbol{\Theta}_g^{(k+1/2)})^{-1}$$
$$\omega_g^{(k+1)} = \frac{1}{p}\mathrm{tr}\left\{(\boldsymbol{\Delta}_g^{(k)})^{-1}(\mathbf{S}^{(k+1)} + \boldsymbol{\Lambda}^{(k+1)}\boldsymbol{\beta}^{(k)}\mathbf{S}^{(k+1)})\right\}$$
$$\boldsymbol{\Delta}^{(k+1)} = \frac{1}{\kappa}\mathrm{diag}\left\{\boldsymbol{\Xi}^{(k+1/2)}\right\}$$

where the matrix $\boldsymbol{\Xi}^{(k+1/2)}$ and the coefficient κ_g are given by,

$$\boldsymbol{\Xi}^{(k+1/2)} = \sum_{g=1}^{G}\frac{n_g^{(k)}}{\omega_g^{(k+1)}}\left[\mathbf{S}_g^{(k+1)} + \boldsymbol{\Lambda}_g^{(k+1)}\boldsymbol{\beta}_g^{(k)}\mathbf{S}_g^{(k+1)}\right]$$
$$\kappa = \left(\prod_{j=1}^{p}\boldsymbol{\Xi}_{jj}^{(k+1/2)}\right)^{1/p}$$

- Model UUCU: $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}'_g + \omega\boldsymbol{\Delta}_g$

$$\boldsymbol{\beta}_g^{(k)} = \boldsymbol{\Lambda}_g^{(k)\prime}(\boldsymbol{\Lambda}_g^{(k)}\boldsymbol{\Lambda}_g^{(k)\prime} + \omega^{(k)}\boldsymbol{\Delta}_g^{(k)})^{-1}$$
$$\boldsymbol{\Lambda}_g^{(k+1)} = \mathbf{S}_g^{(k+1)}\boldsymbol{\beta}_g^{(k)}(\boldsymbol{\Theta}_g^{(k+1/2)})^{-1}$$
$$\omega_g^{(k+1)} = \frac{1}{p}\sum_{g=1}^{G}\pi_g^{(k+1)}\mathrm{tr}\left\{(\boldsymbol{\Delta}_g^{(k)})^{-1}(\mathbf{S}^{(k+1)} + \boldsymbol{\Lambda}^{(k+1)}\boldsymbol{\beta}^{(k)}\mathbf{S}^{(k+1)})\right\}$$
$$\boldsymbol{\Delta}_g^{(k+1)} = \frac{1}{\kappa_g}\mathrm{diag}\left\{\mathbf{S}_g^{(k+1)} + \boldsymbol{\Lambda}_g^{(k+1)}\boldsymbol{\beta}_g^{(k)}\mathbf{S}_g^{(k+1)}\right\},$$

where the coefficient κ_g is given by

$$\kappa_g = \left(\prod_{j=1}^{p}\left[\mathbf{S}_g^{(k+1)} + \boldsymbol{\Lambda}_g^{(k+1)}\boldsymbol{\beta}_g^{(k)}\mathbf{S}_g^{(k+1)}\right]_{jj}\right)^{1/p}$$

- Model UUUU: $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}'_g + \omega_g\boldsymbol{\Delta}_g$

$$\boldsymbol{\beta}_g^{(k)} = \boldsymbol{\Lambda}_g^{(k)\prime}(\boldsymbol{\Lambda}_g^{(k)}\boldsymbol{\Lambda}_g^{(k)\prime} + (\omega_g\boldsymbol{\Delta}_g)^{(k)})^{-1}$$
$$\boldsymbol{\Lambda}_g^{(k+1)} = \mathbf{S}_g^{(k+1)}\boldsymbol{\beta}_g^{(k)}(\boldsymbol{\Theta}_g^{(k+1/2)})^{-1}$$
$$(\omega_g\boldsymbol{\Delta}_g)^{(k+1)} = \boldsymbol{\Psi}_g^{(k+1)} = \mathrm{diag}\left\{\mathbf{S}_g^{(k+1)} + \boldsymbol{\Lambda}_g^{(k+1)}\boldsymbol{\beta}_g^{(k)}\mathbf{S}_g^{(k+1)}\right\}$$

3.4 Computational Details

3.4.1 Initialization

The parameters in the set $\boldsymbol{\vartheta}^{(0)} = \{\boldsymbol{\pi}_g^{(0)}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\alpha}^{(0)}, \boldsymbol{\Sigma}^{(0)}\}$ are initialized via an MSAL model. The MSAL models ability to account for asymmetry provides an initial estimate of the skewness parameter as well as a more accurate initial estimate for our location parameter. The preliminary parameter estimates required for the MSAL model are implemented using the z_{ig} estimates obtained from the `kmeans(...)` function of the `stats` package in R. The contamination parameters are initialized with the fixed values $\eta_g^{(0)} = 1.001$ and $\rho_g^{(0)} = 0.999$ for all $g = 1, \ldots, G$. These values are not set equal to 1 to avoid singularities within the first iteration (Punzo et al. 2020). The initial parameter estimates are then used to obtain initial estimates of all expected values calculated in Section 3.3.1.

Initialization of the parameter set $\boldsymbol{\vartheta}_2^{(1/2)} = \{\boldsymbol{\Lambda}^{(1/2)}, \boldsymbol{\omega}^{(1/2)}, \boldsymbol{\Delta}^{(1/2)}\}$ is conducted subsequent to the E-step of the second alternation in the first iteration. The method of initialization follows from McNicholas and Murphy (2008). The eigen-decomposition of the matrix $\mathbf{S}_g$ is computed using the R function `eigen(...)`. The initial values for the elements in $\boldsymbol{\Lambda}_g$ are then calculated as

$$\lambda_{ij} = \sqrt{d_j}\gamma_{ij},$$

where d_j is the jth largest eigenvalue of $\mathbf{S}_g$ and γ_{ij} is the ith element of the eigenvector corresponding to d_j, where $i = 1, \ldots, p$ and $j = 1, \ldots, q$. The parameters $\{\boldsymbol{\omega}, \boldsymbol{\Delta}\}$ are then initialized as

$$\boldsymbol{\Psi}_g = \text{diag}\{\mathbf{S}_g - \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g'\},$$

$$\omega_g = |\boldsymbol{\Psi}_g|^{1/p}, \qquad \boldsymbol{\Delta}_g = \frac{1}{\omega_g}\boldsymbol{\Psi}_g.$$

3.4.2 Dealing with Infinite Log-Likelihood Values

As documented in Franczak et al. (2014) complications can arise when estimating the location parameter $\boldsymbol{\mu}_g$ for mixtures of SAL distributions. Computational singularities occur when the the parameter $\boldsymbol{\mu}_g$ is equal to some observation vector $\mathbf{x}_i$ in the data. In our PCSALM family, such singularities manifest when calculating the expected values E_{2ig} and $\widetilde{E}_{2ig}$ as the Mahalanobis distance $\delta(\mathbf{x}_i, \boldsymbol{\mu}_g | \boldsymbol{\Sigma}_g)$ will take on a value of zero. To remedy this issue, we stop updating the parameter $\boldsymbol{\mu}_g$ if the euclidean distance between $\boldsymbol{\mu}_g$ and $\mathbf{x}_i$ for $i = 1, \ldots, n$ is less than 10^{-10}. In such an event the estimate for the parameter $\boldsymbol{\alpha}_g$ is also modified to be calculated as

$$\boldsymbol{\alpha}_g^{(k+1)} = \frac{\sum_{i=1}^{n} z_{ig}^{(k)} \left(v_{ig}^{(k)} + \frac{1-v_{ig}^{(k)}}{\sqrt{\eta_g^{(k)}}} \right) (\mathbf{x}_i - \boldsymbol{\mu}_g^{(k)})}{\sum_{i=1}^{n} z_{ig}^{(k)} \left(v_{ig}^{(k)} E_{1ig}^{(k)} + \frac{1-v_{ig}^{(k)}}{\eta_g^{(k)}} \widetilde{E}_{1ig}^{(k)} \right)}$$

where all terms are as given in Section 3.3.2. This solution has been shown to be effective in Franczak et al. (2014) and Morris et al. (2019).

3.4.3 Minimum Component Size

In some cases, the expected component membership labels assigned less than three observations to a mixture component. This creates an issues and when computing the matrices $\mathbf{S}_g$ because the positive definite property does not hold. Therefore, the calculations shown in Section 3.3.5 were compromised. This issue was addressed by restricting the component size to $n_g > 5$ for $g = 1, \ldots, G$ while the AECM algorithm was iterating. From an interpretive perspective, 1 or 2 observations would not comprise a probability distribution, and thus would not serve as a meaningful component.

Chapter 4

Simulated Data Analysis

In this section, four types of simulated data sets were considered:

1. SAL clusters with noise;
2. SAL clusters;
3. Gaussian clusters with noise;
4. Gaussian clusters.

4.1 Simulation 1

In the first simulation, thirty data sets with $n = 1000$ observations and $p = 10$ dimensions were generated to feature $G = 2$ asymmetric components where the size of the components were given by $n_1 = 600$ and $n_2 = 400$. The components in each data set were generated using the R function `rsal(...)` from the package `MixSAL` (Franczak et al. 2018). The parameters used as inputs in the `rsal(...)` function were randomly generated in the following manner:

- Entries in $\boldsymbol{\mu}_1$ were generated by a uniform distribution on $(5, 15)$,
- Entries in $\boldsymbol{\mu}_2$ were generated by a uniform distribution on $(10, 20)$,
- Entries in $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ were generated by a uniform distribution on $(-2, 2)$,

- $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ were generated using the R function `genPositiveDefMat(...)` from the package `clusterGeneration`. (Qui and Joe 2020).

The set of parameters generated was unique to each data set. Contamination was implemented by including an additional 100 observations vectors where each entry was randomly generated from a uniform distribution on $(-10, 50)$, increasing the total number of observations to $n = 1100$. These additional noise observations were unique to each data set. Each scale matrix structure in the PCSALM family was fitted to all data sets with $G = 1, \dots, 4$ components and $q = 1, \dots, 4$ factors. This resulted in a total of 192 models being fitted to each data set.

4.1.1 Results for SAL clusters with noise

Table B.1 provides the complete ARI, sensitivity, specificity, ICL and BIC values for models selected by the ICL from all possible models in the PCSALM family. The classification measures are summarized in Table 4.1. The ARI metric was ineffective at reflecting the capability of a model to correctly identify the additional noise observations as contamination. For this reason, we evaluated the ARI using only the known good points and the sensitivity and specificity were considered. The sensitivity is the proportion of noise observations that were correctly identified as bad points by the model and the specificity is the proportion of good points that were identified as such. The selected models generally provided a strong classification performance with 28 of the selected models providing an ARI and specificity value greater than 0.90. These models were also generally effective at detecting bad points with 22 of the models correctly identifying at least 95% of the noise observations. Despite this strong performance, several of the models suggested by the ICL were unable to accurately identify any bad points and instead opted to consider the noise observations as components.

In general, the preferred models produced by the AECM algorithm tended to be fit to 4 components while keeping observation counts in two of these components

Table 4.1: Summary of classification performance for models selected by ICL for SAL clusters with noise.

Mean ARI	0.95
Mean Sensitivity	0.78
Mean Specificity	0.97
# of Models w/ ARI $\geq$ 0.90	28
# of Models w/ Sensitivity $\geq$ 0.95	22

incredibly low, often at 1 or 0 observations. While technically considered a component by the algorithm, such low observation counts make it easy to justify labelling these points as contamination. An example of such a solution can be seen in the solution proposed by the ICL for the first data set, shown in Table 4.2.

Table 4.2: Contingency table of the suggested clustering solution for data set 1 featuring SAL clusters with noise (0 identifies bad points).

	True Labels		
	0	1	2
0	99	0	1
1	1	0	0
2	0	0	399
3	0	0	0
4	0	600	0

The model selected by the ICL for the 29th data set provided the worst performance with an ARI of 0.29. The proposed solution is provided in Table 4.3a. When evaluating classification performance we can see that the model is severely over-estimating the proportion of contamination in the data, featuring a specificity value of 0.540. A similar issue was present in the proposed solution for the 18th data set, although not as severe. Referring back to the formulation of the ICL in Section 2.5.2, the measure is penalized for the estimated entropy in component membership labels z_{ig}. We applied a similar formulation of entropy to the expected values v_{ig}, to evaluate when there was uncertainty in the bad points identified. The solutions proposed for data sets 18 and 29 featured the highest levels of uncertainty with penalties of -5115.89 and -6161.00 respectively. Comparatively the mean penalty across all 30 data sets was only -1494.07. The values quantifying entropy in v_{ig} corroborate our observation of an over-estimation of the proportion of contamination in the data and could potentially serve as an indication for when this kind of misclassification is present.

Table 4.3: Contingency tables of the clustering solutions for data set 29 featuring SAL clusters with noise (0 identifies bad points).

(a) Model selected by ICL

	True Labels		
	0	1	2
0	98	292	168
1	0	0	232
2	1	0	0
3	1	0	0
4	0	308	0

(b) Model selected by modified ICL

	True Labels		
	0	1	2
0	99	0	2
1	0	0	398
2	1	0	0
3	0	0	0
4	0	600	0

When the selection procedure for data set 29 used a modified ICL that incorporates estimated entropy of the expected values v_{ig} via addition to penalize we obtained the solution given in Table 4.3b. The classification performance has clearly improved, in particular the specificity of the new solution has improved to 0.998. However, when the modified ICL was used as the selection metric for all data sets the mean sensitivity dropped to 0.422. For this reason, we would only suggest using the modified ICL when the entropy term for v_{ig} is sufficiently large and the model selected by the ICL suggests an excessive proportion of bad points. Data set 29 provides ample justification since the initial solution suggests more than 50% of the data is bad points.

Despite poor sensitivity in some of the selected models and the proposed solutions for data sets 18 and 29, these issues are a consequence of ICL's inconsistency in selecting the optimal model and not the PCSALM family's inability to formulate an accurate solution. Table B.2 provides a count of all models for each simulation that were able to satisfy strict classification performance thresholds. Most notably, the PCSALM family was able to produce multiple solutions with an ARI > 0.95 and perfect identification of bad points for all 30 data sets in the simulation. The lowest number of near-perfect solutions occurred in the 28th data set, with 5 models providing such a clustering solution.

Table 4.4: Summary statistics for the classification performance of each scale matrix structure when selected with ICL for SAL clusters with noise.

Model	Mean ARI	S.D. of ARI	Mean Sensitivity	S.D. of Sensitivity
CCCC	0.95	0.15	0.99	0.01
CCCU	0.95	0.16	0.99	0.01
CCUC	0.99	0.01	0.26	0.41
CCUU	0.99	0.01	0.28	0.40
CUCU	0.99	0.01	0.83	0.27
CUUU	0.99	0.01	0.28	0.40
UCCC	0.98	0.01	0.86	0.19
UCUC	0.99	0.01	0.33	0.44
UCCU	0.99	0.01	0.89	0.17
UCUU	0.99	0.01	0.34	0.45
UUCU	0.99	0.01	0.84	0.29
UUUU	0.99	0.01	0.31	0.44

Table 4.4 provides summary statistics of the classification performance of models selected by the ICL for each of the 30 data sets when only comparing models with the same scale matrix structure. The structures "CCCC" and "CCCU" were very capable of identifying bad points, with a mean sensitivity of 0.99 and extremely low variance. However, despite having a high mean ARI of 0.94 and 0.95 respectively the variability in the ARI was much higher than in other models making these structures less reliable for accurate classification performance. The opposite was true for the remaining ten scale matrix structures. All provided highly accurate ARI values with

little variability but were not as capable of identifying contamination. The "UCCU" structure was the most capable at detecting bad points with a mean sensitivity of 0.89.

Table 4.5: Summary statistics for the classification performance of each scale matrix structure when selected with a modified ICL for SAL clusters with noise.

Model	Mean ARI	S.D. of ARI	Mean Sensitivity	S.D. of Sensitivity
CCCC	0.98	0.02	0.99	0.01
CCCU	0.98	0.04	0.99	0.01
CCUC	0.99	0.01	0.23	0.39
CCUU	0.99	0.01	0.25	0.38
CUCU	0.99	0.01	0.81	0.26
CUUU	0.99	0.01	0.28	0.40
UCCC	0.99	0.01	0.84	0.18
UCUC	0.99	0.01	0.23	0.38
UCCU	0.99	0.01	0.88	0.17
UCUU	0.99	0.01	0.24	0.39
UUCU	0.99	0.01	0.83	0.30
UUUU	0.99	0.01	0.21	0.38

When the estimated entropy of the expected values v_{ig} were incorporated into the model selection procedure via addition to the ICL, we obtained the results in Table 4.5. These results are mostly comparable to those selected by the unmodified ICL in Table 4.4, however the ARI values for the scale matrix structures "CCCC" and "CCCU" have a slightly improved mean and a sizeable decrease in the standard deviation. Since their ability to accurately identify bad points remains intact these scale matrix structures in conjunction with this model selection method yielded very strong classification performance for all suggested models across each of the 30 data sets.

4.2 Simulation 2

Thirty data sets were generated in a manner analogous to the generation procedure described in Section 4.1, but were distinguished by the exclusion of any additional noise observations in the data. Similarly, all thirty data sets were fitted to each scale matrix structure in the PCSALM family with $G = 1, \ldots, 4$ components and $q = 1, \ldots, 4$ factors, resulting in a total of 192 models being fit to each data set.

4.2.1 Results for SAL clusters

Models from the PCSALM family were first selected using ICL for each of the 30 data sets. The summary statistics for the ARI and the number of observations assigned to be outliers of these models is provided in Table 4.6. The mean ARI value of 0.78 demonstrates an adequate performance, but the high amounts of variability present compromised the reliability of classification solutions proposed. Additionally, the number of bad points incorrectly identified by the models was much higher than desired and also appeared to be quite volatile.

Table 4.6: Summary statistics of classification performance for models selected by ICL for SAL clusters.

	Mean	Std. Dev.	Minimum	Maximum
ARI	0.78	0.33	-0.00	1.00
Bad points identified	164.27	335.07	0.00	1000.00

Upon further investigation we were able to discover that the high number of observations misclassified as bad points and the large amounts of variation in these metrics was due to six models that grossly over estimated the overall proportion of bad points in their respective data sets. These models were selected for data sets 2, 4, 14, 19, 20 and 23 in the simulation, and were the only models that suggested a specificity below 95%. The number of observations misclassified as bad points by these six models is provided in Table 4.7.

Table 4.7: Bad points incorrectly identified in SAL clusters when specificity does not exceed 0.95.

Data Set	2	4	14	19	20	23
Bad points identified	997	1000	943	650	731	471

These six data sets were the isolated from the rest of the data sets in the simulation and the summary statistics of the ARI were re-analyzed. Table 4.8 contains the updated summary statistics as well as the mean entropy in the expected values v_{ig}. As expected, the models selected for the six isolated data sets provided poor classification performance and feature a high degree of uncertainty in their identification of the bad points. The remaining 24 models featured a much stronger classification performance than our initial analysis with greatly reduced variability.

Table 4.8: Comparison of the certainty in bad points identified and its effect on the classification performance of models selected by ICL for SAL clusters.

	Specificity $< 95\%$	Specificity $\geq 95\%$
Mean v_{ig} Entropy	3085.94	7.80
Mean ARI	0.14	0.94
Std. Dev. ARI	0.16	0.04

Since all of the models that provided poor classification performance for their respective data set also featured a high degree of uncertainty in their identification of the bad points, the model selection process was redone with the aforementioned modification to the ICL to penalize entropy in the expected values v_{ig}. Table 4.9 provides summary statistics of the classification performance for the models selected. The metrics provided here demonstrate a stronger and much more stable classification performance that is similar to the performance of the 24 models in our initial analysis that did not severely over-estimate contamination. Our modified ICL selected a different model in 9 of the 30 data sets simulated, including all six to feature the low specificity that we previously high-lighted.

Table 4.9: Summary statistics of classification performance for models selected by the modified ICL for SAL clusters.

	Mean	Std. Dev.	Minimum	Maximum
ARI	0.94	0.03	0.84	1.00
Bad points identified	3.60	8.49	0.00	34.00

4.3 Simulation 3

The third simulation generated ten data sets featuring Gaussian components and contamination using the method of generation described in Punzo et al. (2020). These data sets were generated with $n = 200$ observations and $p = 10$ dimensions and featured $G = 2$ components and a latent factor structure with $q = 3$ latent factors. The set of parameters were generated in the following manner:

- $\boldsymbol{\mu}_1$ was set at the origin,
- $\boldsymbol{\mu}_2$ was generated from $\mathcal{N}_{10}(\ \mathbf{0}, \mathbf{I}_{10})$,
- Entries in the loading matrices $\boldsymbol{\Lambda}_1,\ \boldsymbol{\Lambda}_2$ were drawn from independent $\mathcal{N}(0,\ 1)$,
- Diagonal entries of the matrcies $\boldsymbol{\Psi}_1,\ \boldsymbol{\Psi}_2$ were generated by a uniform distribution on $(0.5, 1)$,
- π_1 was calculated by generating $n = 200$ values from the uniform distribution on $(0, 1)$, then evaluating the proportion of observations with a value greater than 0.5. π_2 was then given by $1 - \pi_1$.

This set of parameters was then used as inputs for the `rmvnorm(...)` function in R to generate the Gaussian components. Contamination was implemented by including an additional 20 observation vectors where each entry was randomly generated from a uniform distribution on $(-5, 5)$, increasing the total number of observations to $n = 220$. Each scale matrix structure in the PCSALM family was fitted to all data sets with $G = 1, \ldots, 5$ components and $q = 1, \ldots, 9$ latent factors, resulting in a total of 540 models being fit to each data set. $q = 10$ latent factors was not considered as it was in Punzo et al. (2020) since a value $q = p$ led to singularities in our AECM algorithm. The results of the analysis were compared to the results given in Punzo et al. (2020). This comparison was possible since our data was generated to match the data used in their analysis.

4.3.1 Results for Gaussian clusters with noise

A comparison of the classification performance given by the PCSALM family and other mixture of factor analyzers is provided in Section 4.3. The models proposed by the PCSALM family provided a slightly weaker ARI than the MCGFA models, but still had a comparable classification performance for points within the Gaussian clusters to the other mixtures of factor analyzers. The model was not effective at identifying bad points with a mean sensitivity of only 0.075. This is likely because the noise observations included in these data sets were less dispersed than in the previous simulation, and were captured by the skewness parameter of the PCSALM instead. Additionally, the PCSALM model was not effective at recovering the number of latent factors and suggested solutions with $q = 7$.

Table 4.10: Classification performance of mixtures of factor analyzers models on Gaussian clusters with noise.

	PCSALM	MCGFA	MMtFA	EPGMM
G	2	2	2	2
Mean ARI	0.920	**0.936**	0.926	0.902
ARI Std. Dev.	**0.03**	0.06	0.05	0.05
Mean Sensitivity	0.075	**0.965**	N/A	N/A

4.4 Simulation 4

Ten data sets were generated featuring Gaussian components and similarly used the method of generation described in Punzo et al. (2020). These data sets differed from those in Section 4.3 since the data did not include any additional noise observations. These data sets were also fitted to each scale matrix structure in the PCSALM family with $G = 1, \dots, 5$ components and $q = 1, \dots, 9$ latent factors, resulting in a total of 540 models being fit to each data set. Again, the results of the analysis were compared to the results given in Punzo et al. (2020).

4.4.1 Results for Gaussian clusters

Table 4.11 provides a comparison of the classification performance of the PCSALM family for Gaussian clusters to other mixtures of factor analyzers. The PCSALM is able to provide strong and comparable classification performance to the other mixture-models, although it is slightly weaker than its competitors.

Table 4.11: Classification performance of mixtures of factor analyzers models on Gaussian clusters.

	PCSALM	MCGFA	MMtFA	EPGMM
G	2	2	2	2
Mean ARI	0.824	**0.872**	0.867	0.863
ARI Std. Dev.	0.08	**0.04**	0.05	**0.04**

Chapter 5

Real Data Analysis

5.1 AIS data set

The Australian Institute of Sport (AIS) data set (Telford and Cunningham 1991) consists of $p = 11$ numerical measurements of the blood characteristics and physiological features of $n = 202$ athletes, as well as their gender and respective sport. There is a total of 17 different classifications for an athlete's sport. Since the data set only contains 202 observations, there is not enough information to properly define 17 meaningful components. Therefore, our analysis was focused on classifying the data into the gender labels male and female. Each scale matrix constraint in the PCSALM family was fit to $G = 1, \ldots, 4$ components and $q = 1, \ldots, 5$ latent factors. The classification results and a comparison to the results obtained for the same data in Punzo et al. (2020) is provided in Tables 5.1 and 5.2. The model proposed from the PCSALM family was selected with BIC to maintain consistency with the results obtained in Punzo et al. (2020).

Table 5.1: Contingency table of solutions proposed by mixtures of factor analyzers models for the AIS data set.

	PCSALM		MCGFA			MMtFA			EPGMM		
	1	2	1	2	3	1	2	3	1	2	3
female	98	2	64	36	0	65	35	0	80	20	0
male	2	100	3	15	84	3	16	83	1	17	84

The PCSALM family showed significant improvement in classification performance when compared to other mixtures of factors analyzers. The model was able to correctly identify two groups in the data as opposed to the three group solutions by the other models in the comparison. The best fitting PCSALM model featured an ARI of 0.92 and only 4 misclassified observations, which is a noticeable increase over the ARI range of $0.54 - 0.65$ given by the competitors. The AIS data set is known to feature asymmetry in its components (McNicholas 2016), so the PCSALM family's ability to out-preform other mixtures of factors analyzers is unsurprising.

Table 5.2: Classification performance of the AIS data set for mixtures of factor analyzers models.

	PCSALM	MCGFA	MMtFA	EPGMM
Model	UUCU	UCUCU	UCCC	UUU
G	2	3	3	3
q	5	4	5	4
ARI	**0.92**	0.55	0.54	0.65

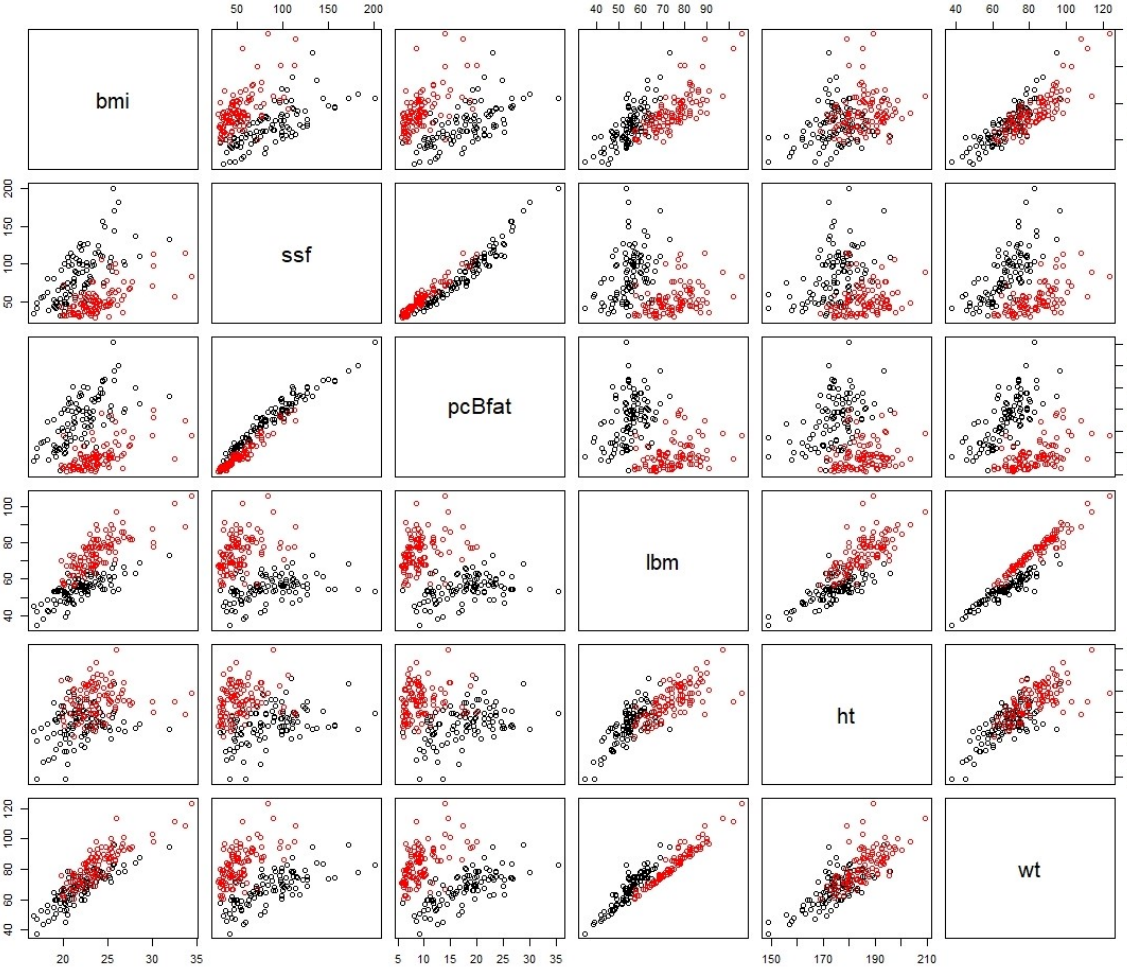
The PCSALM family also showed mild improvement in the classification performance when compared to mixtures of SAL distributions. Table 5.3 provides a comparison between the classification performance of our results and mixtures of SAL distributions. While both models correctly identified two groups in the data, the PCSALM model only misclassified 4 observations as compared to 8 misclassifications by the MSAL model.

Table 5.3: Contingency table of solutions proposed by mixtures of SAL models for the AIS data set.

	PCSALM		MSAL	
	1	2	1	2
female	98	2	100	0
male	2	100	8	94

Figure 5.1 gives the pairs plot for the $p = 6$ physiological measurement variables in the AIS data. These variables were chosen to be featured since there was a clearer visual separation in the points from each gender than in the variables quantifying blood characteristics. Points in the pairs plot were coloured using the model obtained with the PCSALM family. Black points indicate observations classified as female and red points indicate observations classified as male. The proposed solution appears to accurately reflect the separation in the components in many of the given plots. Variable combinations "pcBfat vs. lbm" and "pcBfat vs. wt" are two examples of plots that have particularly clear separation into two components.

Figure 5.1: Pairs plot of the $p = 6$ physiological measurements in the AIS data set, coloured by the PCSALM solution (black $=$ f, red $=$ m).



Chapter 6

Conclusions

6.1 Summary

This thesis introduces a family of parsimonious contaminated shifted asymmetric Laplace mixtures. Applying the constraints developed in McNicholas and Murphy (2008, 2010) to the modified factor analysis decomposition of the scale matrix parameters in mixtures of CSAL distributions led to the formulation of twelve models. Mixtures of factor analyzer models had previously been introduced with component densities that incorporated a contamination protocol with the mixtures of contaminated Gaussian factor analyzers (MCGFA) in Punzo et al. (2020). However, the models in the PCSALM family introduced in this thesis are the first mixtures of factor analyzers developed within the mixture-modelling literature with the capability of accommodating contamination and asymmetric components. The PCSALM models were implemented via an AECM algorithm for parameter estimation with initial values obtained from an embedded MSAL model, and tested in four simulation analyses and an analysis of a real data set provided by the Australian Institute of Sports.

In the first two simulations, data was generated from a mixture of SAL distribution with two components and was considered both with and without noise observations contaminating the data. The PCSALM family was able to provide a strong classification performance and was generally effective at identifying additional noise observations as bad points. In instances of poor specificity a modified ICL measure was introduced that accounts for entropy in the expected values v_{ig}. This model selection criteria was shown to be effective at providing accurate solutions when the ICL suggested an unreasonable proportion of bad points in the data. The scale matrix structures “CCCC” and “CCCU” used in conjunction with the modified ICL for model selection provided highly accurate classification performance for SAL clusters with noise. Additionally, the modified ICL provided strong classification performance for SAL clusters without noise. In this simulation, any limitations in classification performance were a result of inconsistency in the ICL as a model selection metric.

In the third and fourth simulations, the generated data featured two Gaussian components and was considered with and without noise observations contaminating the data. In these simulations the PCSALM family was compared to results from Punzo et al. (2020) for the same data. The PCSALM family was able to provide a competitive clustering performance to other mixtures of factor analyzers for observations within the components. However it was not capable of identifying noise observations as bad points when they were included in the data. Noise observations in simulation 3 were less dispersed than those in simulation 1, so it is likely that PCSALM captured the noise observations through the skewness parameter.

The data set provided by the AIS contains measurements on blood characteristics and physiological features of male and female athletes. The PCSALM family was able to provide an exceptional classification performance for this data set when attempting to distinguish the gender label for athletes, only misclassifying 4 of 202 observations. When compared to established clustering methods the PCSALM model showed improvement over all the other mixture-models considered.

6.2 Future Work

Further testing is always constructive for a models justification. Although we preformed many analyses within this thesis, more tests would allow us to more closely examine the entire scope of the PCSALM family. One of the more significant limitations faced is that the EM algorithms for the MCGFA and CSAL models had not been made publicly available in R at the time of writing this thesis. This hindered our ability to construct comparisons between the PCSALM family and its closest competitors. Additional testing with real data featuring very high dimensions is also required to see how effective the modified factor analyzer structure is in this setting.

While the modified ICL developed in Chapter 4 was shown to be an effective method of correcting for low specificity in PCSALM models, it is still undeveloped as a model selection criteria. The theoretical justification for the criterion's construction must be elaborated upon, and tests must be preformed to evaluate its effectiveness on mixture-models outside of PCSALM family. Furthermore, in this thesis we justify its use when the model identifies a high proportion of bad points that feature high levels of uncertainty in their identification. This metric is quite nebulous and open to subjectivity. It also relies on the interpretation that if the data is sufficiently comprised of bad points, then that identification is open to scrutiny. Refining the recommendation for its use to rely on a more defined combination of the entropy term of the expected values v_{ig} and the proportion of bad points identified is necessary. In the case of the PCSALM family, since several models are fit to each set of data we can obtain the mean value of these measures across all models. Assessing these measures by their relative magnitude is likely to be the most reliable signal for our modified ICL's implementation.

In Punzo et al. (2020) the MCGFA models also considered constraints on the contamination parameters $\{\rho_g, \eta_g\}$ that held them equal across components, yielding four models for each covariance matrix structure used. These constraints were extraneous to the focus of this thesis. However, their implementation in the PCSALM family may yield a stronger classification performance for some data sets.

Bibliography

Aitken, A.C. (1926). “On Bernoulli’s numerical solution of algebraic equations”. In: *Proceedings of the Royal Society of Edimburgh* 46, pp. 289–305.

Barndorff-Nielsen, O., J. Kent, and M. Sørensen (1982). “Normal Variance-Mean Mixtures and z Distributions”. In: *International Statistical Review / Revue Internationale de Statistique* 50.2, pp. 145–159.

Biernacki, C., G. Celeux, and G. Govaert (2000). “Assessing a MixtureModel for Clustering with the Integrated Completed Likelihood”. In: *IEEE Trans. Pattern Analysis and Machine Intelligence* 22.7, pp. 719–725.

Böhning, D. et al. (1994). “The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family”. In: *Annals of the Institute of Statistical Mathematics* 46, pp. 373–388.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). “Maximum likelihood from incomplete data via the EM Algorithm”. In: *Journal of the Royal Statistical Society: Series B* 39.1, pp. 1–38.

Fraley, C. and A. E. Raftery (1998). “How many clusters? Which clustering methods? Answers via model-based cluster analysis”. In: *The Computer Journal* 41.8, pp. 578–588.

Fraley, C. and A. E. Raftery (2002). “Model-Based Clustering, Discriminant Analysis, and Density Estimation”. In: *Journal of the American Statistical Association* 97.458, pp. 611–631.

Franczak, B., R. P. Browne, and P. D. McNicholas (2014). “Mixtures of shifted asymmetric Laplace distributions”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.6, pp. 1149–1157.

Franczak, B. et al. (2018). *MixSAL: Mixtures of Multivariate Shifted Asymmetric Laplace (SAL) Distributions*. R package version, 1.0. URL: https://CRAN.R-project.org/package=MixSAL.

Franczak, Brian et al. (2013). *Parsimonious Shifted Asymmetric Laplace Mixtures*. arXiv: 1311.0317 [stat.ME].

Ghahramani, Z. and G. E. Hinton (1997). *The EM algorithm for factor analyzers*. Tech. rep. CRG-TR-96-1. Toronto, ON: University of Toronto.

Hubert, L. and P. Arabie (1985). “Comparing partitions”. In: *Journal of classification* 2.1, pp. 193–218.

Kotz, S., T. J. Kozubowski, and K. Podgorski (2001). *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance.* 1st. Burkhauser Boston.

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

Lin, Tsung-I (2010). “Robust mixture modeling using multivariate skew t distributions”. In: *Statistics and Computing* 20.3, pp. 343–356.

Marriott, F. (1974). *Interpretation of multiple observations.* 1st. Academic Press.

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

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

McLachlan, G. J. and D. Peel (2000b). “Mixtures of factor analyzers”. In: *Proceedings of the Seventh International Conference on Machine Learning.* San Francisco: Morgan Kaufmann, pp. 599–606.

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

McNicholas, P. D. and T. B. Murphy (2008). “Parsimonious Gaussian mixture models”. In: *Statistics and Computing* 18.3, pp. 285–296.

McNicholas, P. D. and T. B. Murphy (2010). “Model-Based clustering of microarray expression data via latent Gaussian mixture models”. In: *Journal of Statistical Planning and Inference* 26.21, pp. 2705 –2712.

McNicholas, P. D. et al. (2010). “Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models”. In: *Computational Statistics and Data Analysis* 54.3, pp. 711–723.

Meng, X. L. and D. B. Rubin (1993). “Maximum Likelihood Estimation via the ECM algorithm: a general framework”. In: *Biometrika* 80, pp. 267–278.

Meng, X. L. and D. Van Dyk (1997). “The EM algorithm-an old folk song sung to a fast new tune”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 59.3, pp. 511–567.

Morris, K. et al. (2019). “Asymmetric clusters and outliers: Mixtures of multivariate contaminated shifted asymmetric Laplace distributions”. In: *Computational Statistics and Data Analysis* 132, pp. 145–166.

Punzo, A., M. Blostein, and P. D. McNicholas (2020). “High-dimensional unsupervised classification via parsimonious contaminated mixtures”. In: *Pattern Recognition* 98, p. 107031.

Punzo, A. and P. D. McNicholas (2016). “Parsimonious Mixtures of Multivariate Contaminated Normal Distributions”. In: *Biometrical Journal* 58.6, pp. 1506–1537.

Qui, W. and H. Joe (2020). *clusterGeneration: Random Cluster Generation (with Specified Degree of Separation)*. R package version, 1.3.7. URL: https://CRAN.R-project.org/package=clusterGeneration.

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

Schwarz, G. (1978). “Estimating the Dimension of a Model”. In: *The Annals of Statistics* 6.2, pp. 461–464.

Spearman, C. (1904). “The Proof and Measurement of Association Between Two Things”. In: *American Journal of Psychology* 15, pp. 88–103.

Telford, R.D. and R.B. Cunningham (1991). “Sex, sport and body-size dependency of hematology in highly trained athletes”. In: *Medicine and Science in Sports and Exercise* 23, pp. 788–794.

Tipping, T.E. and C.M Bishop (1999a). “Mixtures of probabilistic principal component analysers”. In: *Neural Computation* 11.2, pp. 443–482.

Tipping, T.E. and C.M Bishop (1999b). “Probabilistic principal component analysers”. In: *Journal of the Royal Statistical Society, Series B* 61, pp. 611–622.

Titterington, D. M., A. F. M. Smith, and U. E. Makov (1985). *Statistical Analysis of Finite Mixture Distributions*. Chichester: John Wiley & Sons, pp. x+243. ISBN: 0-471-90763-4.

Tukey, J.W. (1960). “A Survey of Sampling from Contaminated Distributions.” In: *Oklin, I., Ed., Contributions to Probability and Statistics*. Redwood, CA.: Stanford University Press, pp. 448–485.

Woodbury, M.A. (1950). *Inverting Modified Matrices*. Tech. rep. 42. Princeton, N.J.: Princeton University.

Appendix A: First Appendix

A.1 Expected value calculations for alternation 2

$$
\begin{aligned}
&\mathbb{E}\left[Z_{ig}V_{ig}\left(\frac{1-V_{ig}}{\sqrt{\eta_g^{(k+1)}}}\right)\mathbf{U}_{ig}|\ \mathbf{x}_i\right]\\
&= z_{ig}^{(k+1/2)}\boldsymbol{\beta}_g^{(k)}\left(v_{ig}^{(k+1/2)}+\frac{\left(1-v_{ig}^{(k+1/2)}\right)}{\sqrt{\eta_g^{(k+1)}}}\right)(\mathbf{x}_i-\boldsymbol{\mu}_g^{(k+1)})\\
&- z_{ig}^{(k+1/2)}\boldsymbol{\beta}_g^{(k)}\left(v_{ig}^{(k+1/2)}E_{1ig}^{(k+1/2)}+\left(1-v_{ig}^{(k+1/2)}\right)\widetilde{E}_{1ig}^{(k+1/2)}\right)\boldsymbol{\alpha}_g^{(k+1)}
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}\left[Z_{ig}V_{ig}W_{ig}^{-1}\mathbf{U}_{ig}|\ \mathbf{x}_i\right]\\
&= z_{ig}^{(k+1/2)}v_{ig}^{(k+1/2)}\boldsymbol{\beta}_g^{(k)}\left(E_{2ig}^{(k+1/2)}(\mathbf{x}_i-\boldsymbol{\mu}_g^{(k+1)})-\boldsymbol{\alpha}_g^{(k+1)}\right)
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}\left[Z_{ig}\left(\frac{1-V_{ig}}{\eta_g^{(k+1)}}\right)\widetilde{W}_{ig}^{-1}\mathbf{U}_{ig}|\ \mathbf{x}_i\right]\\
&= z_{ig}^{(k+1/2)}\boldsymbol{\beta}_g^{(k)}\left(\frac{1-v_{ig}^{(k+1/2)}}{\eta_g^{(k+1)}}\widetilde{E}_{2ig}^{(k+1/2)}(\mathbf{x}_i-\boldsymbol{\mu}_g^{(k+1)})-\frac{1-v_{ig}^{(k+1/2)}}{\sqrt{\eta_g^{(k+1)}}}\boldsymbol{\alpha}_g^{(k+1)}\right)
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}\left[Z_{ig}V_{ig}W_{ig}^{-1}\mathbf{U}_{ig}\mathbf{U}_{ig}'\mid \mathbf{x}_i\right]\\
&= z_{ig}^{(k+1/2)}v_{ig}^{(k+1/2)}\Bigg[\mathbf{I}_q+\boldsymbol{\beta}_g^{(k)}\boldsymbol{\Lambda}_g^{(k)}+\boldsymbol{\beta}_g^{(k)}\Bigg(E_{2ig}^{(k+1/2)}(\mathbf{x}_i-\boldsymbol{\mu}_g^{(k+1)})(\mathbf{x}_i-\boldsymbol{\mu}_g^{(k+1)})'\\
&- 2(\mathbf{x}_i-\boldsymbol{\mu}_g^{(k+1)})(\boldsymbol{\alpha}_g^{(k+1)})'+E_{1ig}^{(k+1/2)}\boldsymbol{\alpha}_g^{(k+1)}(\boldsymbol{\alpha}_g^{(k+1)})'\Bigg)\boldsymbol{\beta}_g^{(k)'}\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}\left[Z_{ig}\left(\frac{1-V_{ig}}{\eta_g^{(k+1)}}\right)W_{ig}^{-1}\mathbf{U}_{ig}\mathbf{U}_{ig}' \mid \mathbf{x}_i\right]\\
&= z_{ig}^{(k+1/2)}\frac{(1-v_{ig}^{(k+1/2)})}{\eta_g^{(k+1)}}\Bigg[\mathbf{I}_q+\boldsymbol{\beta}_g^{(k)}\boldsymbol{\Lambda}_g^{(k)}+\boldsymbol{\beta}_g^{(k)\prime}\Bigg(\widetilde{E}_{2ig}^{(k+1/2)}(\mathbf{x}_i-\boldsymbol{\mu}_g^{(k+1)})(\mathbf{x}_i-\boldsymbol{\mu}_g^{(k+1)})'\\
&-2\sqrt{\eta_g^{(k+1)}}(\mathbf{x}_i-\boldsymbol{\mu}_g^{(k+1)})(\boldsymbol{\alpha}_g^{(k+1)})'+\widetilde{E}_{1ig}^{(k+1/2)}\eta_g^{(k+1)}\boldsymbol{\alpha}_g^{(k+1)}(\boldsymbol{\alpha}_g^{(k+1)})'\Bigg)\boldsymbol{\beta}_g^{(k)\prime}\Bigg]
\end{aligned}
$$

Appendix B: Second Appendix

B.1 Expanded results for SAL clusters with noise

Table B.1: Classification performance of models selected by ICL for SAL clusters with noise.

Data Set	ARI	Sensitivity	Specificity	ICL	BIC
1	0.998	0.99	0.999	-42486.56	-42486.56
2	1.00	1.00	1.000	-44012.40	-43998.65
3	1.00	1.00	1.000	-42659.86	-42659.54
4	0.945	0.98	0.962	-39616.89	-39594.10
5	0.995	1.00	0.997	-41242.33	-41242.94
6	0.991	0.98	0.996	-51785.42	-51765.15
7	1.000	0.98	1.000	-55140.30	-55113.87
8	1.000	0.00	1.000	-51971.62	-51969.91
9	0.994	0.34	1.000	-53629.29	-53609.36
10	1.000	0.99	1.000	-53669.20	-53656.50
11	1.000	1.00	1.000	-50827.16	-50802.24
12	1.000	0.00	1.000	-51131.58	-51120.59
13	0.995	0.98	0.998	-52404.44	-52384.27
14	0.990	0.98	0.999	-54000.88	-53967.80
15	0.991	0.52	1.000	-53464.67	-53444.32
16	0.979	1.00	0.996	-53548.80	-53545.10
17	0.998	1.00	0.999	-53936.67	-53923.85
18	0.528	0.98	0.635	-52076.65	-52076.65

Data Set	ARI	Sensitivity	Specificity	ICL	BIC
19	0.945	1.00	0.965	-53782.62	53782.58
20	0.977	0.11	1.000	-53101.57	-53089.88
21	0.998	0.98	0.999	-54146.19	-54115.89
22	0.984	0.98	1.000	-52514.57	-52495.91
23	1.000	0.98	1.000	-52732.18	-52693.90
24	1.000	0.00	1.000	-54346.19	-54344.66
25	0.998	0.74	1.000	-53394.01	-53393.33
26	0.966	0.99	0.992	-53565.66	-53557.56
27	0.953	0.98	0.979	-52960.52	-52940.86
28	0.993	0.95	0.997	-52748.44	-52734.11
29	0.295	0.99	0.540	-51123.87	-51110.50
30	0.976	0.00	1.000	-55308.51	-55301.21

Table B.2: Count of models that satisfied performance benchmarks for ARI and sensitivity for SAL clusters with noise.

Data Set	ARI > 0.95	Sensitivity = 1	ARI > 0.95 & Sensitivity = 1
1	136	27	27
2	134	21	21
3	144	11	11
4	138	18	15
5	144	30	30
6	144	16	16
7	144	15	15
8	128	10	10
9	135	18	18
10	141	14	14
11	129	26	26
12	127	8	8

Data Set	ARI > 0.95	Sensitivity = 1	ARI > 0.95 & Sensitivity = 1
13	133	18	13
14	144	19	19
15	135	19	19
16	88	16	8
17	135	22	22
18	120	12	10
19	133	23	22
20	103	14	14
21	144	32	32
22	80	21	17
23	144	36	36
24	136	14	14
25	129	19	19
26	129	19	19
27	137	19	19
28	144	5	5
29	129	9	9
30	144	30	30