Dimension Reduction and Clustering using Non-Elliptical Mixtures

by

Katherine Morris

A Thesis

Presented to

The University of Guelph

In partial fulfilment of requirements for the degree of

Doctor of Philosophy in Mathematics and Statistics

(Applied Statistics)

Guelph, Ontario, Canada

© Katherine Morris, January, 2014
Finite mixtures of non-elliptical distributions (specifically the shifted asymmetric Laplace and the generalized hyperbolic) are considered to introduce dimension reduction methods for model-based clustering. The approaches are based on existing work on reducing dimensionality in the case of finite Gaussian mixtures. The methods rely on identifying a reduced subspace of the data by considering the extent to which group means and group covariances vary. This subspace contains linear combinations of the original data, which are ordered by importance via the associated eigenvalues. Observations can be projected onto the subspace and the resulting set of variables captures most of the clustering structure available in the data. The algorithms are illustrated using simulated and real data.

Furthermore, methods of detecting outliers are developed for model-based clustering using mixtures of contaminated shifted asymmetric Laplace distributions, and mixtures of contaminated skew-normal distributions. The approaches are based on existing work for outlier detection in the context of contaminated Gaussian mix-
tures. The main idea is to introduce a contamination factor which increases the dispersion of the fitted distribution by altering the skewness and covariance parameters. An expectation-conditional maximization algorithm is employed to obtain maximum likelihood estimates for the parameters in the model. Thus each observation is given a posterior probability of belonging to a particular group, and of being an outlier or not. The performance of the methods is tested on simulated and real data.
iv

To Phillip, with love
Acknowledgments

First of all, I would like to express my sincere gratitude to my advisor Prof. Paul McNicholas for his guidance, dedication and support during my years at the University of Guelph. I am very thankful for the opportunities of self development through conferences and workshops that he always encouraged. His dedication towards his students and enthusiasm for research inspired me to continue my studies.

I am also grateful for having a friendly learning environment in the Department of Mathematics & Statistics and the help provided by the members of the McNicholas Research Group. I gratefully acknowledge the financial support obtained through the Dean’s Scholarship and the Ontario Graduate Scholarship.

I would like to thank my examining and advisory committee Prof. Rebecca Nugent, Prof. Paul McNicholas, Prof. Rob Deardon, Prof. Zeny Feng, Prof. Ryan Browne, Prof. Antonio Punzo and Prof. Ed Carter for their time.

Furthermore, I am grateful to my family for their unwavering motivation and support. Last but not least, I am greatly indebted to Phillip for his unconditional love and endless encouragement.
Publications

The following articles based on this work are published, under review, or in preparation.


Table of Contents

List of Tables ix
List of Figures xii

1 Introduction 1
   1.1 Overview ......................................................... 1
   1.2 Background ..................................................... 2
      1.2.1 Dimension reduction for model-based clustering using Gaussian
             mixtures ......................................................... 4
      1.2.2 Dimension reduction for model-based clustering using non-Gaussian
             mixtures ......................................................... 8
      1.2.3 Parameter estimation for model-based clustering .................. 9
      1.2.4 Performance assessment for model-based clustering ........... 11
      1.2.5 Outlier detection for clustering using contaminated Gaussian
             mixtures ......................................................... 11

2 Dimension Reduction and Clustering using Shifted Asymmetric Laplace
   Mixtures 16
   2.1 Methodology ...................................................... 16
   2.2 Applications ..................................................... 21
      2.2.1 Simulated data ............................................... 21
      2.2.2 Real data .................................................... 23
   2.3 Summary ........................................................ 30

3 Dimension Reduction and Clustering using Generalized Hyperbolic
   Mixtures 32
   3.1 Methodology ...................................................... 33
   3.2 Applications ..................................................... 41
      3.2.1 Simulated data ............................................... 41
      3.2.2 Real data .................................................... 46
   3.3 Summary ........................................................ 59

4 Outlier Detection for Clustering using Mixtures of Contaminated
   Shifted Asymmetric Laplace Distributions 62
   4.1 Methodology ...................................................... 62
   4.2 Applications ..................................................... 75
4.2.1 Simulated data ........................................... 75
4.2.2 Real data ............................................... 79
4.3 Summary .................................................. 90

5 Outlier Detection for Clustering using Mixtures of Contaminated Skew Normal Distributions 91
5.1 Methodology ................................................. 91
5.2 Applications ............................................... 100
  5.2.1 Simulated data ........................................ 100
  5.2.2 Real data ............................................. 101
5.3 Summary .................................................. 108

6 Conclusion 110

Bibliography 113

A Appendix 120
  A.1 Outlier detection for clustering using mixtures of contaminated shifted asymmetric Laplace distributions ........................................ 120
  A.1.1 Parameter updates ..................................... 120
  A.1.2 Extension to model-based classification ............... 128
  A.2 Outlier detection for clustering using mixtures of contaminated skew normal distributions ........................................ 137
  A.2.1 Parameter updates ..................................... 137
  A.2.2 Extension to model-based classification ............... 140
  A.3 Dimension Reduction and Clustering using Generalized Hyperbolic Mixtures ........................................ 147
  A.3.1 Selected genes for the colon data. ....................... 147
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Average ARI (with standard errors in brackets) based on 25 simulations.</td>
<td>23</td>
</tr>
<tr>
<td>2.2</td>
<td>Measurements for the Swiss banknote data.</td>
<td>25</td>
</tr>
<tr>
<td>2.3</td>
<td>Model-based clustering results for the best SALMMDR and GMMDR models, in terms of BIC, fitted to the Swiss banknote data.</td>
<td>25</td>
</tr>
<tr>
<td>2.4</td>
<td>Summary of results for the best models fitted to the Swiss banknote data.</td>
<td>25</td>
</tr>
<tr>
<td>2.5</td>
<td>Measurements for the yeast data.</td>
<td>27</td>
</tr>
<tr>
<td>2.6</td>
<td>Model-based clustering results for the best SALMMDR and GMMDR models, in terms of BIC, fitted to the yeast data.</td>
<td>28</td>
</tr>
<tr>
<td>2.7</td>
<td>Summary of results for the best models fitted to the yeast data.</td>
<td>28</td>
</tr>
<tr>
<td>2.8</td>
<td>Measurements for the AIS data.</td>
<td>29</td>
</tr>
<tr>
<td>2.9</td>
<td>Model-based clustering results for the best SALMMDR and GMMDR models, in terms of BIC, fitted to the Australian Institute of Sport data.</td>
<td>30</td>
</tr>
<tr>
<td>2.10</td>
<td>Summary of results for the best models fitted to the Australian Institute of Sport data.</td>
<td>30</td>
</tr>
<tr>
<td>3.1</td>
<td>Summary of results for the HMMDR and GMMDR models fitted to the simulated data in Scenario I without noise, based on 300 runs.</td>
<td>44</td>
</tr>
<tr>
<td>3.2</td>
<td>Summary of results for the HMMDR and GMMDR models fitted to the simulated data in Scenario I with noise, based on 300 runs.</td>
<td>44</td>
</tr>
<tr>
<td>3.3</td>
<td>Summary of clustering results for the best HMMDR and GMMDR models for the simulated high-dimensional data.</td>
<td>46</td>
</tr>
<tr>
<td>3.4</td>
<td>Details about the component covariance structures for each of the four MMDR methods used.</td>
<td>48</td>
</tr>
<tr>
<td>3.5</td>
<td>Measurements taken for the female vole data.</td>
<td>50</td>
</tr>
<tr>
<td>3.6</td>
<td>Summary of model-based clustering results for the bank data.</td>
<td>51</td>
</tr>
<tr>
<td>3.7</td>
<td>Summary of model-based clustering results for the voles data.</td>
<td>51</td>
</tr>
<tr>
<td>3.8</td>
<td>Summary of model-based classification and discriminant analysis results for the bank data, based on 25 runs.</td>
<td>51</td>
</tr>
<tr>
<td>3.9</td>
<td>Summary of model-based classification and discriminant analysis results for the voles data, based on 25 runs.</td>
<td>51</td>
</tr>
<tr>
<td>3.10</td>
<td>Chemical and physical properties available for the wine data in gclus.</td>
<td>52</td>
</tr>
<tr>
<td>3.11</td>
<td>Model-based clustering, classification, and discriminant analysis results for our HMMDR approach fitted to the wine data. Model-based classification and discriminant analysis results are based on 25 runs.</td>
<td>52</td>
</tr>
<tr>
<td>3.12</td>
<td>Summary of model-based clustering results for the wine data.</td>
<td>52</td>
</tr>
</tbody>
</table>
3.13 Summary of model-based classification and discriminant analysis results for the wine data, based on 25 runs. 
3.14 Tissue sample properties of the Wisconsin breast cancer data.
3.15 Model-based clustering, classification, and discriminant analysis results for our HMMDR approach on the breast cancer data. Model-based classification and discriminant analysis results are based on 25 runs.
3.16 Summary of model-based clustering results for the breast cancer data.
3.17 Summary of model-based classification and discriminant analysis results for the breast tissue data, based on 25 runs.
3.18 Model-based clustering, classification, and discriminant analysis results for the colon data. Model-based classification and discriminant analysis results are based on 25 runs.
3.19 Summary of model-based clustering results for the best models fitted to the colon data.
3.20 Summary of model-based classification and discriminant analysis results for the colon data, based on 25 runs.

4.1 Free parameters in the contaminated SAL mixture model.
4.2 Parameters used to simulate data for a two component SAL mixture.
4.3 Clustering results for the best model, in terms of ARI, fitted to the simulated SAL data in scenario one.
4.4 Clustering results for the best model, in terms of BIC, fitted to the simulated SAL data in scenario two.
4.5 Clustering results for the best model, in terms of BIC, fitted to the AIS data using a contaminated SAL mixture.
4.6 Clustering results for the model fitted to the AIS data using \texttt{tclust} with $\alpha = 0.00495$.
4.7 Model-based clustering results fitted to the wine data using a contaminated SAL mixture.
4.8 Clustering results for the model fitted to the AIS data using \texttt{tclust} with $\alpha = 0.0056$.
4.9 Clustering results for the best model, in terms of BIC, fitted to the yeast data using a contaminated SAL mixture.
4.10 Clustering results for the model fitted to the yeast data using \texttt{tclust} with $\alpha = 0.0032$.
4.11 Fatty acids present in the olive oil data.
4.12 Clustering results for the best model, in terms of BIC, fitted to the olive oil data using a contaminated SAL mixture.
4.13 Summary of estimates for $\rho$ and $\lambda$ obtained while fitting contaminated SAL mixtures to the real data.

5.1 Parameters used to simulate data for a two component SN mixture.
5.2 Clustering results for the best model, in terms of ARI, fitted to the simulated SN data.
5.3 Clustering results for the best model, in terms of BIC, fitted to the bank data using a contaminated SN mixture. 103
5.4 Clustering results for the model fitted to the bank data using tclust with $\alpha = 0.01$. 104
5.5 Clustering results for the best model, in terms of BIC, fitted to the voles data using a contaminated SN mixture. 106
5.6 Clustering results for the model fitted to the voles data using tclust with $\alpha = 0.0102$. 107
5.7 Summary of estimates for $\rho$ and $\lambda$ obtained while fitting contaminated SN mixtures to the real data. 107
5.8 Summary of clustering results for the best model, in terms of BIC, fitted to the real data using contaminated mixtures. 109

A.1 Genes selected with SAM for the colon data. 147
List of Figures

2.1 Scatterplot of simulated data for Scenario I indicating the cluster structure. 22
2.2 Scatterplot of simulated data for Scenario II indicating the cluster structure. 23
2.3 Scatterplots illustrating the SALMMDR directions for the Swiss banknote data; the numbers (left-hand plot) or shapes (right-hand plot) signify true classification and the colour indicates estimated SALMMDR cluster allocation. 26
2.4 Scatterplots illustrating the SALMMDR directions for the yeast data; the numbers (left-hand plot) or shapes (right-hand plot) signify true classification and colour indicates the estimated SALMMDR cluster allocation. 29

3.1 Diagram showing the relationship between the model-based methods discussed. 35
3.2 Pairs plot illustrating a typical example of the simulated multivariate Gaussian components in Scenarios I. 42
3.3 Pairs plot illustrating a typical example of the simulated multivariate Gaussian components in Scenarios II. Only some of the noise components are shown. 43
3.4 Pairs plot illustrating a typical example of the simulated multivariate hyperbolic components in $p = 20$ dimensions. 45
3.5 Plots of some of the estimated HMMDR directions for the wine data obtained from model-based clustering (direction 2 vs. 4 on the left-hand side, and direction 4 vs. 5 on the right-hand side). Symbols indicate true cluster membership and colours indicate the estimated HMMDR cluster allocation. The edge histograms depict the estimated distributions of the observations in each cluster. 53
3.6 Plots of some of the estimated HMMDR directions for the breast cancer data obtained from model-based clustering (direction 1 vs. 3 on the left-hand side, and direction 2 vs. 3 on the right-hand side). Symbols indicate true cluster membership and colours indicate the estimated HMMDR cluster allocation. The edge histograms depict the estimated distributions of the observations in each cluster. 56
3.7 Plots of some of the estimated HMMDR directions for the colon data obtained from model-based clustering (direction 2 vs. 3 on the left-hand side, and direction 1 vs. 2 on the right-hand side). Symbols indicate true cluster membership and colours indicate the estimated HMMDR cluster allocation. The edge histograms depict the estimated distributions of the observations in each cluster. .......................... 59

4.1 Contour plot of a contaminated SAL mixture superimposed on its uncontaminated SAL mixture counterpart. .......................... 64
4.2 Scatterplot of simulated data for scenario one, with added contours. Color indicates the estimated component membership and shape indicates the known membership. .......................... 77
4.3 Scatterplot of simulated data for scenario one, with added contours. Color indicates the estimated component membership and shape indicates the known membership. Simulated outliers are depicted by the + symbol. .......................... 78
4.4 Principal components for the AIS data indicating the cluster structure and artificial outlier. Colour represents known component membership. .......................... 81
4.5 Scatterplot of the first two discriminant coordinates for the AIS data obtained with \textit{tclust}. Colour and shape represent estimated cluster membership. .......................... 82
4.6 Principal components for the wine data indicating the cluster structure and artificial outlier. Colour represents known component membership. .......................... 83
4.7 Scatterplot of the first two discriminant coordinates for the wine data obtained with \textit{tclust}. Colour and shape represent estimated cluster membership. .......................... 84
4.8 Principal components for the yeast data indicating the cluster structure and natural outliers. Colour represents known component membership. .......................... 86
4.9 Scatterplot of the first two discriminant coordinates for the yeast data obtained with \textit{tclust}. Colour and shape represent estimated cluster membership. .......................... 86
4.10 Figure depicts the diagnostic plot from the ROBPCA procedure for the olive oil data, indicating possible outliers. .......................... 88
4.11 Figure shows the pairs plot of the data projected on the first four robust PCs. Colour indicates known cluster membership. .......................... 88

5.1 Contour plot of a contaminated SN mixture superimposed on its uncontaminated SAL mixture counterpart. .......................... 95
5.2 Scatterplot of simulated data for scenario one, with added contours. Color indicates the estimated component membership and shape indicates the known membership. .......................... 102
5.3 Principal components for the bank data indicating the cluster structure and natural outliers. Colour represents known component membership. 104
5.4 Scatterplot of the first two discriminant coordinates for the bank data obtained with \texttt{tclust}. Colour and shape represent estimated cluster membership. ................................................................. 105

5.5 Principal components for the voles data indicating the cluster structure and artificial outlier. Colour represents known component membership. 107

5.6 Scatterplot of the first two discriminant coordinates for the voles data obtained with \texttt{tclust}. Colour and shape represent estimated cluster membership. ................................................................. 108
Chapter 1

Introduction

1.1 Overview

This thesis is an exposition of three recent pieces of work within the general framework of model-based clustering for finite mixtures. Approaches to dimension reduction using non-elliptical mixtures are proposed. The methods incorporate the use of mixture modelling with shifted asymmetric Laplace distributions and generalized hyperbolic distributions. Furthermore, models based on mixtures of contaminated non-elliptical distributions, specifically shifted asymmetric Laplace and skew-normal distributions, are introduced as a convenient avenue to modelling and detecting outliers.

The thesis is organized into six chapters. Chapter 1 gives an overview of model-based clustering and defines the objectives of the research. It also provides the background for dimension reduction and outlier detection within the model-based clustering framework of Gaussian mixtures.

Chapter 2 develops the methodology for dimension reduction for model-based clustering via mixtures of shifted asymmetric Laplace distributions (Franczak et al., 2012). The proposed algorithm, SALMMDR, is obtained by following the work
of Scrucca (2010), where the dimension reduction method for Gaussian mixtures (GMMDR) was first introduced.

Chapter 3 proposes the analogue of GMMDR for mixtures of generalized hyperbolic distributions (Browne and McNicholas, 2013b). The method, called HM-MDR, is derived similarly to SALMMDR. Extensions to model-based classification and discriminant analysis are also introduced.

Chapter 4 introduces a method for outlier detection using mixtures of contaminated shifted asymmetric Laplace distributions. This is derived by following the work of Punzo and McNicholas (2013) for contaminated Gaussian mixture models.

Chapter 5 extends the outlier detection method for mixtures of shifted asymmetric Laplace distributions to mixtures of contaminated skew-normal distributions. The derivation of the model is done analogously to Punzo and McNicholas (2013) by considering the skew-normal distribution model proposed by Sahu et al. (2003).

Chapter 6 presents concluding remarks and ideas for future work.

1.2 Background

Modern data sets used in many practical applications have grown in size and complexity, compelling the use of clustering and classification algorithms based on probability models. The model-based approach assumes that data are generated by a finite mixture of probability distributions. A \( p \)-dimensional random vector \( \mathbf{X} \) is said to arise from a parametric finite mixture distribution if its density is a convex
set of probability densities, i.e.,

\[ f(x|\vartheta) = \sum_{g=1}^{G} \pi_g f_d(x|\theta_g), \quad (1.1) \]

where \( G \) is the number of components, \( \pi_g \) are mixing proportions, so that \( \sum_{g=1}^{G} \pi_g = 1 \) and \( \pi_g > 0 \), and \( \vartheta = (\pi_1, \ldots, \pi_G, \theta_1, \ldots, \theta_G) \) is the parameter vector. The \( f_d(x|\theta_g) \) are called component densities and \( f(x|\vartheta) \) is formally referred to as a \( G \)-component parametric finite mixture distribution. The likelihood for the mixture model in (1.1) is defined as

\[ L(\vartheta|x) = \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g f_d(x|\theta_g). \quad (1.2) \]

The use of mixture models in clustering applications can be traced back a half-century to an application of Gaussian mixture models (Wolfe, 1963). Gaussian mixture model-based approaches have been very popular due to their mathematical tractability, and until recently, they dominated literature in the field. Extensive details on finite mixture models are given by Everitt and Hand (1981), McLachlan and Basford (1988), and McLachlan and Peel (2000).

Many dimension reduction methods summarize the information available through a reduced combination of the original variables. However, in terms of visualization, they do not always provide adequate information on the potential structure of the data at hand. The methods proposed in Chapter 2 and 3 address this issue by revealing the underlying data clusters. At the same time, using heavy-tailed distributions, such as the generalized hyperbolic, to model data can be advantageous because they assign correct weights to extreme events (McNeil et al., 2005).

Often, real data are “contaminated” by outliers that affect the estimation of
component means and covariance matrices, as discussed for example in Bock (2002). Thus an important practical application is the development of methods capable of detecting the outliers and performing robust parameter estimation when they are present. Several approaches that discuss ways of coping with such issues are: mixtures of multivariate skew-t distributions (for e.g., Wang et al., 2009; Lin, 2010; Lee and McLachlan, 2012; Vrbik and McNicholas, 2012), mixtures of multivariate t-distributions with the Box-Cox transformation (Lo and Gottardo, 2012), mixtures of multivariate normal inverse Gaussian distributions (Karlis and Santourian, 2009), and mixtures of generalized hyperbolic distributions (Browne and McNicholas, 2013b). However, in the methods cited, there is no automatic way of outlier detection unless one defines some exogenous trimming rule. Chapter 4 and 5 address this issue by introducing convenient ways of detecting and modelling outliers.

1.2.1 Dimension reduction for model-based clustering using Gaussian mixtures

Scrucca (2010) proposed a method of dimension reduction for model-based clustering within the Gaussian mixtures framework, called GMMDR. Consider a $G$-component Gaussian mixture model

$$f(x|\theta) = \sum_{g=1}^{G} \pi_g f_N(x|\mu_g, \Sigma_g),$$  \hspace{1cm} (1.3)

with component density

$$f_N(x|\mu_g, \Sigma_g) = \frac{\exp \left\{-\frac{1}{2}(x - \mu_g)^\top \Sigma_g^{-1}(x - \mu_g)\right\}}{(2\pi)^{\frac{p}{2}} |\Sigma_g|^{\frac{1}{2}}}. \hspace{1cm} (1.4)$$
The goal of GMMDR is to estimate a subspace that captures most of the clustering structure contained in the data. At the core of the method lies the sliced inverse regression (SIR) work of Li (1991, 2000), which reduces data dimensionality by considering the variation in group means to identify the subspace. Scrucca (2010) extended the SIR ideas to also include variation of group covariances. The members of the subspace arise through linear combinations of the original data, and are ordered by importance via their associated eigenvalues.

Consider a \((p \times 1)\) vector \(X\) of random variables and a discrete random variable \(Y\) taking \(G\) distinct values to indicate the \(G\) clusters. Let \(\beta\) denote a fixed \((p \times d)\) matrix with \(d \leq p\) such that

\[
Y \perp \perp X | \beta^\top X. \tag{1.5}
\]

The conditional independence in (1.5) tells us that the distribution of \(Y|X\) is the same as the distribution of \(Y|\beta^\top X\) for all values of \(X\) in its marginal sample space. As a consequence, we can replace the \((p \times 1)\) vector \(X\) with the \((d \times 1)\) vector \(\beta^\top X\) without loss of clustering information. If \(d < p\) then we have reduced the dimension of the predictor vector.

Li (1991) defines the basis for the subspace \(S(\beta)\) given by \(\beta\) as a dimension-reduction subspace (DRS) for the regression of \(Y\) on \(X\). A minimum DRS may not necessarily be unique but if several such subspaces exist, then they all have the same dimension.

The assumption in (1.5) implies that \(P(Y = g|X) = P(Y = g|\beta^\top X)\) so the
density for the $g$-th component (1.4) of the mixture model in (1.3) can be written as

$$f_N(x) = \frac{P(Y = g|x) f_X(x)}{P(Y = g)} = \frac{P(Y = g|\beta^\top x) f_X(x)}{P(Y = g)} = f_N(\beta^\top x) \frac{f_X(x)}{f_{\beta^\top x}(\beta^\top x)}.$$  \hspace{1cm} (1.6)

Also, for any two groups, $i$ and $j$ say, the ratio of the Gaussian mixture model densities becomes

$$\frac{f_i(x)}{f_j(x)} = \frac{f_i(\beta^\top x)}{f_j(\beta^\top x)}.$$  \hspace{1cm} (1.7)

Given (1.5), the ratio of the conditional densities is the same whether it is computed on the original variables space or on the DRS. Thus the clustering information is contained completely in $S(\beta)$.

Given that an observation is assigned to a cluster $g$ by the MAP (1.16), i.e., to the cluster for which the conditional probability given the data is a maximum such that

$$\arg \max_g P(Y = g|x) = \frac{\pi_g f_N(x)}{\sum_{j=1}^G \pi_j f_j(x)},$$  \hspace{1cm} (1.8)

which is equivalent to

$$\arg \max_g P(Y = g|\beta^\top x) = \frac{\pi_g f_N(\beta^\top x)}{\sum_{j=1}^G \pi_j f_j(\beta^\top x)},$$  \hspace{1cm} (1.9)

by (1.6). Hence, the assignment of an observation to a cluster is unchanged when performed on the DRS.

To determine the smallest subspace that captures the clustering information contained within the data, one needs to identify those directions where the cluster
means $\mu_g$ and the cluster covariances $\Sigma_g$ vary as much as possible, provided that each direction is $\Sigma$-orthogonal to the others. Finding these directions is achieved through the generalized eigen-decomposition of the kernel matrix $M$, defined by Scrucca (2010) as

$$Mv_i = l_i \Sigma v_i,$$

where $l_1 \geq l_2 \geq \cdots \geq l_d > 0$ and $v_i^\top \Sigma v_j = 1$ if $i = j$ and $v_i^\top \Sigma v_j = 0$ otherwise. Note that there are $d \leq p$ directions that span the subspace. This kernel contains the variations in cluster means

$$M_I = \sum_{g=1}^{G} \pi_g (\mu_g - \mu)(\mu_g - \mu)^\top,$$

and variations in cluster covariances

$$M_{II} = \sum_{g=1}^{G} \pi_g (\Sigma_g - \bar{\Sigma})\Sigma^{-1}(\Sigma_g - \bar{\Sigma})^\top,$$

such that $M = M_I \Sigma^{-1} M_I + M_{II}$.

Here, $\mu = \sum_{g=1}^{G} \pi_g \mu_g$ is the global mean, $\Sigma = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^\top$ is the covariance matrix, and $\bar{\Sigma} = \sum_{g=1}^{G} \pi_g \Sigma_g$ is the pooled within-cluster covariance matrix. Fitting of the Gaussian mixtures within GMMDR is done using the MCLUST family (Fraley and Raftery, 1999). The MCLUST family is a subset of ten of the fourteen Gaussian parsimonious clustering models (GPCMs) introduced by Celeux and Govaert (1995). The GPCMs arise from the imposition of various constraints on eigen-decomposed component covariance matrices (cf. Banfield and Raftery, 1993; Fraley and Raftery, 2002).
1.2.2 Dimension reduction for model-based clustering using non-Gaussian mixtures

The work of Scrucca (2010) has already been extended to one other non-Gaussian mixture settings. In the context of model-based clustering, Morris et al. (2013) proposed a $t$-distribution analogue of GMMDR, called $t$MMDR. This approach uses the $t$EIGEN family of models (Andrews and McNicholas, 2012a), which is a $t$-analogue of the GPCM family of models. The most general, i.e., unconstrained, member of the $t$EIGEN family is a mixture model with component density

$$f_t(x|\mu_g, \Sigma_g, \nu_g) = \frac{\Gamma\left(\frac{\nu_g + p}{2}\right)|\Sigma_g|^{-\frac{1}{2}}}{\pi^{\frac{p}{2}} \Gamma\left(\frac{\nu_g}{2}\right)(1 + \frac{\delta(x, \mu_g|\Sigma_g)}{\nu_g})^{\frac{\nu_g + p}{2}}}, \tag{1.13}$$

where $\mu_g$ is the mean, $\Sigma_g$ is a scale matrix, $\nu_g$ is the number of degrees of freedom, and $\delta(x, \mu_g|\Sigma_g) = (x - \mu_g)^\top \Sigma_g^{-1}(x - \mu_g)$ is the squared Mahalanobis distance between $x$ and $\mu_g$. Of course, a mixture model with component density (1.13) is just a mixture of multivariate $t$-distributions, which has been applied for robust clustering for some time (McLachlan and Peel, 1998; Peel and McLachlan, 2000).

The $t$MMDR algorithm performed favourably when compared to its Gaussian analogue, as well as other well-established dimension reduction methods. In Chapter 3, we also apply $t$MMDR to model-based classification and discriminant analysis.
1.2.3 Parameter estimation for model-based clustering

To facilitate discussion of parameter estimation, we introduce $z_{ig}$ to denote component membership, so that

$$z_{ig} = \begin{cases} 
1 & \text{if observation } x_i \text{ belongs to component } g, \text{ and} \\
0 & \text{otherwise.}
\end{cases} \tag{1.14}$$

Parameter estimation for the mixture models discussed in Chapters 2 and 3 is carried out using the expectation-maximization (EM) algorithm (Baum et al., 1970; Orchard and Woodbury, 1972; Sundberg, 1974; Dempster et al., 1977). The EM algorithm is an iterative procedure for finding maximum likelihood estimates when data are incomplete or are treated as being incomplete. The EM algorithm is based on the complete-data log-likelihood of a mixture model, where the complete-data comprise the observed $x_i$ and the missing $z_{ig}$.

Two steps are iterated until convergence is reached. In the expectation step (E-step), the expected value of the complete-data log-likelihood is computed. Then in the maximization step (M-step), the expected value of the complete-data log-likelihood is maximized with respect to the model parameters.

Convergence is determined via an asymptotic estimate of the log-likelihood for a mixture model at iteration $k + 1$, namely

$$l_{\infty}^{k+1} = l^k + \frac{l^{k+1} - l^k}{1 - a^k}, \tag{1.15}$$

where $a^k = (l^{k+1} - l^k)/(l^k - l^{k-1})$ denotes the Aitken acceleration (Aitken, 1926).

After convergence, component memberships are usually estimated based on
the maximum a posteriori (MAP) classification given by

\[
\text{MAP}\{\hat{z}_{ig}\} = \begin{cases} 
1 & \text{if } \max_g \{\hat{z}_{ig}\} \text{ occurs at component } g, \text{ and} \\
0 & \text{otherwise.}
\end{cases}
\] (1.16)

Note that the MAP classification is used to give the predicted classifications (clusterings) in the data analyses presented herein.

In our applications we assume that the number of components \( G \) is unknown. This is not unusual in real model-based clustering applications, where a criterion is often used to determine \( G \). The Bayesian information criterion (BIC; Schwarz, 1978) is the most popular choice and is given by

\[
\text{BIC} = 2l(x, \hat{\theta}) - r \log n,
\] (1.17)

where \( l(x, \hat{\theta}) \) is the maximized log-likelihood, \( \hat{\theta} \) denotes the maximum likelihood estimate of \( \theta \), \( r \) represents the number of free parameters, and \( n \) is the number of observations.

Similarly, parameter estimation for the mixture models discussed in Chapters 4 and 5 is carried out using the expectation-conditional maximization (ECM) algorithm (Meng and Rubin, 1993). In the ECM algorithm the M-step is replaced by a number of conditional maximization steps (CM-step) that are more computationally efficient. Extensive details on the EM algorithm and its variants are given by McLachlan and Krishnan (2008).
1.2.4 **Performance assessment for model-based clustering**

We can compare our predicted classifications to the true class labels in each data analysis case. To do this, we use the adjusted Rand index (ARI; Hubert and Arabie, 1985), which is the Rand index (Rand, 1971) corrected for chance agreement. The Rand index is based on pairwise agreement and takes a value between 0 and 1, where 1 indicates perfect agreement between two partitions. The correction that leads to the ARI accounts for the fact that random classification is expected to result in some correct agreements; accordingly, the ARI has an expected value of 0 under random classification and, as with the Rand index, perfect classification corresponds to a value of 1. Negative ARI values are possible and indicate classification results worse than would be expected by random classification.

1.2.5 **Outlier detection for clustering using contaminated Gaussian mixtures**

Punzo and McNicholas (2013) introduced a contaminated Gaussian distribution, having two parameters indicating the proportion of outliers and degree of contamination, which offers a natural way of modelling and detecting outliers. For a finite mixture of form (1.1), Punzo and McNicholas (2013) adopt the component density

$$f_d(x|\bar{\theta}_g) = \lambda_g f_N(x|\mu_g, \Sigma_g) + (1 - \lambda_g) f_N(x|\mu_g, \rho_g \Sigma_g),$$

(1.18)

where \(\lambda_g \in [0, 1]\) denotes the proportion of outliers, \(\rho_g > 1\) represents the contamination factor, and \(f_N\) indicates the distribution of a \(p\)-variate Gaussian component.
appearing in (1.4). The result is the mixture of contaminated Gaussian distributions defined by

\[ f(x|\theta) = \sum_{g=1}^{G} \pi_g \left\{ \lambda_g f_N(x|\mu_g, \Sigma_g) + (1 - \lambda_g) f_N(x|\mu_g, \rho_g \Sigma_g) \right\}. \]  

(1.19)

Since there are \( p(p+1)/2 \) free parameters for each \( \Sigma_g \), Punzo and McNicholas (2013) introduce parsimony in the model (1.19) by following Celeux and Govaert (1995), and considering the eigen-decomposition

\[ \Sigma_g = \delta_g \Gamma_g \Delta_g \Gamma_g^\top, \]  

(1.20)

where \( \delta_g = |\Sigma_g|^{1/p} \), \( \Delta_g \) is the scaled (\(|\Delta_g| = 1\)) diagonal matrix of eigenvalues of \( \Sigma_g \) (sorted in decreasing order), and \( \Gamma_g \) is a \( p \times p \) orthogonal matrix where the columns are the normalized eigenvectors of \( \Sigma_g \) (ordered according to their eigenvalues).

By imposing constraints on \( \delta_g, \Gamma_g \) and \( \Delta_g \), Punzo and McNicholas (2013) generate a family of fourteen parsimonious mixtures of contaminated Gaussian distributions (PMCGD) models. Full details on this family appears in Table 1 of Punzo and McNicholas (2013). The ECM algorithm (Meng and Rubin, 1993) is used to fit these PMCGD models. A brief discussion of the ECM appears earlier in this chapter.

There are two sources of missing data here, namely

1. the component memberships, \( z_{ig} \) (1.14), and

2. not knowing whether an observation in component \( g \) is “good” or “bad”.

The parameter \( v_{ig} \) is then introduced to denote the second source of missing data,
such that
\[
v_{ig} = \begin{cases} 
1 & \text{if observation } x_i \text{ in group } g, \text{ is “good”} \\
0 & \text{if observation } x_i \text{ in group } g, \text{ is “bad”} 
\end{cases}
\] (1.21)

Punzo and McNicholas (2013) develop the parameter updates in the general context of model-based classification. By letting the known/labelled observations be represented by \( \tilde{z}_{ig} \), the complete-data log-likelihood equation corresponding to the contaminated Gaussian mixture model in (1.19) becomes
\[
l(\vartheta | x) = l_1(\vartheta | x) + l_2(\vartheta | x) + l_3(\vartheta | x),
\] (1.22)

where
\[
l_1(\vartheta | x) = \sum_{i=1}^{k} \sum_{g=1}^{G} \tilde{z}_{ig} \log(\pi_g) + \sum_{i=k+1}^{n} \sum_{g=1}^{G} z_{ig} \log(\pi_g),
\] (1.23)

and
\[
l_2(\vartheta | x) = \sum_{i=1}^{k} \sum_{g=1}^{G} \tilde{z}_{ig} [v_{ig} \log(\lambda_g) + (1 - v_{ig}) \log(1 - \lambda_g)] + \sum_{i=k+1}^{n} \sum_{g=1}^{G} z_{ig} [v_{ig} \log(\lambda_g) + (1 - v_{ig}) \log(1 - \lambda_g)],
\] (1.24)

and
\[
l_3(\vartheta | x) = -\frac{1}{2} \sum_{i=1}^{k} \sum_{g=1}^{G} \left\{ \tilde{z}_{ig} \log |\Sigma_g| + p \tilde{z}_{ig} (1 - v_{ig}) \log(\rho_g) \right\} + \tilde{z}_{ig} \left( v_{ig} + \frac{1 - v_{ig}}{\rho_g} \right) \delta(x_i, \mu_g | \Sigma_g) \left\{ \right\}
- \frac{1}{2} \sum_{i=k+1}^{n} \sum_{g=1}^{G} \left\{ z_{ig} \log |\Sigma_g| + p z_{ig} (1 - v_{ig}) \log(\rho_g) \right\} + z_{ig} \left( v_{ig} + \frac{1 - v_{ig}}{\rho_g} \right) \delta(x_i, \mu_g | \Sigma_g). 
\] (1.25)
Here $\delta(x_i, \mu_g | \Sigma_g)$ denotes the squared Mahalanobis distance between $x$ and $\mu_g$. At the E-step of the ECM algorithm, the conditional expectation of $l(\theta | x)$ is computed. This leads to the following updates for $z_{ig}$ and $v_{ig}$

$$
\hat{z}_{ig} = \frac{\pi_g[\lambda_g f_N(x_i | \mu_g, \Sigma_g) + (1 - \lambda_g) f_N(x_i | \mu_g, \rho_g \Sigma_g)]}{\sum_{g=1}^{G} \pi_g[\lambda_g f_N(x_i | \mu_g, \Sigma_g) + (1 - \lambda_g) f_N(x_i | \mu_g, \rho_g \Sigma_g)]}.
$$

(1.26)

Then, at the first CM-step of the ECM algorithm, the updates for $\lambda_g$, $\pi_g$, $\mu_g$ and $\Sigma_g$ are derived. For $\pi_g$ and $\lambda_g$ they are

$$
\hat{\pi}_g = \frac{n_g}{n}, \quad \hat{\lambda}_g = \frac{1}{n} \left( \sum_{i=1}^{k} \hat{z}_{ig} v_{ig} + \sum_{i=k+1}^{n} z_{ig} v_{ig} \right),
$$

(1.27)

where $n_g = \sum_{i=1}^{k} \hat{z}_{ig} + \sum_{i=k+1}^{n} z_{ig}$. The updates for $\mu_g$ and $\Sigma_g$ are

$$
\hat{\mu}_g = \frac{1}{s_g} \left[ \sum_{i=1}^{k} \hat{z}_{ig} \left( v_{ig} + \frac{1 - v_{ig}}{\rho_g} \right) x_i + \sum_{i=k+1}^{n} z_{ig} \left( v_{ig} + \frac{1 - v_{ig}}{\rho_g} \right) x_i \right],
$$

$$
\hat{\Sigma}_g = \frac{1}{n_g} \left\{ \sum_{i=1}^{k} \hat{z}_{ig} \left( v_{ig} + \frac{1 - v_{ig}}{\rho_g} \right) (x_i - \mu_g) (x_i - \mu_g)^\top \right. \\
+ \sum_{i=k+1}^{n} z_{ig} \left( v_{ig} + \frac{1 - v_{ig}}{\rho_g} \right) (x_i - \mu_g) (x_i - \mu_g)^\top \left\},
$$

(1.28)

where

$$
\hat{s}_g = \sum_{i=1}^{k} \hat{z}_{ig} \left( v_{ig} + \frac{1 - v_{ig}}{\rho_g} \right) + \sum_{i=k+1}^{n} z_{ig} \left( v_{ig} + \frac{1 - v_{ig}}{\rho_g} \right).
$$

(1.29)

Finally, at the second CM-step of the ECM algorithm, the update for the contamination factor $\rho_g$ is computed as

$$
\hat{\rho}_g = \frac{\sum_{i=1}^{k} \hat{z}_{ig} (1 - v_{ig}) \delta(x_i, \mu_g | \Sigma_g) + \sum_{i=k+1}^{n} z_{ig} (1 - v_{ig}) \delta(x_i, \mu_g | \Sigma_g)}{p \sum_{i=1}^{k} \hat{z}_{ig} (1 - v_{ig}) + p \sum_{i=k+1}^{n} z_{ig} (1 - v_{ig})}.
$$

(1.30)
Punzo and McNicholas (2013) also provide a discussion on the comparison between the contaminated Gaussian mixture model (1.19) and the trimmed clustering approach implemented in the R package \texttt{tclust} by Fritz et al. (2012), which is discussed in more detail in Chapter 4. The contaminated Gaussian framework introduced has several advantages over \texttt{tclust}, such as

1. giving each observation a posterior probability of belonging to a particular group, and within each group, detecting whether it is an outlier;

2. not requiring the pre-specification of the proportion of observations to be trimmed;

3. the ability to use the method for high-dimensional data; and

4. the ability to use the method in the context of model-based clustering and model-based classification.
Chapter 2

Dimension Reduction and Clustering using Shifted Asymmetric Laplace Mixtures

In this chapter we discuss a dimension reduction technique for model-based clustering using the class of multivariate shifted asymmetric Laplace distributions (SAL) proposed by Franczak et al. (2012). Our method is the SAL analogue (SALM-MDR; Morris and McNicholas, 2013) of the dimension reduction via Gaussian mixtures developed by Scrucca (2010). SAL mixtures are amongst a few methods that show promise for clustering data with asymmetric components. They can be used instead of Gaussian mixtures when performance cannot necessarily be improved by merging fitted clusters.

2.1 Methodology

We consider the mixture of SAL distributions introduced by Franczak et al. (2012)

\[
f(x|\vartheta) = \sum_{g=1}^{G} \pi_g f_{\text{SAL}}(x|\mu_g, \Sigma_g, \alpha_g),
\]

(2.1)
with component density

$$f_{\text{SAL}}(x|\alpha_g, \Sigma_g, \mu_g) = \frac{2}{(2\pi)^{p/2} |\Sigma_g|^{1/2}} \left( \frac{\delta(x, \mu_g|\Sigma_g)}{2 + \alpha_g^\top \Sigma_g^{-1} \alpha_g} \right)^{p/2} K_\nu(u), \quad (2.2)$$

where $\mu_g$ is the location, $\Sigma_g$ is a scale matrix, and $\alpha \in \mathbb{R}^p$ denotes the skewness. Here, $\delta(x, \mu_g|\Sigma_g) = (x - \mu_g)\Sigma_g^{-1}(x - \mu_g)$ denotes the squared Mahalanobis distance between $x$ and $\mu_g$, and $u = \sqrt{(2 + \alpha_g^\top \Sigma_g^{-1} \alpha_g)\delta(x, \mu_g|\Sigma_g)}$. Also, $K_\nu$ is the modified Bessel function of the third kind with index $\nu = (2 - p)/2$. Furthermore, Franczak et al. (2012) show that the random variable $X$ in (2.2) can be generated through the relationship

$$X = \mu + W\alpha + \sqrt{W}Y, \quad (2.3)$$

where $W \sim \text{Exp}(\text{rate} = 1)$ and $Y \sim \mathcal{N}(0, \Sigma)$, thus $X|W = w \sim \mathcal{N}(\mu + w\alpha, w\Sigma)$.

We focus on a clustering scenario where none of the observations have known component membership. The SAL mixture likelihood becomes

$$\mathcal{L}(\vartheta|x) = \prod_{i=1}^n \sum_{g=1}^G \pi_g f_{\text{SAL}}(x_i|\mu_g, \Sigma_g, \alpha_g). \quad (2.4)$$

For a SAL mixture model the complete-data comprise the observed $x$, the missing component memberships $z_{ig}$ (1.14) and the latent $w_{ig}$. Thus the complete-data log-likelihood can be written as

$$l(\vartheta|x) = \sum_{i=1}^n \sum_{g=1}^G z_{ig} \log(\pi_g) + \sum_{i=1}^n \sum_{g=1}^G z_{ig} \log(f_{\text{SAL}}(x_i|\mu_g, \alpha_g, \Sigma_g)), \quad (2.5)$$

and the EM algorithm is used to find the maximum likelihood estimates for (2.5), as outlined in Chapter 1.

Given a SAL mixture (2.1), we wish to find a subspace $\mathcal{S}(\beta)$ where the cluster means and cluster covariances vary the most. Although $\mu_g$ is a mean and $\Sigma_g$...
is a covariance matrix in (2.2), note that they are not the mean and covariance matrix of the random variable $X$ with the density in (2.2), except for the special case where $\alpha_g = 0$. The true mean of $X$ in (2.2) is $\tilde{\mu}_g := \mu_g + \alpha_g$, and the true covariance of $X$ in (2.2) is $\tilde{\Sigma}_g := \Sigma_g + \alpha_g\alpha_g^\top$.

**Definition 2.1.1** We define the kernel matrix $M_{\text{SAL}}$ for SAL mixtures to be

$$M_{\text{SAL}} = \sum_{g=1}^{G} \pi_g (\tilde{\mu}_g - \mu)(\tilde{\mu}_g - \mu)^\top \Sigma^{-1} \sum_{g=1}^{G} \pi_g (\tilde{\mu}_g - \mu)(\tilde{\mu}_g - \mu)^\top + \sum_{g=1}^{G} \pi_g (\tilde{\Sigma}_g - \Sigma)\Sigma^{-1}(\tilde{\Sigma}_g - \Sigma)^\top,$$

(2.6)

where $\Sigma = \frac{1}{n} \sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^\top$ denotes the overall covariance matrix and $\tilde{\Sigma} = \sum_{g=1}^{G} \pi_g \tilde{\Sigma}_g$ is the pooled within-cluster covariance matrix.

**Proposition 2.1.2** The directions where the cluster means $\tilde{\mu}_g$ and the cluster covariances $\tilde{\Sigma}_g$ vary the most are obtained from the eigen-decomposition

$$M_{\text{SAL}}v_i = l_i \Sigma v_i,$$

(2.7)

where $l_1 \geq l_2 \geq \cdots \geq l_d > 0$ and $v_i^\top \Sigma v_j = 1$ if $i = j$ and $v_i^\top \Sigma v_j = 0$ otherwise. The eigenvectors $[v_1, \ldots, v_d] \equiv \beta$, with $d \leq p$, form the basis of the dimension reduction subspace $S(\beta)$. These eigenvectors are defined as the SALMMDR directions.

**Proposition 2.1.3** Let $S(\beta)$ be the subspace spanned by the SALMMDR directions obtained from the eigen-decomposition of $M_{\text{SAL}}$ (2.7).

The projections of the parameters onto $S(\beta)$ are given by $\beta^\top \tilde{\mu}_g$ and $\beta^\top \tilde{\Sigma}_g \beta$, respectively.
ii The projections of the $n \times p$ data matrix $X$ onto the subspace $S(\beta)$ are computed from $X\beta$. These projections are defined as the SALMMDR variables.

For an $n \times p$ data matrix $X$, the kernel $M_{\text{SAL}}$ (2.6) is obtained using the estimates from the fit of a SAL mixture model on $X$, via the EM algorithm. Then the SALMMDR directions are calculated from the generalized eigen-decomposition of $M_{\text{SAL}}$ (2.6) with respect to the overall covariance matrix $\Sigma$. The SALMMDR directions are ordered based on eigenvalues, which means that directions associated with eigenvalues close to zero can be disregarded in practical applications because clusters will superimpose greatly along these directions.

As in the case of GMMDR, the estimation of the SALMMDR variables can be viewed as a form of feature extraction where the components are reduced through a set of linear combinations of the original variables. This set of features may contain estimated SALMMDR variables that provide no clustering information but require parameter estimation. Thus, the next step in the process of model-based clustering is to detect and remove these unnecessary SALMMDR variables.

Scrucca (2010) used the subset selection method of Raftery and Dean (2006) to prune the subset of GMMMDR features. We will also use this approach to select the most appropriate SALMMDR features. We compare two subsets of features, $s$ and $s' = \{s \setminus i\} \subset s$, using the BIC difference

$$
\text{BIC}_{\text{diff}}(Z_{i \in s}) = \text{BIC}_{\text{clust}}(Z_s) - \text{BIC}_{\text{not clust}}(Z_s)
$$

$$
= \underbrace{\text{BIC}_{\text{clust}}(Z_s)}_{1} - \underbrace{[\text{BIC}_{\text{clust}}(Z_{s'}) + \text{BIC}_{\text{reg}}(Z_i \mid Z_{s'})]}_{2} + \underbrace{\text{BIC}_{\text{reg}}(Z_i \mid Z_{s'})}_{3},
$$

(2.8)

where
(1) is the BIC value for the best clustering model fitted using features in \(s\),

(2) is the BIC value for the best clustering model fitted using features in \(s'\), and

(3) is the BIC value for the regression of the \(i\)-th feature on the remaining features in \(s'\).

Now, the space of all possible subsets contains \(2^d - 1\) elements so an exhaustive search is not feasible. To bypass this issue, we employ the greedy search algorithm of Scrucca (2010) to find a local optimum in the model space, which is based on the forward-backward search algorithm of Raftery and Dean (2006). The greedy search from Scrucca (2010) is a forward-only procedure; a backward step is not necessary because the SALMMDR variables are \(\Sigma\)-orthogonal. Because a backward step is not needed, computing time is decreased.

**Algorithm 2.1** Local optimum estimation (Scrucca, 2010)

1. Select the first feature to be the one which maximizes the BIC difference in (2.8) between the best clustering model and the model which assumes no clustering, i.e., a single component.

2. Select the next feature amongst those not previously included, to be the one which maximizes the BIC difference in (2.8).

3. Iterate the previous step until all the BIC differences for the inclusion of a variable become negative.

At each step, the search over the model space is performed with respect to
the model parametrization and the number of clusters. A summary of our new method of dimension reduction for model-based clustering via SAL mixtures, SALMMDR, appears below.

**Algorithm 2.2** SALMMDR estimation and feature selection

1. Fit a SAL mixture model (2.1) to the data using the EM algorithm.

2. Estimate the SALMMDR directions: identify directions where the cluster means and cluster variances vary the most, provided each direction is $\Sigma$-orthogonal to the others. This is done through the eigen-decomposition of the kernel matrix $M_{\text{SAL}}$ (2.6).

3. Select the SALMMDR variables: compute the set of features by projecting the data onto the estimated subspace and use the greedy search algorithm to discard the ones which provide no clustering information.

4. Fit a SAL mixture model (2.1) on the selected SALMMDR variables and return to step 2.

5. Repeat steps 2–4 until none of the remaining features can be dropped.

**2.2 Applications**

**2.2.1 Simulated data**

We employed a data simulation scheme to test the SALMMDR algorithm and compare its performance to that of the GMMDR procedure. Two scenarios were
considered. In Scenario I we generated two variables from a multivariate normal distribution with means $\mu_1 = (0, -2)$ and $\mu_2 = (0, 5)$, and covariance matrices

$$
\Sigma_1 = \begin{bmatrix}
1 & 0.5 \\
0.5 & 1
\end{bmatrix} \quad \text{and} \quad \Sigma_2 = \begin{bmatrix}
1 & 0 \\
0 & 1
\end{bmatrix}.
$$

In Scenario II, we modified Scenario I by adding three noise variables generated from independent standard normal distributions. Figure 2.1 and Figure 2.2 depict scatterplots of the simulated data and its underlying cluster structure for Scenario I and Scenario II respectively.

Figure 2.1: Scatterplot of simulated data for Scenario I indicating the cluster structure.

To ascertain the performance of both clustering methods under varying data dimensions, each scenario was run 25 times for three data sets, consisting of 100, 500, and 1000 data points, respectively, generated according to the schemes described earlier. We evaluated the clustering results by computing the adjusted Rand index (ARI;
Figure 2.2: Scatterplot of simulated data for Scenario II indicating the cluster structure.

Hubert and Arabie, 1985) for each simulation; higher values of ARI correspond to better performance, with the value 1 reflecting perfect class agreement. The simulation results (Table 2.1) show that the SALDMMDR approach gave ARI values similar to those from GMMDR.

<table>
<thead>
<tr>
<th></th>
<th>Scenario I</th>
<th></th>
<th>Scenario II</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n = 100</td>
<td>n = 500</td>
<td>n = 1000</td>
<td>n = 100</td>
</tr>
<tr>
<td>GMMDR</td>
<td>0.9606</td>
<td>1</td>
<td>0.9996</td>
<td>0.9853</td>
</tr>
<tr>
<td></td>
<td>(0.0838)</td>
<td>(0)</td>
<td>(0.0013)</td>
<td>(0.0347)</td>
</tr>
<tr>
<td>SALMMMDR</td>
<td>1</td>
<td>0.9984</td>
<td>0.9988</td>
<td>0.9793</td>
</tr>
<tr>
<td></td>
<td>(0)</td>
<td>(0.0109)</td>
<td>(0.0034)</td>
<td>(0.0427)</td>
</tr>
</tbody>
</table>

2.2.2 Real data

In order to benchmark the performance of our algorithm for real data we compare the results with other known clustering approaches outlined below.

2. Mixtures of Gaussian distributions (Browne and McNicholas, 2013a) with the R package `mixture`: an implementation of all 14 Gaussian parsimonious clustering models (GPCMs) for model-based clustering.

3. Parsimonious Gaussian mixture models (McNicholas and Murphy, 2008): Gaussian mixtures of factor analyzers are used to parametrize the covariance structure, implemented with the R package `pgmm` (McNicholas et al., 2011).


5. For non model-based methods of comparison, we selected $k$-means and partitioning around medoids (PAM), computed with the R functions `kmeans` and `pam`, respectively.

For the real data analyses herein, we `scale()` data prior to analysis and use the hierarchical agglomerative clustering procedure for initialization (available from `mclust`).

**Swiss banknote data**

Flury and Riedwyl (1988) presented six measurements (Table 2.2) taken from Swiss banknotes, available through the R package `gclus` (Hurley, 2010).
Our goal is to cluster the 100 genuine and 100 counterfeit notes. Table 2.3 gives the classifications for SALMMDR and GMMDR. While SALMMDR performs better (ARI = 0.98) than GMMDR (ARI = 0.67), it is clear that one possible merging of the Gaussian components would result in similar classification performance. The comparison with the other methods in Table 3.6 shows that $k$-means performs the best on these data (ARI = 1), with PAM matching the results of SALMMDR.

Table 2.3: Model-based clustering results for the best SALMMDR and GMMDR models, in terms of BIC, fitted to the Swiss banknote data.

<table>
<thead>
<tr>
<th></th>
<th>SALMMDR</th>
<th>GMMDR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(4 features)</td>
<td>(3 features)</td>
</tr>
<tr>
<td>Genuine</td>
<td>99 0</td>
<td>25 74 1 0</td>
</tr>
<tr>
<td>Counterfeit</td>
<td>1 100</td>
<td>0 0 15 85</td>
</tr>
</tbody>
</table>

Table 2.4: Summary of results for the best models fitted to the Swiss banknote data.

<table>
<thead>
<tr>
<th>Method</th>
<th>ARI</th>
<th>Comp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SALMMDR</td>
<td>0.98</td>
<td>2</td>
</tr>
<tr>
<td>GMMMDR</td>
<td>0.67</td>
<td>4</td>
</tr>
<tr>
<td>mixture</td>
<td>0.68</td>
<td>4</td>
</tr>
<tr>
<td>pgmm</td>
<td>0.85</td>
<td>3</td>
</tr>
<tr>
<td>teigen</td>
<td>0.68</td>
<td>4</td>
</tr>
<tr>
<td>kmeans</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>pam</td>
<td>0.98</td>
<td>2</td>
</tr>
</tbody>
</table>

Scatterplots of the estimated SALMMDR directions (Figure 2.3) show that the genuine and counterfeit banknotes are clearly distinguished by the first two direc-
tions. The blue and red shaded areas on the right side plot in Figure 2.3 represent the MAP (1.16) classification regions corresponding to the first two estimated directions. Here, $\hat{z}_{ig}$ represents the estimated cluster membership of an observation. Note that the plots on the right side of Figures 2.3 and 2.4 are based on plots produced by the R package mclust (Fraley et al., 2012b).

Figure 2.3: Scatterplots illustrating the SALMMDR directions for the Swiss banknote data; the numbers (left-hand plot) or shapes (right-hand plot) signify true classification and the colour indicates estimated SALMMDR cluster allocation.

Yeast data

Horton and Nakai (1996) provided data on cellular localization of yeast proteins. For our analysis, we used the set available from the UCI Machine Learning Repository, with variables as indicated in Table 2.5.

We considered the three variables analyzed by Franczak et al. (2012): Mc-Geochs method for signal sequence recognition (mcg), the score of the ALOM mem-
Table 2.5: Measurements for the yeast data.

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>McGeochs method for signal sequence recognition (mcg)</td>
<td></td>
</tr>
<tr>
<td>Score of ALOM membrane spanning region prediction program (alm)</td>
<td></td>
</tr>
<tr>
<td>Score of discriminant analysis of amino acid content of vacuolar and extracellular proteins (vac)</td>
<td></td>
</tr>
<tr>
<td>Score of discriminant analysis of amino acid content of N-terminal region of mitochondrial and non-mitochondrial proteins (mit)</td>
<td></td>
</tr>
<tr>
<td>Score of discriminant analysis of nuclear localization signals of nuclear and non-nuclear proteins (nuc)</td>
<td></td>
</tr>
<tr>
<td>Presence of HDEL substring (erl)</td>
<td></td>
</tr>
<tr>
<td>Peroxiomial targeting signal in C-terminus (pox)</td>
<td></td>
</tr>
<tr>
<td>von Heijne’s method for signal sequence recognition (gvh)</td>
<td></td>
</tr>
</tbody>
</table>

brane spanning region prediction program (alm), and the score of discriminant analysis of the amino acid content of vacuolar and extracellular proteins (vac). We chose these three variables for two reasons. First, to enable comparisons with the analysis of Franczak et al. (2012). Second, they seem to be the most reliable (e.g. pox variable consists of mostly zero values, nuc variable exhibits mostly values of 0.22) and show potential for clustering. The goal of our cluster analysis is to distinguish between the two localization sites, CYT (cytosolic or cytoskeletal) and ME3 (membrane protein, no N-terminal signal), which total 626 observations. Table 2.6 details the classifications obtained with SALMMDR and GMMDR. Although the performance of the GMMDR (ARI = 0.45) approach can be improved by merging the components, even the most generous merging results in an inferior performance to the SALMMDR (ARI = 0.825) approach. The summary of the results obtained with the comparator methods, Table 2.7, indicates that SALMMDR outperforms them all, with PAM (ARI = 0.77) offering the next best results.
Table 2.6: Model-based clustering results for the best SALMMDR and GMMDR models, in terms of BIC, fitted to the yeast data.

<table>
<thead>
<tr>
<th></th>
<th>SALMMDR</th>
<th>GMMDR (2 features)</th>
<th>GMMDR (merged)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(2 features)</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>CYT</td>
<td>453</td>
<td>17</td>
<td>25</td>
</tr>
<tr>
<td>ME3</td>
<td>10</td>
<td>146</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 2.7: Summary of results for the best models fitted to the yeast data.

<table>
<thead>
<tr>
<th>Method</th>
<th>ARI</th>
<th>Comp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SALMMDR</td>
<td>0.825</td>
<td>2</td>
</tr>
<tr>
<td>GMMDR</td>
<td>0.45</td>
<td>4</td>
</tr>
<tr>
<td>mixture</td>
<td>0.44</td>
<td>4</td>
</tr>
<tr>
<td>pgmm</td>
<td>0.43</td>
<td>4</td>
</tr>
<tr>
<td>teigen</td>
<td>0.42</td>
<td>4</td>
</tr>
<tr>
<td>kmeans</td>
<td>0.43</td>
<td>2</td>
</tr>
<tr>
<td>pam</td>
<td>0.77</td>
<td>2</td>
</tr>
</tbody>
</table>

Scatterplots of the estimated SALMMDR directions (Figure 2.4) show that the procedure gives good separation in the two localization sites. The right side plot depicts the uncertainty, $u_i = 1 - \max_g(\hat{z}_{ig})$, projected onto the first and third estimated SALMMDR directions, where $\hat{z}_{ig}$ is the expected value of $z_{ig}$ based on the parameter estimates; the misclassified observations lie in regions of higher uncertainty, which are represented by darker shaded areas.
The numbers (left-hand plot) or shapes (right-hand plot) signify true classification and colour indicates the estimated SALMMDR cluster allocation.

### Australian Institute of Sport data

Cook and Weisberg (1994) provided eleven body and blood measurements collected from athletes at the Australian Institute of Sport (AIS), courtesy of Richard Telford and Ross Cunningham.

<table>
<thead>
<tr>
<th>Sex</th>
<th>Sport</th>
<th>Red cell count</th>
<th>White cell count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hematocrit</td>
<td>Hemoglobin</td>
<td>Plasma ferritin</td>
<td>Body mass index</td>
</tr>
<tr>
<td>Sum of skin folds</td>
<td>Body fat percent</td>
<td>Lean body mass</td>
<td>Height</td>
</tr>
<tr>
<td>Weight</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Our aim here is to cluster the 102 male and 100 female athletes. The classifications obtained with SALMMDR and GMMMDR appear in Table 2.9. While SALMMDR (ARI = 0.94) outperformed GMMMDR (ARI = 0.49), Table 2.10 shows that it also did much better than all of the comparator methods, with k-means (ARI = 0.85) as the next best clustering option.
Table 2.9: Model-based clustering results for the best SALMMDR and GMMDR models, in terms of BIC, fitted to the Australian Institute of Sport data.

<table>
<thead>
<tr>
<th></th>
<th>SALMMDR</th>
<th>GMMDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>(7 features)</td>
<td>1 2 1 2</td>
<td>3 4 4</td>
</tr>
<tr>
<td>Female</td>
<td>100 3 15</td>
<td>51 34 0</td>
</tr>
<tr>
<td>Male</td>
<td>0 99 16</td>
<td>16 0 5</td>
</tr>
</tbody>
</table>

Table 2.10: Summary of results for the best models fitted to the Australian Institute of Sport data.

<table>
<thead>
<tr>
<th>Method</th>
<th>ARI</th>
<th>Comp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SALMMDR</td>
<td>0.94</td>
<td>2</td>
</tr>
<tr>
<td>GMMDR</td>
<td>0.49</td>
<td>4</td>
</tr>
<tr>
<td>mixture</td>
<td>0.65</td>
<td>4</td>
</tr>
<tr>
<td>pgmm</td>
<td>0.49</td>
<td>4</td>
</tr>
<tr>
<td>teigen</td>
<td>0.52</td>
<td>3</td>
</tr>
<tr>
<td>kmeans</td>
<td>0.85</td>
<td>2</td>
</tr>
<tr>
<td>pam</td>
<td>0.83</td>
<td>2</td>
</tr>
</tbody>
</table>

2.3 Summary

This chapter introduced a dimension reduction approach for model-based clustering via SAL mixtures. Our method, known as SALMMDR, is a SAL analogue of the Gaussian procedure called GMMDR (Scrucca, 2010). The SALMMDR approach was illustrated using simulated and real data, where it outperformed its Gaussian analogue. The procedure was also benchmarked against other clustering methods, namely GPCM, PGMM, tEIGEN, k-means clustering, and k-medoids clustering. With the exception of the Swiss banknotes data, for which k-means produced a perfect classification, SALMMDR gave better clustering results than the rest of
the comparator methods. While merging GMMDR components improved its performance, they still did not match that of SALMMDR. Note that the term “analogue” is used a little loosely herein because we do not consider decomposed covariance structures for the SAL mixture models. Future work could focus on dimension reduction using a SAL analogue of the GPCM models.
Chapter 3

Dimension Reduction and Clustering using

Generalized Hyperbolic Mixtures

In this chapter we discuss a dimension reduction technique for model-based clustering using the class of multivariate generalized hyperbolic distributions (GH) proposed by Browne and McNicholas (2013b). The method introduced here is the GH analogue of the dimension reduction via Gaussian mixtures developed by Scrucca (2010).

Generalized hyperbolic distributions were introduced by Barndorff-Nielsen (1977) and used to model eolian sand deposits, i.e., sand deposits arising from the action of wind. The name of the distribution was derived from the fact that its log-density has the shape of a hyperbola. Properties of generalized hyperbolic densities were discussed in Barndorff-Nielsen and Halgreen (1977) and Blæsild (1978) and, more recently, mixtures of these distributions appear in McNeil et al. (2005) and Härdle and Simar (2011). Generalized hyperbolic distributions can effectively model extreme values, making them very useful in the context of financial and risk management applications, where the normal distribution does not offer a good description of reality. The multivariate generalized hyperbolic family is extremely flexible and
contains many special and limiting cases, such as the inverse Gaussian, Laplace, and skew-$t$ distributions.

### 3.1 Methodology

Browne and McNicholas (2013b) proposed a multivariate generalized hyperbolic mixture model,

$$ f(x|\theta) = \sum_{g=1}^{G} \pi_g f_h(x|\lambda_g, \omega_g, \mu_g, \Sigma_g, \alpha_g), \quad (3.1) $$

where $\pi_g$ are the mixing proportions and the $g$th component density is

$$ f_h(x|\lambda_g, \omega_g, \mu_g, \Sigma_g, \alpha_g) = \frac{K_{\lambda_g-p/2} \left( \sqrt{[\omega_g + \alpha_g^\top \Sigma_g^{-1} \alpha_g]}(\omega_g + \delta(x, \mu_g|\Sigma_g)] \right)}{(2\pi)^{p/2}|\Sigma_g|^{1/2}} K_{\lambda_g}(\omega_g) \exp(-\mu_g^\top \Sigma_g^{-1} \alpha_g), \quad (3.2) $$

with index parameter $\lambda_g$, concentration parameter $\omega_g$, skewness parameter $\alpha_g$, location $\mu_g$, and scale matrix $\Sigma_g$. Here, $\delta(x, \mu_g|\Sigma_g) = (x - \mu_g)^\top \Sigma_g^{-1}(x - \mu_g)$ is the squared Mahalanobis distance between $x$ and $\mu_g$, with $K_{\lambda_g}$ and $K_{\lambda_g-p/2}$ denoting modified Bessel functions of the third kind.

The evaluation of modified Bessel functions in the density (3.2) sometimes leads to numerical overflow or underflow. To avoid these issues, we use asymptotic expansions from Abramowitz and Stegun (1972), i.e., for large $x$ or $\lambda$,

$$ K_\lambda(\lambda x) = \sqrt{\frac{\pi}{2\lambda}} \exp^{-\lambda x} \left[ 1 + \sum_{k=1}^{\infty} (-1)^k \frac{\tau^k}{\lambda^k} \right], $$
where
\[ \rho = \sqrt{1 + x^2} + \ln \left( \frac{x}{1 + \sqrt{1 + x^2}} \right). \]

Here, \( u_k(\tau) \) is the Debye polynomial represented by \( u_0(\tau) = 1 \) and
\[ u_{k+1}(\tau) = \frac{1}{2} \tau^2 (1 - \tau^2) u'_k(\tau) + \frac{1}{8} \int_0^\tau (1 - 5s^2) u_k(s) ds, \]
for \( k = 1, 2, \ldots \) and \( \tau = 1/\sqrt{1 + x^2} \).

The parametrization in (3.2) is one of several available for generalized hyperbolic distributions (cf. McNeil et al., 2005). In this case, the \( p \)-dimensional random vector \( X \) is generated by combining a generalized inverse Gaussian (GIG) random variable \( Y \) with a latent multivariate Gaussian random variable \( U \sim \mathcal{N}(0, \Sigma) \). Note that the density of \( Y \sim \text{GIG}(\omega, \eta, \lambda) \) is
\[ h(y|\omega, \eta, \lambda) = \frac{(y/\eta)^{\lambda-1}}{2\eta K_\lambda(\omega)} \exp \left\{ -\frac{\omega}{2} \left( \frac{y}{\eta} + \frac{\eta}{y} \right) \right\}. \quad (3.3) \]
We fix \( \eta = 1 \) and use the relationship \( X = \mu + Y \alpha + \sqrt{Y} U \). Full details on the derivation of this parametrization and its use in parameter estimation are given by Browne and McNicholas (2013b).

In the following sections we discuss using the generalized hyperbolic distribution for model-based methods in the context of unsupervised (clustering), semi-supervised (classification) and supervised (discriminant analysis) learning. Figure 3.1 shows the relationship between these learning approaches.
Model-based clustering

Consider a clustering scenario in which none of the observations have known component membership. The generalized hyperbolic model-based clustering likelihood is

\[ L(\vartheta|\mathbf{x}) = \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g f_h(x_i|\lambda_g, \omega_g, \mu_g, \Sigma_g, \alpha_g). \]  

(3.4)

To facilitate discussion of parameter estimation, introduce \( z_{ig} \) to denote component membership, so that \( z_{ig} = 1 \) if observation \( \mathbf{x}_i \) belongs to component \( g \) and \( z_{ig} = 0 \) otherwise.

Parameter estimation for generalized hyperbolic mixtures is carried out using the EM algorithm, based on the complete-data log-likelihood (3.5), where the complete-data comprise the observed \( \mathbf{x}_i \), the missing \( z_{ig} \), and the latent \( y_{ig} \). Our
complete-data log-likelihood is given by

$$l(\vartheta) = \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \left[ \log(\pi_{g}) + \sum_{j=1}^{p} \log[\phi(x_i|\mu_g + y_{ig}\alpha_g, y_{ig}\Sigma_g)] + \log[h(y_{ig}|\omega_g, \lambda_g)] \right],$$

(3.5)

where $X_i|y_{ig} \sim N(\mu_g + y_{ig}\alpha_g, y_{ig}\Sigma_g)$, $Y_{ig} \sim GIG(\omega_g, 1, \lambda_g)$. Extensive details of the EM algorithm for generalized hyperbolic mixtures are given by Browne and McNicholas (2013b). Convergence is determined via an asymptotic estimate of the log-likelihood (3.5) using the Aitken acceleration and (1.15). After convergence, component memberships are usually estimated based on MAP classification (1.16).

In our applications (Section 3.2), we assume that the number of components $G$ is unknown. This is not unusual in real model-based clustering applications, where the BIC (1.17) is often used to determine $G$.

**Model-based classification and discriminant analysis**

Model-based classification, or partial classification (cf. McLachlan, 1982), is a semi-supervised analogue of model-based clustering that has historically received much less attention within the literature. However, model-based classification has garnered increased attention over the past few years and some authors (e.g., Dean et al., 2006; McNicholas, 2010) have demonstrated that model-based classification can give excellent performance in real applications. Model-based discriminant analysis (Hastie and Tibshirani, 1996) is a supervised analogue of model-based clustering that has similarly received much less attention until recently (e.g., Andrews and McNicholas, 2011, 2012a). Model-based classification and discriminant analysis are
best explained through the associated likelihoods.

Consider the classification scenario where there are $n$ observations, $k$ of which have known group memberships. Under a model-based classification framework, all $n$ observations are used to estimate the group memberships for the $n - k$ observations with unknown group memberships. Following McNicholas (2010), we arrange the data so that the first $k$ observations have known group memberships; therefore, the likelihood is given by

$$
L(\theta | x) = \prod_{i=1}^{k} \prod_{g=1}^{G} [\pi_g f_h(x_i | \lambda_g, \omega_g, \mu_g, \Sigma_g, \alpha_g)]^{z_{ig}} \prod_{j=k+1}^{n} \sum_{s=1}^{G} \pi_s f_h(x_j | \lambda_s, \omega_s, \mu_s, \Sigma_s, \alpha_s).
$$

(3.6)

As in the case of model-based clustering, parameter estimation is carried out using the EM algorithm. From (3.4) and (3.6), we see that model-based clustering can be viewed as a special case of model-based classification that arises by considering (3.6) with $k = 0$.

For model-based discriminant analysis, we again have $n$ observations, $k$ of which have known group memberships. Again, we arrange the data so that the first $k$ observations have known group memberships; however, instead of using all $n$ observations to estimate the unknown group memberships, we only use the first $k$ observations (i.e., the labelled observations). First, we form the likelihood

$$
L(\theta | x) = \prod_{i=1}^{k} \prod_{g=1}^{G} [\pi_g f_h(x_i | \lambda_g, \omega_g, \mu_g, \Sigma_g, \alpha_g)]^{z_{ig}}.
$$

(3.7)

Then, the parameter estimates are computed via the EM algorithm and the a posteriori expected values of the $z_{ig}$ are used to estimate the group memberships of the remaining $n - k$ observations.
Fraley and Raftery (2002) discuss MCLUST DA, where they allow for multiple components per known group by using the BIC to choose the number of components as well as the best MCLUST model (i.e., covariance structure) for each group. Scrucca (2013) extended dimension reduction for Gaussian mixtures to this MCLUST DA framework. However, for the discriminant analyses herein, we restrict HMMDR to one component per known group. In part, this is done because of the large number of parameters to be estimated and the relatively small number of observations in the real data sets we consider. However, it is also done because we believe that the flexibility inherent in generalized hyperbolic components makes it far less likely, relative to Gaussian components, that multiple components would be needed to model a class. This latter point will be investigated as part of future work.

As discussed in Chapter 1 and 2, two other non-Gaussian extensions of GMMDR exist. The approach introduced herein aims to combine the robustness offered by the tMMDR algorithm with the elegance and asymmetry afforded by the SALMMDR procedure.

The dimension reduction approach of Scrucca (2010) is extended through development of a generalized hyperbolic analogue. We will also develop methods for model-based classification and discriminant analysis for GMMMDR and all of its non-Gaussian analogues considered herein. As mentioned earlier, our GMMMDR DA extension differs from that of Scrucca (2013) because we only consider one mixture component per class.

Given a generalized hyperbolic mixture (3.2), we wish to find a subspace $S(\beta)$ where the cluster means and cluster covariances vary the most. Although $\mu_g$ is
a mean and $\Sigma_g$ is a covariance matrix in (3.2), note that they are not the mean and covariance matrix of the random variable $X$ with the density in (3.2), except for the special case where $\alpha_g = 0$. The true mean of $X$ in (3.2) is $\tilde{\mu}_g := \mu_g + \alpha_g$, and the true covariance of $X$ in (3.2) is $\tilde{\Sigma}_g := \Sigma_g + \alpha_g\alpha_g^\top$.

**Definition 3.1.1** We define the kernel matrix $M_{GH}$ for GH mixtures to be

$$M_{GH} = \sum_{g=1}^{G} \pi_g (\tilde{\mu}_g - \mu)(\tilde{\mu}_g - \mu)^\top \Sigma^{-1} \sum_{g=1}^{G} \pi_g (\tilde{\mu}_g - \mu)(\tilde{\mu}_g - \mu)^\top + \sum_{g=1}^{G} \pi_g (\tilde{\Sigma}_g - \bar{\Sigma})\Sigma^{-1}(\tilde{\Sigma}_g - \bar{\Sigma})^\top,$$

(3.8)

where $\Sigma = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^\top$ denotes the overall covariance matrix and $\bar{\Sigma} = \sum_{g=1}^{G} \pi_g \tilde{\Sigma}_g$ is the pooled within-cluster covariance matrix.

**Proposition 3.1.2** The directions where the cluster means $\tilde{\mu}_g$ and the cluster covariances $\tilde{\Sigma}_g$ vary the most are obtained from the eigen-decomposition

$$M_{GH} v_i = l_i \Sigma v_i,$$

(3.9)

where $l_1 \geq l_2 \geq \cdots \geq l_d > 0$ and $v_i^\top \Sigma v_j = 1$ if $i = j$ and $v_i^\top \Sigma v_j = 0$ otherwise.

The eigenvectors $[v_1, \ldots, v_d] \equiv \beta$, with $d \leq p$, form the basis of the dimension reduction subspace $S(\beta)$. These eigenvectors are defined as the HMMDR directions.

**Proposition 3.1.3** Let $S(\beta)$ be the subspace spanned by the HMMDR directions obtained from the eigen-decomposition of $M_{GH}$ (3.9).

i The projections of the parameters onto $S(\beta)$ are given by $\beta^\top \tilde{\mu}_g$ and $\beta^\top \tilde{\Sigma}_g \beta$, respectively.
ii The projections of the \( n \times p \) data matrix \( X \) onto the subspace \( S(\beta) \) are computed from \( X \beta \). These projections are defined as the HMMDR variables.

For an \( n \times p \) data matrix \( X \), the kernel \( M_{GH} \) (3.8) is obtained using the estimates from the fit of an HMM on \( X \), via the EM algorithm. Then the HMMDR directions are calculated from the generalized eigen-decomposition of \( M_{GH} \) (3.8) with respect to the overall covariance matrix \( \Sigma \). The HMMDR directions are ordered based on eigenvalues, which means that directions associated with eigenvalues close to zero can be disregarded in practical applications because clusters will superimpose greatly along these directions.

Similar to GMMDR, the estimation of the HMMDR variables can be interpreted as feature selection, where the members are reduced through a set of linear combinations of the original variables. It is possible that this set of features contains estimated HMMDR variables that do not offer any clustering information but require parameter estimation. Scrucca (2010) uses the selection method of Raftery and Dean (2006) to prune the subset of GMMDR features. We follow this approach (2.8) described in Chapter 2.

Because the space of all possible subsets contains \( 2^d - 1 \) elements, where \( d \leq p \), a full feature search is not usually feasible. To this end, we employ the forward greedy search algorithm of Scrucca (2010) (2.1) to find a local optimum in the model space. At each stage, the search over the model space is performed with respect to the model parameterization and the number of clusters. This algorithm can be applied to the three frameworks under consideration: model-based clustering, classification, and
discriminant analysis, by modifying the likelihood functions (3.4), (3.6), and (3.7) in the EM procedure accordingly. Similar to the SALMMDR algorithm (2.2), our new method, called HMMDR can be summarized as follows.

**Algorithm 3.1** HMMDR estimation and feature selection

1. Fit a GH mixture (3.1) to the data using the EM algorithm.

2. Estimate the HMMDR directions: identify directions where the cluster means and cluster variances vary the most, provided each direction is $\Sigma$-orthogonal to the others. This is done through the eigen-decomposition of the kernel matrix $M_{GH}$ in (3.8).

3. Select the HMMDR variables: compute the set of features by projecting the data onto the estimated subspace and use the greedy search algorithm to discard the ones that provide no clustering information.

4. Fit a GH mixture (3.1) to the selected HMMDR variables and return to step 2.

5. Repeat steps 2–4 until none of the remaining features can be discarded.

### 3.2 Applications

#### 3.2.1 Simulated data

First, we employ a data simulation scheme based on two scenarios to test the HMMDR algorithm. In Scenario I, we generate three variables from a multivariate normal distribution with means $\mu_1 = (0, -2, 0)$, $\mu_2 = (2, 4, 0)$, $\mu_3 = (-2, -4, 2)$
and common covariance matrix $\Sigma = \text{diag}(0.5)$. Also, we use three different sample sizes, $n \in \{100, 500, 1000\}$. In Scenario II, we modify Scenario I by adding five noise variables generated from standard normal distributions. Figures 3.2 and 3.3 illustrate a typical example of the three components.

![Pairs plot illustrating a typical example of the simulated multivariate Gaussian components in Scenarios I.](image)

Figure 3.2: Pairs plot illustrating a typical example of the simulated multivariate Gaussian components in Scenarios I.

We run each scenario 25 times for each possible combination of data dimension and model framework. For model-based classification and discriminant analysis we assumed that each observation had a 50% probability of being known. This resulted in the number of known observations $k$ being close to $n/2$ but varying slightly from run to run (more details on the selection of known observations appear in the next section). The ARI for both model-based classification and discriminant analysis was computed based only on the $n - k$ unknown observations. The results are given
in Tables 3.1 and 3.2. HMMDR generally exhibits high ARI values across both scenarios and all sample sizes for model-based clustering, classification, and discriminant analysis. We notice a slight drop in classification performance for Scenario II with 100 observations; however, performance on noisy data improves for larger $n$. One or two features are selected throughout Scenario I, and as expected, the addition of noise tends to lead to a slight increase in the number of features for Scenario II.
Table 3.1: Summary of results for the HMMDR and GMMMDR models fitted to the simulated data in Scenario I without noise, based on 300 runs.

<table>
<thead>
<tr>
<th>Scenario I (no noise)</th>
<th>HMMDR</th>
<th>GMMMDR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n = 100$</td>
<td>$n = 500$</td>
</tr>
<tr>
<td>Clust. Avg. ARI</td>
<td>0.9701</td>
<td>0.9634</td>
</tr>
<tr>
<td>Std. dev.</td>
<td>0.0349</td>
<td>0.0173</td>
</tr>
<tr>
<td>Features</td>
<td>1–2</td>
<td>1–2</td>
</tr>
<tr>
<td>Avg. feat.</td>
<td>1.137</td>
<td>1.003</td>
</tr>
<tr>
<td>Class. Avg. ARI</td>
<td>0.9877</td>
<td>0.9829</td>
</tr>
<tr>
<td>Std. dev.</td>
<td>0.0223</td>
<td>0.0123</td>
</tr>
<tr>
<td>Features</td>
<td>1–2</td>
<td>1–2</td>
</tr>
<tr>
<td>Avg. feat.</td>
<td>1.07</td>
<td>1.033</td>
</tr>
<tr>
<td>DA Avg.ARI</td>
<td>0.9274</td>
<td>0.9815</td>
</tr>
<tr>
<td>Std. dev.</td>
<td>0.1031</td>
<td>0.0181</td>
</tr>
<tr>
<td>Features</td>
<td>1–2</td>
<td>1–2</td>
</tr>
<tr>
<td>Avg. feat.</td>
<td>1.303</td>
<td>1.027</td>
</tr>
</tbody>
</table>

Table 3.2: Summary of results for the HMMMDR and GMMMDR models fitted to the simulated data in Scenario I with noise, based on 300 runs.

<table>
<thead>
<tr>
<th>Scenario II (noise)</th>
<th>HMMDR</th>
<th>GMMMDR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n = 100$</td>
<td>$n = 500$</td>
</tr>
<tr>
<td>Clust. Avg. ARI</td>
<td>0.9239</td>
<td>0.9810</td>
</tr>
<tr>
<td>Std. dev.</td>
<td>0.0692</td>
<td>0.0120</td>
</tr>
<tr>
<td>Features</td>
<td>1–4</td>
<td>1–3</td>
</tr>
<tr>
<td>Avg. feat.</td>
<td>2.609</td>
<td>2.287</td>
</tr>
<tr>
<td>Class. Avg. ARI</td>
<td>0.9507</td>
<td>0.9296</td>
</tr>
<tr>
<td>Std. dev.</td>
<td>0.0533</td>
<td>0.0129</td>
</tr>
<tr>
<td>Features</td>
<td>1–4</td>
<td>1–3</td>
</tr>
<tr>
<td>Avg. feat.</td>
<td>1.07</td>
<td>1.65</td>
</tr>
<tr>
<td>DA Avg.ARI</td>
<td>0.7299</td>
<td>0.9662</td>
</tr>
<tr>
<td>Std. dev.</td>
<td>0.1280</td>
<td>0.0247</td>
</tr>
<tr>
<td>Features</td>
<td>1–3</td>
<td>1–3</td>
</tr>
<tr>
<td>Avg. feat.</td>
<td>1.557</td>
<td>1.97</td>
</tr>
</tbody>
</table>

Next we compare the performance of HMMMDR against the existing GMMMDR approach for high dimensional data. We generated three-component data sets with $n_g = 40$ observations per component from generalized hyperbolic distributions with random covariance matrices (produced using the R package clusterGeneration (Qiu and Joe, 2006) and $\alpha = -1$. The means were drawn from a standard multivariate
normal and multiplied by a small integer. A typical sample is illustrated in Figure 3.4.

Figure 3.4: Pairs plot illustrating a typical example of the simulated multivariate hyperbolic components in $p = 20$ dimensions.

These simulations were performed with $p \in \{20, 30, 40, 50\}$. Some of the runs resulted in error due to lack of convergence, combining of groups, or other numerical errors. Table 3.3 gives the averages over the successful runs. We note that our implementation of HMMDR is slightly more robust than our version of GMMDR, with more successful runs except when $p > n_g$. We note numerical difficulties for $p > n_g$ as the HMMDR covariance matrices will generally be numerically singular; this problem is mitigated somewhat in GMMDR since various restricted covariance decomposition structures are used while fitting. HMMDR performs reasonably well, when successful, for high dimensional data ($p = 40, 50$) but requires many more features than GMMDR. We anticipate the performance of HMMDR would improve with a more complex approach to covariance estimation, taking into account restricted
covariance structures and other methods to avoid numerical singularities.

Table 3.3: Summary of clustering results for the best HMMDR and GMMDR models for the simulated high-dimensional data.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>HMMDR</td>
<td>20</td>
<td>30</td>
<td>30</td>
<td>0.43</td>
<td>0.41</td>
<td>9.8</td>
<td>2.93</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>30</td>
<td>29</td>
<td>0.72</td>
<td>0.88</td>
<td>19</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>50</td>
<td>44</td>
<td>0.94</td>
<td>0.96</td>
<td>25.1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>60</td>
<td>31</td>
<td>0.97</td>
<td>1</td>
<td>29.5</td>
<td>3</td>
</tr>
<tr>
<td>GMMDR</td>
<td>20</td>
<td>30</td>
<td>26</td>
<td>0.88</td>
<td>0.91</td>
<td>4.7</td>
<td>3.2</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>30</td>
<td>26</td>
<td>0.96</td>
<td>1</td>
<td>3.2</td>
<td>3.1</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>50</td>
<td>43</td>
<td>0.98</td>
<td>1</td>
<td>3.8</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>60</td>
<td>50</td>
<td>0.99</td>
<td>1</td>
<td>3.7</td>
<td>3</td>
</tr>
</tbody>
</table>

3.2.2 Real data

To gauge the performance of our algorithm on real data, we compare results with the eight methods outlined below. Except for k-means, we choose these particular comparator methods because they provide model-based analyses while implicitly reducing the dimension space.

1. Robust principal component analysis (ROBPCA) (Hubert et al., 2005) paired with t-mixtures via the tEIGEN family: principal components analysis resistant to outliers, with robust loadings computed by using projection-pursuit techniques and the minimum covariance determinant method. We use the R package rrcov (Todorov and Filzmoser, 2009) for the ROBPCA computations as well as the teigen package (Andrews and McNicholas, 2012b).

2. Parsimonious Gaussian mixture models (McNicholas and Murphy, 2008): Gaussian mixtures of factor analyzers are used to parameterize the covariance struc-
ture. The R package pgmm (McNicholas et al., 2011) is used to derive the results.

3. Mixtures of common factor analyzers (Baek et al., 2010): model-based density estimation, where common component-factor loadings are used to reduce the number of parameters. The R package mcfa (Baek et al., 2009) is used herein.

4. FisherEM (Bouveyron and Brunet, 2012): a subspace clustering method based on Gaussian mixtures, where an EM-like algorithm estimates both the discriminative subspace and the parameters of the model. The R package FisherEM (Bouveyron and Brunet, 2012) is employed to carry out the computations.

5. Clustvarsel (Scrucca et al., 2013) paired with Gaussian mixtures via the MCLUST family: a greedy procedure to find the (locally) optimal subset of variables in a data set. The R package clustvarsel is used to derive the results.

6. k-means: partitions $n$ observations into $k$ clusters, where each observation belongs to the cluster with the nearest mean. The R function kmeans is used herein.

7. GMMDR: the approach of Scrucca (2010) based on Gaussian mixtures. While the dimension reduction procedure of GMMDR is available in the R package mclust (Fraley et al., 2012a), the subset selection procedure is not currently available.

8. tMMDR (Morris et al., 2013): the $t$-analogue of GMMDR. Fitting of the $t$-mixtures was carried out with the R package teigen.

9. SALMMDR (Morris and McNicholas, 2013): the SAL analogue of GMMDR.
For the analyses in this section, we fit HMMDR and comparator methods to the scaled version of each data set. Where appropriate, we initialize the algorithms with the Gaussian hierarchical agglomerative procedure from MCLUST (cf. Fraley and Raftery, 1999). In the case of HMMDR and its analogues, we allow the number of components to vary between \( G = 1 \) and \( G = 6 \). Note that we use the term ‘analogue’ somewhat loosely here, because we do not consider decomposed covariance structures for either generalized hyperbolic mixtures or shifted asymmetric Laplace mixtures (Table 3.4).

Table 3.4: Details about the component covariance structures for each of the four MMDR methods used.

<table>
<thead>
<tr>
<th>Method</th>
<th>Covariance</th>
<th>Eigen-decomposed</th>
<th>Model family</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMMDR</td>
<td>( \Sigma_g + \alpha_g \alpha_g^\top )</td>
<td>No</td>
<td>–</td>
</tr>
<tr>
<td>SALMMDR</td>
<td>( \Sigma_g + \alpha_g \alpha_g^\top )</td>
<td>No</td>
<td>–</td>
</tr>
<tr>
<td>( t )MMDR</td>
<td>( \nu_g \Sigma_g, \nu_g &gt; 2 )</td>
<td>Yes</td>
<td>( t )EIGEN</td>
</tr>
<tr>
<td>GMMDR</td>
<td>( \Sigma_g )</td>
<td>Yes</td>
<td>MCLUST</td>
</tr>
</tbody>
</table>

In the context of model-based classification and discriminant analysis, we use the approach of McNicholas (2010) to simulate a situation in which some of the group memberships are unknown. For each observation \( \mathbf{x}_i \), a random number is generated from a uniform distribution on \([0, 1]\). If the random number is less than 0.5, then \( \mathbf{x}_i \) is taken as known; otherwise, \( \mathbf{x}_i \) is taken as unknown. To make sure that all the classes are represented, we repeated this procedure within each group until at least one known observation was produced before moving onto the next group.
We utilize the functionality of \texttt{teigen} to fit both Gaussian mixtures and $t$-mixtures for model-based classification (cf. Andrews and McNicholas, 2012a). Similarly, \texttt{teigen} and \texttt{mclust} (Fraley et al., 2012b) were employed for model-based discriminant analysis. For each data set, the procedures were run 25 times, using hierarchical agglomerative starting values of the unknown $\hat{z}_{ig}$. Note that the number of unknown observations varied from run to run, and so each run was applied to data sets with different numbers of observations. This is evidenced by the results of the model-based classification and discriminant analysis (Tables 3.11, 3.15, and 3.18).

Note that we choose each real data set on the basis that it has previously been used to illustrate the performance of some of the comparator methods. We consider that this approach facilitates a very fair comparison. In all cases, we illustrate model-based clustering, classification, and discriminant analysis. For model-based classification and discriminant analysis, the ARI is computed based only on unlabelled (unknown) observations. We perform random subset cross-validation, training on 25 different subsets consisting of roughly half the number of observations. This is more challenging than other well known procedures such as 10-fold cross-validation.

Swiss bank notes data

Flury and Riedwyl (1988) present six measurements taken from Swiss bank notes (length, diagonal, left, right, top, and bottom) and available through the R package \texttt{gclus} (Hurley, 2010). Each bank note is either genuine or counterfeit. In terms of model-based clustering, HMMDR and its comparators were fitted to these data and the resulting MAP classifications show very high ARI values for most meth-
ods (Table 3.6), with HMMDR and \( k \)-means being the only methods to cluster the bank notes data perfectly. For model-based classification, HMMDR, SALMMDR, and \( t \)MMDR produce perfect classifications of the unknown observations (Table 3.8). However, only HMMDR provides perfect model-based discriminant analysis results on the bank notes. We note that HMMDR selected the minimum possible number of features in all three scenarios, i.e., one feature.

Female voles data

Flury (1997, Table 5.3.7) discuss seven measurements of female voles from two species (\textit{Microtus californicus} and \textit{Microtus ochrogaster}, Table 3.5) originally studied by Airoldi and Hoffmann (1984). The data are available within the \texttt{R} package \texttt{flury} (Flury, 2010).

Table 3.5: Measurements taken for the female vole data.

<table>
<thead>
<tr>
<th>Age in days</th>
<th>Incisive foramen length</th>
<th>Zygomatic width</th>
</tr>
</thead>
<tbody>
<tr>
<td>Condylo incisive length</td>
<td>Skull height</td>
<td>Interorbital width</td>
</tr>
<tr>
<td>Alveolar length of upper molar tooth row</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Tables 3.7 and 3.9 indicate that, out of all of the procedures fitted to the voles data, HMMDR is the only one giving perfect classification results in all three paradigms.
Table 3.6: Summary of model-based clustering results for the bank data.

<table>
<thead>
<tr>
<th>Method</th>
<th>ARI</th>
<th>Feat.</th>
<th>Comp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMMDR</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>SALMMDR</td>
<td>0.98</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>tMMDR</td>
<td>0.98</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>GMMDR</td>
<td>0.98</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>ROBPCA</td>
<td>0.98</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>FisherEM</td>
<td>0.98</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>clustvarsel</td>
<td>0.85</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>mcfa</td>
<td>0.98</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>pgmm</td>
<td>0.82</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>kmeans</td>
<td>1</td>
<td>-</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 3.7: Summary of model-based clustering results for the voles data.

<table>
<thead>
<tr>
<th>Method</th>
<th>ARI</th>
<th>Feat.</th>
<th>Comp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMMDR</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>SALMMDR</td>
<td>0.95</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>tMMDR</td>
<td>0.91</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>GMMDR</td>
<td>0.91</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ROBPCA</td>
<td>0.91</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>FisherEM</td>
<td>0.66</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>clustvarsel</td>
<td>0.91</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>mcfa</td>
<td>0.91</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>pgmm</td>
<td>0.91</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>kmeans</td>
<td>0.74</td>
<td>-</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 3.8: Summary of model-based classification and discriminant analysis results for the bank data, based on 25 runs.

<table>
<thead>
<tr>
<th>Method</th>
<th>ARI</th>
<th>Feat.</th>
<th>Comp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMMDR class.</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>SALMMDR class.</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>tMMDR class.</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>GMMDR class.</td>
<td>0.96</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>HMMDR disc.</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>SALMMDR disc.</td>
<td>0.96</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>tMMDR disc.</td>
<td>0.95</td>
<td>1–2</td>
<td>2</td>
</tr>
<tr>
<td>GMMDR disc.</td>
<td>0.98</td>
<td>1–4</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 3.9: Summary of model-based classification and discriminant analysis results for the voles data, based on 25 runs.

<table>
<thead>
<tr>
<th>Method</th>
<th>ARI</th>
<th>Feat.</th>
<th>Comp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMMDR class.</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>SALMMDR class.</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>tMMDR class.</td>
<td>0.95</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>GMMDR class.</td>
<td>0.96</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>HMMDR disc.</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>SALMMDR disc.</td>
<td>0.91</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>tMMDR disc.</td>
<td>0.91</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>GMMDR disc.</td>
<td>0.88</td>
<td>1–4</td>
<td>2</td>
</tr>
</tbody>
</table>

Wine data

Forina et al. (1986) recorded several chemical and physical properties for three types of Italian wines: Barolo, Grignolino, and Barbera. As shown in Table 3.10, thirteen properties for 178 wines are available from the R package gclus (Hurley, 2004).

Within the model-based clustering framework, HMMDR is the best performer (ARI = 0.97, Table 3.12), with only two misclassified observations (Table 3.11). The model-based classification scenario (Table 3.13) reveals that HMMDR,
Table 3.10: Chemical and physical properties available for the wine data in gclus.

<table>
<thead>
<tr>
<th>Property</th>
<th>Alcohol</th>
<th>Proline</th>
<th>OD280/OD315 of diluted wines</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malic acid</td>
<td>Ash</td>
<td>Alkalinity of ash</td>
<td></td>
</tr>
<tr>
<td>Hue</td>
<td>Total phenols</td>
<td>Magnesium</td>
<td></td>
</tr>
<tr>
<td>Color intensity</td>
<td>Nonflavonoid phenols</td>
<td>Proanthocyanins</td>
<td></td>
</tr>
<tr>
<td>Flavonoids</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

SALMMDR, and tMMDR produce perfect classification results. With an ARI of 0.92, HMMDR gives the best performance within the model-based discriminant analysis paradigm.

Table 3.11: Model-based clustering, classification, and discriminant analysis results for our HMMDR approach fitted to the wine data. Model-based classification and discriminant analysis results are based on 25 runs.

<table>
<thead>
<tr>
<th>Wine</th>
<th>HMMDR clust.</th>
<th>HMMDR class.</th>
<th>HMMDR disc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barolo</td>
<td>59 0 0</td>
<td>875 0 0</td>
<td>625 50 0</td>
</tr>
<tr>
<td>Grignolino</td>
<td>0 68 2</td>
<td>0 925 0</td>
<td>0 825 0</td>
</tr>
<tr>
<td>Barbera</td>
<td>0 0 48</td>
<td>0 0 500</td>
<td>0 0 550</td>
</tr>
<tr>
<td>ARI; Features</td>
<td>0.97; 7</td>
<td>1; 5</td>
<td>0.92; 8</td>
</tr>
</tbody>
</table>

Table 3.12: Summary of model-based clustering results for the wine data.

<table>
<thead>
<tr>
<th>Method</th>
<th>ARI</th>
<th>Feat.</th>
<th>Comp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMMDR</td>
<td>0.97</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>SALMMDR</td>
<td>0.92</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>tMMDR</td>
<td>0.93</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>GMMMDR</td>
<td>0.85</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>ROBPCA</td>
<td>0.83</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>FisherEM</td>
<td>0.91</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>clustvarsel</td>
<td>0.78</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>mcfa</td>
<td>0.90</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>pgmm</td>
<td>0.79</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>kmeans</td>
<td>0.90</td>
<td>-</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 3.13: Summary of model-based classification and discriminant analysis results for the wine data, based on 25 runs.

<table>
<thead>
<tr>
<th>Method</th>
<th>ARI</th>
<th>Feat.</th>
<th>Comp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMMDR class.</td>
<td>1</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>SALMMDR class.</td>
<td>1</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>tMMDR class.</td>
<td>1</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>GMMMDR class.</td>
<td>0.93</td>
<td>3–8</td>
<td>3</td>
</tr>
<tr>
<td>HMMDR disc.</td>
<td>0.92</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>SALMMDR disc.</td>
<td>0.88</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>tMMDR disc.</td>
<td>0.85</td>
<td>1–8</td>
<td>3</td>
</tr>
<tr>
<td>GMMMDR disc.</td>
<td>0.85</td>
<td>2–8</td>
<td>3</td>
</tr>
</tbody>
</table>

Figure 3.5 illustrates three of the estimated HMMDR directions obtained
from our model-based clustering of the wine data (Table 3.11). The edge histograms depict the distribution of the observations from the estimated directions, coloured by estimated cluster allocation. The plot on the left-hand side reveals quite clearly the underlying cluster structure in the data. Although there is some overlap in these two directions, only two wines were misclassified by the HMMDR method and so some of the other dimensions must give additional clarity.

![Figure 3.5: Plots of some of the estimated HMMDR directions for the wine data obtained from model-based clustering (direction 2 vs. 4 on the left-hand side, and direction 4 vs. 5 on the right-hand side). Symbols indicate true cluster membership and colours indicate the estimated HMMDR cluster allocation. The edge histograms depict the estimated distributions of the observations in each cluster.](image)

**Wisconsin breast cancer data**

*Mangasarian et al. (1995)* presented a study of breast cancer from Wisconsin, undertaken to establish whether fine needle aspiration of breast tissue samples could
classify tumour status. Several attributes are recorded (Table 3.14) for 681 cases of potentially cancerous tumours, of which 238 were actually malignant. These data are available in the R package faraway (Faraway, 2011).

Table 3.14: Tissue sample properties of the Wisconsin breast cancer data.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Marginal adhesion</th>
<th>Epithelial cell size</th>
<th>Clump thickness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bare nuclei</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mitoses</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bland chromatin</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal nucleoli</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Within the model-based clustering framework, HMMDR selected three features and gave the best classification performance (ARI = 0.89, Table 3.15). The tMMDR (ARI = 0.86) and k-means (ARI = 0.84) approaches are close behind; however, the rest of the comparators did not produce particularly good results on these data (Table 3.16). For the model-based classification and discriminant analysis scenarios, HMMDR gave classification performance similar to clustering and was again the best performer over the 25 runs (Table 3.17).

Table 3.15: Model-based clustering, classification, and discriminant analysis results for our HMMDR approach on the breast cancer data. Model-based classification and discriminant analysis results are based on 25 runs.

<table>
<thead>
<tr>
<th>Status</th>
<th>HMMDR clust.</th>
<th>HMMDR class.</th>
<th>HMMDR disc.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Malignant</td>
<td>236</td>
<td>2925</td>
<td>2575</td>
</tr>
<tr>
<td>Benign</td>
<td>17</td>
<td>100</td>
<td>50</td>
</tr>
<tr>
<td>ARI; Features</td>
<td>0.89; 3</td>
<td>0.89; 2</td>
<td>0.87; 6</td>
</tr>
</tbody>
</table>

Figure 3.6 illustrates three of the estimated HMMDR directions obtained from the model-based clustering output shown in Table 3.15. The plots depict quite clearly the inherent cluster structure in the data, and we notice that the malignant
breast tissues are quite tightly packed in their cluster. The red histogram on the top edge of the left-hand side of Figure 3.6 is a nice illustration of a skew distribution along a HMMDR direction.
Figure 3.6: Plots of some of the estimated HMMDR directions for the breast cancer data obtained from model-based clustering (direction 1 vs. 3 on the left-hand side, and direction 2 vs. 3 on the right-hand side). Symbols indicate true cluster membership and colours indicate the estimated HMMDR cluster allocation. The edge histograms depict the estimated distributions of the observations in each cluster.

Colon data

Alon et al. (1999) analyzed gene expression data from microarray experiments of colon tissue, probed by oligonucleotide arrays. These data contain 62 tissue samples with 2000 genes: 40 tumour tissues and 22 normal tissues. The data are available in the R package plsgenomics (Boulesteix et al., 2011). The challenge here is dealing with the large number of gene expression levels compared with the small number of tissue samples. As is the case with microarray data in general, there are many non-informative genes that can obstruct the clustering of the samples. Thus, gene filtering was carried out prior to further analysis. While (McLachlan et al., 2002) and McNicholas and Murphy (2010) used the EMMIX-GENE procedure (McLachlan...
et al., 2002) to reduce the dimensionality of the colon data, we considered a different method. This was done in order to understand how the dimension reduction mixes the original directions, because EMMIX-GENE uses factor analyzers, so we would have a two-layered dimension reduction (i.e. two layers of variable mixing).

Our gene filtering approach is to find differentially expressed genes based on modified $t$-tests, using the R Bioconductor package \texttt{siggenes} (Schwender, 2012). This package contains the function \texttt{sam}, which implements the Significance Analysis of Microarrays (SAM) method proposed by Tusher et al. (2001). SAM computes a statistic $d_i$ for each gene $i$, measuring the strength of the relationship between gene expression and the response variable (which is the class variable in our case). It uses repeated permutations of the data to determine if the expression of any gene is significantly related to the response. The cutoff for significance is determined by a tuning parameter $\Delta$, chosen by the user based on the false positive rate. We employed 100 permutations and chose $\Delta = 2.4$, which yielded 23 genes for analysis (Table A.1, Appendix A).

Even with the dimensionality reduced to 23 genes, the analysis of the colon data was quite challenging. When HMMDR was fitted to these data within the model-based clustering paradigm, five observations were misclassified (Table 3.18), corresponding to an ARI of 0.70. This was the best result, with $t$MMDR (ARI = 0.64) being the next best performer (Table 3.19). For model-based classification and discriminant analysis, HMMDR gave better performance with ARI values of 0.86 and 0.83, respectively (Table 3.20). We note that SALMMDR, $t$MMDR, and GMMDR also gave improved performance within the model-based classification and
discriminant analysis paradigms; however, HMMDR was the best approach across all paradigms.

Table 3.18: Model-based clustering, classification, and discriminant analysis results for the colon data. Model-based classification and discriminant analysis results are based on 25 runs.

<table>
<thead>
<tr>
<th></th>
<th>HMMDR clust.</th>
<th>HMMDR class.</th>
<th>HMMDR disc.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Normal</td>
<td>22</td>
<td>0</td>
<td>225</td>
</tr>
<tr>
<td>Tumour</td>
<td>5</td>
<td>35</td>
<td>25</td>
</tr>
<tr>
<td>ARI; Features</td>
<td>0.70; 3</td>
<td>0.86; 5</td>
<td>0.83; 6</td>
</tr>
</tbody>
</table>

Table 3.19: Summary of model-based clustering results for the best models fitted to the colon data.

<table>
<thead>
<tr>
<th>Method</th>
<th>ARI</th>
<th>Feat.</th>
<th>Comp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMMDR</td>
<td>0.70</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>SALMMDR</td>
<td>0.59</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>tMMDR</td>
<td>0.64</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>GMMDR</td>
<td>0.59</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ROBPCA</td>
<td>0.36</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>FisherEM</td>
<td>0.59</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>clustvarsel</td>
<td>0.35</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>mcfai</td>
<td>0.64</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>pgmm</td>
<td>0.40</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>kmeans</td>
<td>0.59</td>
<td>-</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 3.20: Summary of model-based classification and discriminant analysis results for the colon data, based on 25 runs.

<table>
<thead>
<tr>
<th>Method</th>
<th>ARI</th>
<th>Feat.</th>
<th>Comp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMMDR class.</td>
<td>0.86</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>SALMMDR class.</td>
<td>0.71</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>tMMDR class.</td>
<td>0.75</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>GMMDR class.</td>
<td>0.73</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>HMMDR disc.</td>
<td>0.83</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>SALMMDR disc.</td>
<td>0.70</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>tMMDR disc.</td>
<td>0.74</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>GMMDR disc.</td>
<td>0.70</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

It is interesting to note that McNicholas and Murphy (2010) obtained similar clustering results to HMMDR for the colon data. They considered a subset of 461 genes and the best model fitted to these data had six latent factors, with five misclassified tissues and an ARI of 0.70. Although equal to two significant figures, we point out for completeness that our HMMDR approach has a very slightly higher ARI value (0.699 vs. 0.697). McLachlan et al. (2002) also analyzed these data; they selected 446 genes and identified five types of clusterings on this subset. However,
these clusterings did not correspond to the tissue type.

Figure 3.7 illustrates three of the estimated HMMDR directions obtained from the model-based clustering output from Table 3.18. The plots depict the inherent cluster structure in the colon tissues and, as we would expect, the misclassified tissues are generally close to the cluster boundaries.

![Figure 3.7: Plots of some of the estimated HMMDR directions for the colon data obtained from model-based clustering (direction 2 vs. 3 on the left-hand side, and direction 1 vs. 2 on the right-hand side). Symbols indicate true cluster membership and colours indicate the estimated HMMDR cluster allocation. The edge histograms depict the estimated distributions of the observations in each cluster.](image)

3.3 Summary

This chapter introduced an effective dimension reduction technique for model-based clustering, classification, and discriminant analysis using multivariate mixtures
of generalized hyperbolic distributions. Our method, known as HMMDR, focused on identifying the smallest subspace of the data that captured the inherent cluster structure. The HMMDR approach was illustrated using simulated and real data, where it performed favourably compared to its existing special cases, i.e., GMMDR, $t$MMDR, and SALMMMDR. In clustering applications, HMMDR consistently outperformed several other model-based dimension reduction methods (ROBPCA, pgmm, FisherEM, clustvarsel, and mcfa).

One limitation sometimes encountered using our current approach is singularities while fitting full $p \times p$ covariance matrices for each mixture component when there are $n_g < p$ observations for the component. One purely numerical approach we have used to mitigate singularities was to regularize the updated covariance matrices at each iteration in the EM algorithm using a suitably chosen eigenvalue cutoff. Future work could include restricted models for the covariance structure or some alternative approaches for dealing with singularity problems in dimension reduction methods. This could include working on distinct subspaces for fitting each covariance matrix and only injecting into a common subspace when forming the kernel matrix $M_{\text{HMM}}$.

The real data sets used for our illustrations were selected because they were previously used to illustrate the performance of some of the comparator methods. Therefore, it is encouraging that HMMDR consistently outperformed all comparator methods on all data sets. Future work could study whether incorporating generalized hyperbolic analogues of the GPCM models into our approach is beneficial. While such an approach is taken within GMMDR and $t$MMDR, we believe that careful
study is needed before determining whether adding these analogues is desirable. This also ties in with whether multiple components should be available to represent a class in HMMDR discriminant analysis, another topic deserving of further work. Finally, the application of our approach within the fractionally-supervised classification framework (Vrbik and McNicholas, 2013) could also be a subject of future work.
Chapter 4

Outlier Detection for Clustering using Mixtures of Contaminated Shifted Asymmetric Laplace Distributions

In this chapter we discuss the detection of outliers for model-based clustering using the class of multivariate shifted asymmetric Laplace distributions (SAL) proposed by Franczak et al. (2012). As outlined in Chapter 2, SAL mixtures are an effective method of clustering asymmetric data. A method similar to that of Punzo and McNicholas (2013) for contaminated Gaussian mixtures is developed for clustering with contaminated SAL mixtures.

4.1 Methodology

Recall that the density of a component in a SAL mixture model appeared in (2.2)

\[
  f_{SAL}(x|\alpha_g, \Sigma_g, \mu_g) = \frac{2}{(2\pi)^{p/2}|\Sigma_g|^{1/2}} \exp\left\{ -\frac{(x - \mu_g)^\top \Sigma_g^{-1} \alpha_g}{2 + \alpha_g^\top \Sigma_g^{-1} \alpha_g} \right\} K_\nu(u),
\]

(4.1)

where \( \mu_g, \Sigma_g, \alpha, \delta(x, \mu_g|\Sigma_g), u, \) and \( K_\nu \) are as defined in Chapter 2.
We follow the approach of Punzo and McNicholas (2013) for contaminated Gaussian mixtures to build the framework for model-based clustering with contaminated SAL mixtures. To this end, we employ the contamination scheme where $\Sigma_g$ gets scaled to $\rho_g \Sigma_g$ and $\alpha_g$ gets scaled to $\sqrt{\rho_g} \alpha_g$, with $\rho_g > 0$ denoting the contamination factor. This leads to the following contaminated SAL distribution

$$C_{SAL}(x|\vartheta) = \lambda_g f_{SAL}(x_i|\mu_g, \alpha_g, \Sigma_g) + (1 - \lambda_g) f_{SAL}(x_i|\mu_g, \sqrt{\rho_g} \alpha_g, \rho_g \Sigma_g), \quad (4.2)$$

where $\lambda_g \in [0, 1]$ denotes the proportion of non-outliers. The corresponding contaminated SAL mixture model is defined by

$$f_{SAL}(x|\vartheta) = \sum_{g=1}^{G} \pi_g \{ \lambda_g f_{SAL}(x_i|\mu_g, \alpha_g, \Sigma_g) + (1 - \lambda_g) f_{SAL}(x_i|\mu_g, \sqrt{\rho_g} \alpha_g, \rho_g \Sigma_g) \}. \quad (4.3)$$

There are two sources of missing data: the group memberships, $z_{ig}$, and the fact that we do not know whether an observation in group $g$ is “good” or “bad”. Using the notation of Punzo and McNicholas (2013), we define this second source of missing data to be $v_{ig}$, such that

$$v_{ig} = \begin{cases} 
1 & \text{if observation } x_i \text{ in group } g \text{ is “good” and} \\
0 & \text{if observation } x_i \text{ in group } g \text{ is “bad”}. 
\end{cases} \quad (4.4)$$

Note that we do not consider covariance decomposed structures for the model in (4.3), unlike the method of Punzo and McNicholas (2013). Instead, we focus on the most general case by letting the covariance vary between components, which is akin to the Gaussian mixture model VVV of Celeux and Govaert (1995).
We can visualize the effect of our contamination scheme on a typical SAL mixture. Figure 4.1 depicts this for a SAL mixture with $\alpha = (1.5, 1.5)^T$, $\mu = (0, 0)^T$, $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$, and a contamination factor of $\rho = 4$. One of the advantages of employing the contamination scheme described above is that the mode in the contaminated SAL density is the same as in the original density. This can be seen in Figure 4.1.

Figure 4.1: Contour plot of a contaminated SAL mixture superimposed on its uncontaminated SAL mixture counterpart.

The complete-data likelihood for the mixture of contaminated SAL distri-
The complete-data log-likelihood corresponding to (4.7) is

\[ l(\theta | x) = \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} v_{ig} \log(\pi_g) + \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} v_{ig} \log(\lambda_g) \]

\[ + \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} v_{ig} \log(f_{\text{SAL}}(x_i | \mu_g, \alpha_g, \Sigma_g)) \]

\[ + \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \log(1 - v_{ig}) \log(\pi_g) + \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \log(1 - v_{ig}) \log(1 - \lambda_g) \]

\[ + \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \log(f_{\text{SAL}}(x_i | \mu_g, \sqrt{\rho_g} \alpha_g, \rho_g \Sigma_g)) \]  

(4.6)

Computationally, it is more efficient to use the relationship between the SAL and Gaussian distributions outlined in Chapter 2 to rewrite the above complete-data log-likelihood as

\[ l(\theta | x) = \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} v_{ig} \log(\pi_g) + \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} v_{ig} \log(\lambda_g) \]

\[ + \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} v_{ig} \log(\phi(x_i | \mu_g + w_{ig} \alpha_g, w_{ig} \Sigma_g) h(w_{ig})) \]

\[ + \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \log(1 - v_{ig}) \log(\pi_g) + \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \log(1 - v_{ig}) \log(1 - \lambda_g) \]

\[ + \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \log(\phi(x_i | \mu_g + w_{ig} \sqrt{\rho_g} \alpha_g, w_{ig} \rho_g \Sigma_g) h(w_{ig})). \]  

(4.7)
Furthermore, one can simplify (4.7) to

\[
I(\theta | \mathbf{x}) = \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} v_{ig} \log(\pi_g) + \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} v_{ig} \log(\lambda_g) \\
+ \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} v_{ig} \log \left( \frac{1}{(2\pi)^{p/2} |w_{ig} \Sigma_g|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{x}_i - \mu_g - w_{ig} \alpha_g) ^\top \right\} \right) \\
\times (w_{ig} \Sigma_g)^{-1} (\mathbf{x}_i - \mu_g - w_{ig} \alpha_g) \exp(-w_{ig})) \\
+ \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} (1 - v_{ig}) \log(\pi_g) + \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} (1 - v_{ig}) \log(1 - \lambda_g) \\
+ \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} (1 - v_{ig}) \log \left( \frac{1}{(2\pi)^{p/2} |w_{ig} \rho_{g} \Sigma_g|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{x}_i - \mu_g - w_{ig} \sqrt{\rho_g} \alpha_g) ^\top \right\} \right) \\
\times (w_{ig} \rho_{g} \Sigma_g)^{-1} (\mathbf{x}_i - \mu_g - w_{ig} \sqrt{\rho_g} \alpha_g) \exp(-w_{ig})) \right) .
\]

(4.8)

In the context of estimating parameters for the contaminated model, we aim to establish their identifiability. This is an important requirement that the asymptotic theory needs to hold for maximum likelihood estimation of the model parameters. We follow the approach of Punzo and McNicholas (2013) to show this. First, we assume that a mixture of SAL distributions is identifiable. However, a mixture of mixtures is not identifiable in general. In order to apply the usual theory of identifiability to both the contaminated and uncontaminated parts of our model at the same time, it is convenient to rewrite the model density

\[
f_{\text{SAL}}(\mathbf{x} | \theta) = \sum_{g=1}^{G} \pi_g \{ \lambda_g f_{\text{SAL}}(\mathbf{x}_i | \mu_g, \alpha_g, \Sigma_g) + (1 - \lambda_g) f_{\text{SAL}}(\mathbf{x}_i | \mu_g, \sqrt{\rho_g} \alpha_g, \rho_g \Sigma_g) \}.
\]

(4.9)
as
\[
f_{\text{SAL}}(x|\theta) = \sum_{g=1}^{G} \sum_{h=1}^{2} \pi_g \lambda_{gh} f_{\text{SAL}}(x_i|\mu_g, \sqrt{\rho_{gh}} \alpha_g, \rho_{gh} \Sigma_g) \tag{4.10}
\]
where we have set \(\lambda_{g1} = \lambda_g, \lambda_{g2} = 1 - \lambda_g\) and \(\rho_{g1} = 1, \rho_{g2} = \rho_g\). In the next proposition we consider the identifiability of this form of our model.

**Proposition 4.1.1** Let
\[
f_{\text{SAL}}(x|\theta) = \sum_{g=1}^{G} \sum_{h=1}^{2} \pi_g \lambda_{gh} f_{\text{SAL}}(x_i|\mu_g, \sqrt{\rho_{gh}} \alpha_g, \rho_{gh} \Sigma_g)
\]
and
\[
f_{\text{SAL}}(x|\tilde{\theta}) = \sum_{s=1}^{\tilde{G}} \sum_{t=1}^{2} \tilde{\pi}_s \tilde{\lambda}_{st} f_{\text{SAL}}(x_i|\tilde{\mu}_s, \sqrt{\tilde{\rho}_{st}} \tilde{\alpha}_s, \tilde{\rho}_{st} \tilde{\Sigma}_s)
\]
be two different parametrizations of the same contaminated SAL mixture. If for \(g \neq g^*\) we have
\[
\|\mu_g - \mu_{g^*}\|_2^2 + \|\Sigma_g - a \Sigma_{g^*}\|_2^2 > 0 \tag{4.11}
\]
for all \(a > 0\), where \(\|\cdot\|_2^2\) is the Frobenius norm, then \(G = \tilde{G}\) and there exists a relabelling such that
\[
\pi_g = \tilde{\pi}_g, \lambda_{gh} = \tilde{\lambda}_{gh}, \mu_g = \tilde{\mu}_g, \alpha_g = \tilde{\alpha}_g, \Sigma_g = \tilde{\Sigma}_g \text{ and } \rho_{gh} = \tilde{\rho}_{gh}.
\]

**Proof.** From the identifiability of finite mixtures of SAL distributions we have that \(2G = 2\tilde{G}\) and for each pair \((g,h)\) there is a corresponding \((s,t)\) with
\[
\pi_g \lambda_{gh} = \tilde{\pi}_s \tilde{\lambda}_{st}, \mu_g = \tilde{\mu}_s, \sqrt{\rho_{gh}} \alpha_g = \sqrt{\tilde{\rho}_{st}} \tilde{\alpha}_s \text{ and } \rho_{gh} \Sigma_g = \tilde{\rho}_{st} \tilde{\Sigma}_s \tag{4.12}
\]
Let us consider the correspondence of \((g,1)\) and \((s_1,t_1)\) together with that of \((g,2)\) and \((s_2,t_2)\). Now \(\mu_g = \tilde{\mu}_{s_1} = \tilde{\mu}_{s_2}\) so that \((4.11)\) implies that \(s_1 = s_2\), which
we will call \( s \) from now on. Since \( \mu_g = \tilde{\mu}_s \) the assumption (4.11) also implies that the parameters \( \Sigma_g \) and \( \tilde{\Sigma}_s \) are not proportional, and so \( \Sigma_g = \rho_{g1}\Sigma_g = \tilde{\rho}_{st1}\tilde{\Sigma}_s \) implies \( \tilde{\rho}_{st1} = 1 \) and \( \Sigma_g = \tilde{\Sigma}_s \). So \( \rho_{g1} = \tilde{\rho}_{st1} \) and \( \rho_{g2} = \tilde{\rho}_{st2} \), hence \( \alpha_g = \tilde{\alpha}_s \).

Given that we have determined the correspondence for the \( \rho_{gh} \), and so which of the pair of distributions is contaminated

\[
1 \neq \rho_{g2} = \tilde{\rho}_{st2}, \text{ for some } t_2 \in \{1, 2\}
\]

we can recover \( \tilde{\pi}_s \) and \( \tilde{\lambda}_{st2} \) from the joint probabilities \( \tilde{\rho}_{st} = \tilde{\pi}_s\tilde{\lambda}_{st} = \pi_g\lambda_{gh} \) as follows

\[
\tilde{\pi}_s = \tilde{\rho}_{s1} + \tilde{\rho}_{s2} = \pi_g, \quad \tilde{\lambda}_{st2} = \frac{\tilde{\rho}_{st2}}{\tilde{\pi}_s} = \frac{\pi_g\lambda_{g2}}{\pi_g} = \lambda_{g2}.
\]

Using \( \tilde{\lambda}_s1 + \tilde{\lambda}_{s2} = 1 \) we have also \( \lambda_{g1} = \tilde{\lambda}_{st1} \). This extends the correspondence for the \( g \)-th pair to all the parameters, which we can then do for each \( g \) to complete the proof. \( \square \)

We employ the ECM algorithm to find the maximum likelihood estimates in (4.8). At the E-step of the algorithm the Q-function, \( Q(\theta|\hat{\theta}) = E[l(\theta|x)|x, \hat{\theta}] \) is computed. This Q-function denotes the conditional expectation of (4.8) given the observed data \( x \) and the current estimated parameters \( \hat{\theta} \). For simplicity we let

\[
l(\theta|x) = l_1(\theta|x) + l_2(\theta|x), \text{ where } l_1(\theta|x) \text{ contains the first three terms from (4.8) and } l_2(\theta|x) \text{ contains the last three terms from (4.8).}
\]

Then we can write the Q-function as

\[
Q = E[l_1(\theta|x)] + E[l_2(\theta|x)],
\]

(4.13)
thus for the first term in the Q-function (4.13) we obtain

\[ E[l_1(\theta | x)] = \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} \hat{v}_{ig} \log(\pi_g) + \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} \hat{v}_{ig} \log(\lambda_g) - \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} \hat{v}_{ig} \frac{D}{2} \log(2\pi) + \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \hat{v}_{ig} \log(\Sigma^{-1}) - \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} \hat{v}_{ig} E[\log(W_{ig} | x_i, z_{ig} = 1)] \]

\[ \times E[|x_i - \mu_g|^2 \Sigma^{-1} | x_i, z_{ig} = 1] - \frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} \hat{v}_{ig} E[|W_{ig}| \Sigma^{-1} (x_i - \mu_g)] - \frac{1}{2} n \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} \hat{v}_{ig} E[|W_{ig} | x_i, z_{ig} = 1] \alpha_g \Sigma^{-1} \alpha_g. \]

(4.14)

Since the terms \(-\frac{p}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} \hat{v}_{ig} E[\log(W_{ig}) | x_i, z_{ig} = 1] \) and \(-\sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} \hat{v}_{ig} \times E[|W_{ig}| x_i, z_{ig} = 1] \) are constant with respect to the model parameters, we can omit them from our calculations. The second term in the Q-function (4.13) has the following structure
\[
E[l_2(\theta | x)] = \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig}) \log(1 - \lambda_g) + \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig}) \log(\pi_g)
\]

\[
- \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig}) \frac{p}{2} \log(2\pi) + \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig}) \frac{1}{2} \log(||(\rho_g \Sigma_g)^{-1}||)
\]

\[
- \frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig}) E[\hat{\bar{W}}_{ig}^{-1}|x_i, z_{ig} = 1](x_i - \mu_g)^\top(\rho_g \Sigma_g)^{-1}(x_i - \mu_g)
\]

\[
- \frac{p}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig}) E[\log(\hat{\bar{W}}_{ig})|x_i, z_{ig} = 1]
\]

\[
+ \frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig})(x_i - \mu_g)^\top(\rho_g \Sigma_g)^{-1}(\sqrt{\rho_g} \alpha_g)
\]

\[
+ \frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig})(\sqrt{\rho_g} \alpha_g)^\top(\rho_g \Sigma_g)^{-1}(x_i - \mu_g)
\]

\[
- \frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig}) E[\hat{\bar{W}}_{ig}|x_i, z_{ig} = 1](\sqrt{\rho_g} \alpha_g)^\top(\rho_g \Sigma_g)^{-1}(\sqrt{\rho_g} \alpha_g)
\]

\[
- \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig}) E[\hat{\bar{W}}_{ig}|x_i, z_{ig} = 1].
\]

(4.15)

Here \(\hat{W}\) is the analogue of \(W\) when \(\Sigma_g \rightarrow \rho_g \Sigma_g\) and \(\alpha_g \rightarrow \sqrt{\rho_g} \alpha_g\). Once again two terms can be omitted from the calculations, namely \(-\frac{p}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig}) E[\log(\hat{\bar{W}}_{ig})|x_i, z_{ig} = 1]\) and \(-\sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig}) E[\hat{\bar{W}}_{ig}|x_i, z_{ig} = 1]\), because they are constant with respect to the model parameters.

Now, at the E-step of the ECM algorithm we compute the update for \(z_{ig}\)

\[
\hat{z}_{ig} = \frac{\pi_g c_{\text{sal}}(x_i|\vartheta_g)}{\sum_{g=1}^{G} \pi_g c_{\text{sal}}(x_i|\vartheta_g)} = \frac{\pi_g [\lambda_g f_{\text{sal}}(x_i|\mu_g, \alpha_g, \Sigma_g) + (1 - \lambda_g) f_{\text{sal}}(x_i|\mu_g, \sqrt{\rho_g} \alpha_g, \rho_g \Sigma_g)]}{\sum_{g=1}^{G} \pi_g [\lambda_g f_{\text{sal}}(x_i|\mu_g, \alpha_g, \Sigma_g) + (1 - \lambda_g) f_{\text{sal}}(x_i|\mu_g, \sqrt{\rho_g} \alpha_g, \rho_g \Sigma_g)]},
\]

(4.16)
and also derive the update for $v_{ig}$

$$
\hat{v}_{ig} = \frac{\lambda_g f_{\text{SAL}}(x_i | \vartheta_g)}{C_{\text{SAL}}(x_i | \vartheta_g)} = \frac{\lambda_g f_{\text{SAL}}(x_i | \mu_g, \alpha_g, \Sigma_g)}{\lambda_g f_{\text{SAL}}(x_i | \mu_g, \alpha_g, \Sigma_g) + (1 - \lambda_g) f_{\text{SAL}}(x_i | \mu_g, \sqrt{\rho_g} \alpha_g, \rho_g \Sigma_g)}.
$$

(4.17)

At the first CM-step of the ECM algorithm we update $\pi_g$ and $\lambda_g$ with

$$
\hat{\pi}_g = \frac{\sum_{i=1}^n \hat{z}_{ig}}{n}, \quad \hat{\lambda}_g = \frac{\sum_{i=1}^n \hat{z}_{ig} \hat{v}_{ig}}{\sum_{i=1}^n \hat{z}_{ig}},
$$

respectively. Then we use the Q-function (4.13) to compute parameter updates for $\mu_g, \Sigma_g, A_g$ and $\rho_g$. In what follows we introduce some notation to simplify the calculation of these updates. Full details on their mathematical derivation appear in Section A.1 of the Appendix A. First, we set

$$
E_{1ig} := E[W_{ig} | x_i, z_{ig} = 1] = \sqrt{\frac{a}{b}} K_w(u),
$$

$$
E_{2ig} := E[W_{ig}^{-1} | x_i, z_{ig} = 1] = \sqrt{\frac{b}{a}} K_w(u) - \frac{2\nu}{a},
$$

$$
E_{3ig} := E[\hat{W}_{ig} | x_i, z_{ig} = 1] = \sqrt{\frac{a}{\tilde{b}}} K_w(\tilde{u}),
$$

$$
E_{4ig} := E[\hat{W}_{ig}^{-1} | x_i, z_{ig} = 1] = \sqrt{\frac{\tilde{b}}{a}} K_w(\tilde{u}) - \frac{2\nu}{\tilde{a}},
$$

(4.18)

where $a = (x_i - \mu_g)^\top \Sigma_g^{-1} (x_i - \mu_g)$ and $b = 2 + \alpha_g^\top \Sigma_g^{-1} \alpha_g$, as defined previously in Chapter 2, with $\tilde{a}$ and $\tilde{b}$ their analogues when $\Sigma_g^{-1} \rightarrow (\rho_g \Sigma_g)^{-1}$ and $\alpha_g \rightarrow \sqrt{\rho_g} \alpha_g$. Also, we set

$$
u = \sqrt{ab} \quad \text{and} \quad \tilde{u} = \sqrt{\tilde{a} \tilde{b}}.
$$

(4.19)
Furthermore, we use the following

\[
A = \sum_{i=1}^{n} \hat{z}_{ig}(\hat{v}_{ig}E_{2ig} + \frac{1 - \hat{v}_{ig}}{\rho_g}E_{4ig}), \\
B = \sum_{i=1}^{n} \hat{z}_{ig}(\hat{v}_{ig}E_{1ig} + (1 - \hat{v}_{ig})E_{3ig}), \\
C = \sum_{i=1}^{n} \hat{z}_{ig}(\hat{v}_{ig} + \frac{1 - \hat{v}_{ig}}{\rho_g}), \\
D = \sum_{i=1}^{n} \hat{z}_{ig}(\hat{v}_{ig} + \frac{1 - \hat{v}_{ig}}{\sqrt{\rho_g}}),
\]

(4.20)

to simplify the notation in the calculations. It turns out that the update for \( \mu_g \) becomes

\[
\hat{\mu}_g = \frac{B \left( \sum_{i=1}^{n} \hat{z}_{ig}(\hat{v}_{ig}E_{2ig} + \frac{1 - \hat{v}_{ig}}{\rho_g}E_{4ig})x_i \right) - C \left( \sum_{i=1}^{n} \hat{z}_{ig}(\hat{v}_{ig} + \frac{1 - \hat{v}_{ig}}{\rho_g})x_i \right)}{BA - C^2},
\]

(4.21)

and the update for \( \alpha_g \) is

\[
\hat{\alpha}_g = \frac{A \left( \sum_{i=1}^{n} \hat{z}_{ig}(\hat{v}_{ig} + \frac{1 - \hat{v}_{ig}}{\sqrt{\rho_g}})x_i \right) - D \left( \sum_{i=1}^{n} \hat{z}_{ig}(\hat{v}_{ig}E_{2ig} + \frac{1 - \hat{v}_{ig}}{\rho_g}E_{4ig})x_i \right)}{BA - D^2}.
\]

(4.22)

Following the derivations in Section A.1 of the Appendix A, we find the update for \( \Sigma_g \) to be

\[
\hat{\Sigma}_g = \frac{1}{\sum_{i=1}^{n} \hat{z}_{ig}} A(x_i - \hat{\mu}_g)(x_i - \hat{\mu}_g)^\top \\
- \frac{1}{\sum_{i=1}^{n} \hat{z}_{ig}} \sum_{i=1}^{n} \hat{z}_{ig} \left( \hat{v}_{ig} + \frac{1 - \hat{v}_{ig}}{\sqrt{\rho_g}} \right)(x_i - \hat{\mu}_g)\hat{\alpha}_g^\top \\
- \frac{1}{\sum_{i=1}^{n} \hat{z}_{ig}} \sum_{i=1}^{n} \hat{z}_{ig} \left( \hat{v}_{ig} + \frac{1 - \hat{v}_{ig}}{\sqrt{\rho_g}} \right)\hat{\alpha}_g(x_i - \hat{\mu}_g)^\top \\
+ \frac{1}{\sum_{i=1}^{n} \hat{z}_{ig}} \alpha_g\alpha_g^\top B.
\]

(4.23)
At the second CM-step of the ECM algorithm we update the contamination factor \( \rho_g \). Looking at the Q-function in (4.13) we observe that there are two possible ways of updating \( \rho_g \).

1. Use the \( R \) function \texttt{optim()} to find a maximum for \( Q \), as \( \rho_g \) varies in the interval \((1, \rho_g^{\text{max}})\), by considering

\[
Q = \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig}(1 - v_{ig}) \left( -\frac{p}{2} \log(\rho_g) \right) \\
- \frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig}) E_{4ig}(x_i - \mu_g)^\top (\rho_g \Sigma_g)^{-1}(x_i - \mu_g) \\
+ \frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig}) (\sqrt{\rho_g} \alpha_g)^\top (\rho_g \Sigma_g)^{-1}(x_i - \mu_g) \\
+ \frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig}) (\sqrt{\rho_g} \alpha_g)^\top (\rho_g \Sigma_g)^{-1}(x_i - \mu_g) \\
+ \text{terms independent of } \rho_g. \tag{4.24}
\]

2. Solve the equation \( \frac{\partial Q}{\partial \rho_g} = 0 \) for \( \rho_g \) directly,

\[
0 = \rho_g p \sum_{i=1}^{n} \hat{z}_{ig}(1 - \hat{v}_{ig}) + \rho_g^{1/2} \left[ \frac{1}{4} \sum_{i=1}^{n} \hat{z}_{ig}(1 - \hat{v}_{ig})(x_i - \mu_g)^\top \Sigma_g^{-1} \alpha_g \\
+ \frac{1}{4} \sum_{i=1}^{n} \hat{z}_{ig}(1 - \hat{v}_{ig}) \alpha_g^\top \Sigma_g^{-1}(x_i - \mu_g) \right] \\
- \frac{1}{2} \sum_{i=1}^{n} \hat{z}_{ig}(1 - \hat{v}_{ig}) E_{4ig}(x_i - \mu_g)^\top \Sigma_g^{-1}(x_i - \mu_g), \tag{4.25}
\]

which is a quadratic equation of the form \( S \rho_g + T \rho_g^{1/2} + U = 0 \) that can be easily solved for \( \rho_g \) (where \( S, T, U \) are constants).

We found it beneficial to implement both methods of computing the \( \rho_g \) update because they can be used interchangeably in case of numerical errors.
Within the model-based clustering framework, we allow the number of components, $G$ to vary between $G = 2$ and $G = (\text{no. known components}+2)$, and the best model is selected using the BIC (1.17). In order to find the BIC in the case of contaminated SAL mixtures, one needs to determine the number of free parameters in the model (Table 4.1). Thus, we have $\text{BIC} = 2l(\mathbf{x}, \hat{\theta}) - (3G-1+2Gp+Gp(p+1)/2) \log(n)$.

Table 4.1: Free parameters in the contaminated SAL mixture model.

<table>
<thead>
<tr>
<th></th>
<th>$\pi$</th>
<th>$\lambda$</th>
<th>$\rho$</th>
<th>$\alpha$</th>
<th>$\mu$</th>
<th>$\Sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free parameters</td>
<td>$G-1$</td>
<td>$G$</td>
<td>$Gp$</td>
<td>$Gp$</td>
<td>$\frac{Gp(p+1)}{2}$</td>
<td></td>
</tr>
</tbody>
</table>

One can extend the model-based clustering paradigm for contaminated SAL mixtures to model-based classification, using an analogous process. In this context, the likelihood equation corresponding to our model in (4.3) becomes

$$
\mathcal{L}(\theta|\mathbf{x}) = \prod_{i=1}^{k} \prod_{g=1}^{G} \left\{ \pi_g \left[ \lambda_g f_{SAL}(\mathbf{x}_i|\mu_g, \alpha_g, \Sigma_g) \right] \left[ (1-\lambda_g) f_{SAL}(\mathbf{x}_i|\mu_g, \sqrt{\rho_g} \alpha_g, \rho_g \Sigma_g) \right] \right\}^{z_{ig}} \\
\times \prod_{i=k+1}^{n} \sum_{g=1}^{G} \pi_h \left[ \lambda_h f_{SAL}(\mathbf{x}_i|\mu_h, \alpha_h, \Sigma_h) \right] \left[ (1-\lambda_h) f_{SAL}(\mathbf{x}_i|\mu_h, \sqrt{\rho_h} \alpha_h, \rho_h \Sigma_h) \right].
$$

(4.26)

The details regarding this extension and parameter updates appear in Section A.1.2 of the Appendix A.
4.2 Applications

4.2.1 Simulated data

We generated \( n = 180 \) data points from a SAL mixture with two components of equal size \( (n_1 = n_2 = 90) \) using the relationship \( \mathbf{X} = \mathbf{\mu} + W\mathbf{\alpha} + \sqrt{W}\mathbf{Y} \), as discussed in (2.3). The specific parameters used to simulate these data appear in Table 4.2. Then we considered two simulation schemes to test our contaminated SAL model (4.3) : scenario one has noise added, and scenario two incorporates artificial “outliers”.

Table 4.2: Parameters used to simulate data for a two component SAL mixture.

<table>
<thead>
<tr>
<th>Component 1</th>
<th>Component 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>( W_1 \sim \exp(1) )</td>
<td>( W_2 \sim \exp(1) )</td>
</tr>
<tr>
<td>( \mathbf{\alpha}_1 = \begin{pmatrix} 1 \ 1 \end{pmatrix} )</td>
<td>( \mathbf{\alpha}_2 = \begin{pmatrix} -1 \ -1 \end{pmatrix} )</td>
</tr>
<tr>
<td>( \mathbf{\mu}_1 = \begin{pmatrix} 1 \ -2 \end{pmatrix} )</td>
<td>( \mathbf{\mu}_2 = \begin{pmatrix} 5 \ -5 \end{pmatrix} )</td>
</tr>
<tr>
<td>( \Sigma_1 = \begin{bmatrix} 1 &amp; 0.5 \ 0.5 &amp; 1 \end{bmatrix} )</td>
<td>( \Sigma_2 = \begin{bmatrix} 1 &amp; 0 \ 0 &amp; 1 \end{bmatrix} )</td>
</tr>
</tbody>
</table>

For the first simulation scenario we added 20 noise observations to each mixture component by generating them from a uniform distribution over the range \([-10, 10]\). We fit 100 of these simulated mixtures with noise using the contaminated SAL model (4.3).

To gauge the performance of the fitted contaminated mixture, we use the
values of \( v_{ig} (4.4) \) to determine if a point should be assigned to the “good” or “bad” categories. This resulted in 3 fitted groups, namely Component 1, Component 2, and Noise. The observation assignments to these groups were compared against the known labels of the simulated data. Table 4.3 shows the best fit observed among the 100 runs, with an ARI of 0.9349. Overall, we obtain an average ARI of 0.8093 with a standard deviation of 0.0734.

Table 4.3: Clustering results for the best model, in terms of ARI, fitted to the simulated SAL data in scenario one.

<table>
<thead>
<tr>
<th>Known</th>
<th>Fitted</th>
<th>Comp. 1</th>
<th>Comp. 2</th>
<th>Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comp. 1</td>
<td>90</td>
<td>0</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Comp. 2</td>
<td>0</td>
<td>90</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Noise</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4.2 plots the results of the clustering with contaminated SAL mixtures for scenario one. Colour indicates the estimated group membership and shape indicates the known membership.

For the second simulation scenario we added 20 artificial “outliers” by generating data using \( X = \mu + W\alpha + \sqrt{W}Y \) with the same parameters, but increasing the size of \( Y \) by a scaling factor to produce more widely dispersed values. Note that not all of these added data points will end up being visually recognised as outliers. We proceeded to fit 100 of these simulated mixtures with outliers using the contaminated SAL mixture model (4.3).

To determine an ARI for the fitted contaminated mixture, we use the val-
Figure 4.2: Scatterplot of simulated data for scenario one, with added contours. Color indicates the estimated component membership and shape indicates the known membership.

values of $v_{ig}$ (4.4) to determine if a point should be assigned to the “good” or “bad” categories. This resulted in 3 fitted groups, namely Component 1, Component 2, and Outliers. The observation assignments to these groups were compared against the known labels of the simulated data. Table 4.4 shows the best fit observed among the 100 runs, with an ARI of 0.9051. Overall, the average ARI was 0.8199 with a standard deviation of 0.0495.

Figure 4.3 plots the results of the clustering with contaminated SAL mixtures for scenario two. Colour indicates the estimated group membership and shape indicates the known membership. The + symbol denotes the simulated outliers which,
Table 4.4: Clustering results for the best model, in terms of BIC, fitted to the simulated SAL data in scenario two.

<table>
<thead>
<tr>
<th></th>
<th>Fitted</th>
<th>Known</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Comp. 1</td>
<td>Comp. 2</td>
</tr>
<tr>
<td>Comp. 1</td>
<td>90</td>
<td>0</td>
</tr>
<tr>
<td>Comp. 2</td>
<td>0</td>
<td>90</td>
</tr>
<tr>
<td>Outliers</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

as mentioned earlier, do not all have the visual appearance of outliers. Indeed those “outliers” near the fitted components have been assigned to one of the components as indicated by their color.

Figure 4.3: Scatterplot of simulated data for scenario one, with added contours. Color indicates the estimated component membership and shape indicates the known membership. Simulated outliers are depicted by the + symbol.
4.2.2 Real data

In this section we explore the effectiveness of the contaminated SAL mixture model in two ways, as appropriate for specific data sets. One option is to try detecting any natural occurring outliers in the data. Another method is to introduce an artificial outlier and attempt to detect it. To this end, we pick an observation and add to it suitable multiples of the first and/or second principal component vectors. The PCs are determined using robust principal component analysis (ROBPCA; Hubert et al., 2005), as implemented in the R package rrcoov (Todorov and Filzmoser, 2009).

We compare our results with those obtained via the trimmed clustering approach of Fritz et al. (2012), as implemented in the R package tclust. With this procedure, a proportion, \( \alpha \), of the most outlying observations is trimmed, by considering Gaussian mixture components. In particular, the tclust() function implements several algorithms aimed at maximizing the likelihood of the spurious outliers model introduced by Gallegos (2002) and Gallegos and Ritter (2005). For \( f(\cdot|\mu, \Sigma) \), denoting the probability density function of the \( p \)-variate normal distribution with mean \( \mu \) and covariance matrix \( \Sigma \), this model is defined through likelihoods of the form

\[
\left[ \prod_{j=1}^{k} \prod_{i \in R_j} f(x_i | \mu_j, \Sigma_j) \right] \left[ \prod_{i \in R_0} g_i(x_i) \right].
\]  

(4.27)

Here \( \{R_0, \ldots, R_k\} \) represents a partition of the set of indices \( \{1, 2, \ldots, n\} \) such that \( \#R_0 = \lceil n\alpha \rceil \), and \( R_0 \) are the indices of the “non-regular” observations generated by other (not necessarily normal) probability density functions \( g_i \). “Non-regular” observations can then be considered “outliers” if certain assumptions for the \( g_i \) are made (see Gallegos, 2002; Gallegos and Ritter, 2005).
The comparison between our proposed contaminated SAL mixtures and the \texttt{tclust} procedure might not seem appropriate at first, especially when analyzing asymmetric data. However, \texttt{tclust} is the most closely aligned comparator available at the moment, given the setup of outlier detection for model-based clustering we consider herein. Note that the plots shown in Figure 4.5, Figure 4.7 and Figure 4.9 were produced using the functionality from \texttt{tclust}.

\textbf{Australian Institute of Sport data}

\textit{Cook and Weisberg} (1994) provided eleven body and blood measurements collected from athletes at the Australian Institute of Sport (AIS; Table 2.8). This data set was discussed in detail in Chapter 2. We introduce an artificial outlier by adding a linear combination of the first two PC vectors to the first observation. The outlier is detected correctly by the contaminated SAL model, as indicated in Table 4.5. In this case, the model-based clustering resulted in an ARI of 0.8932, which is similar to the ARI = 0.8935 obtained with \texttt{tclust}, as shown in Table 4.6.

Table 4.5: Clustering results for the best model, in terms of BIC, fitted to the AIS data using a contaminated SAL mixture.

<table>
<thead>
<tr>
<th></th>
<th>Fitted</th>
<th>Known</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>99</td>
<td>5</td>
</tr>
<tr>
<td>Male</td>
<td>0</td>
<td>97</td>
</tr>
<tr>
<td>Outliers</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 4.4 represents the AIS data projected onto its first three PCs, and
Table 4.6: Clustering results for the model fitted to the AIS data using \texttt{tclust} with $\alpha = 0.00495$.

<table>
<thead>
<tr>
<th></th>
<th>Known</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fitted</td>
<td>Female</td>
<td>Male</td>
</tr>
<tr>
<td>Female</td>
<td>97</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>3</td>
<td>99</td>
<td></td>
</tr>
<tr>
<td>Trimmed</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

our correctly classified artificial outlier being drawn with a vertical link to the underlying plane. Figure 4.5 represents the first two discriminant coordinates for the AIS data obtained with \texttt{tclust}, with colour and shape indicating the estimated cluster membership.

![Figure 4.4](image)

Figure 4.4: Principal components for the AIS data indicating the cluster structure and artificial outlier. Colour represents known component membership.
Wine data

Forina et al. (1986) recorded several chemical and physical properties for three types of Italian wines, namely Barolo, Grignolino, and Barbera (Table 3.10). The details of these data were examined in Chapter 3. Similar to the AIS data, we contaminate the first observation by adding a linear combination of the first two PCs to it. The artificial outlier is detected by the model, and the clustering results from Table 4.7 yield an ARI = 0.9226. This is much higher than the ARI = 0.2231 we obtained with tclust (Table 4.8).

Figure 4.6 shows the wine data projected onto its first three PCs. The artificial outlier which was classified correctly is represented with a vertical link to its underlying plane. Figure 4.7 represents the first two discriminant coordinates for
Table 4.7: Model-based clustering results fitted to the wine data using a contaminated SAL mixture.

<table>
<thead>
<tr>
<th>Known</th>
<th>Fitted</th>
<th>Barolo</th>
<th>Grignolino</th>
<th>Barbera</th>
<th>Outliers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barolo</td>
<td>57</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Grignolino</td>
<td>1</td>
<td>67</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Barbera</td>
<td>0</td>
<td>3</td>
<td>48</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Outliers</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4.8: Clustering results for the model fitted to the AIS data using \texttt{tclust} with $\alpha = 0.0056$.

<table>
<thead>
<tr>
<th>Known</th>
<th>Fitted</th>
<th>Barolo</th>
<th>Grignolino</th>
<th>Barbera</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barolo</td>
<td>38</td>
<td>24</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Grignolino</td>
<td>21</td>
<td>34</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Barbera</td>
<td>0</td>
<td>12</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td>Trimmed</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

the wine data obtained with \texttt{tclust}, with colour and shape indicating the estimated cluster membership.

Figure 4.6: Principal components for the wine data indicating the cluster structure and artificial outlier. Colour represents known component membership.
Yeast data

Horton and Nakai (1996) provided data on cellular localization of yeast proteins (Table 2.5) and we analyzed them in Chapter 2. On closer inspection, we observe that the “Score of ALOM membrane spanning region prediction program (alm)” shows two outliers with values of 1 at the CYT site. Thus our aim here is to try and detect these natural occurring outliers using the contaminated SAL model proposed in (4.3).

The outliers are detected correctly by the contaminated SAL model, as indicated in Table 4.9. In this case, the model-based clustering resulted in an ARI of 0.8219, which is much better than the ARI = −0.0586 for the clustering results obtained with tclust, which appear in Table 4.10.
Table 4.9: Clustering results for the best model, in terms of BIC, fitted to the yeast data using a contaminated SAL mixture.

<table>
<thead>
<tr>
<th></th>
<th>Fitted</th>
<th>Known</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYT</td>
<td>451</td>
<td>16</td>
</tr>
<tr>
<td>ME3</td>
<td>10</td>
<td>147</td>
</tr>
<tr>
<td>Outliers</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4.10: Clustering results for the model fitted to the yeast data using \texttt{tclust} with $\alpha = 0.0032$.

<table>
<thead>
<tr>
<th></th>
<th>Fitted</th>
<th>Known</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYT</td>
<td>427</td>
<td>163</td>
</tr>
<tr>
<td>ME3</td>
<td>34</td>
<td>0</td>
</tr>
<tr>
<td>Trimmed</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

In order to better visualize the yeast data, we project the observations onto the first three principal components. Figure 4.8 depicts the PCs and the two outlying observations are marked with vertical links to the underlying plane. Figure 4.9 represents the first two discriminant coordinates for the wine data obtained with \texttt{tclust}, with colour and shape indicating the estimated cluster membership.
Figure 4.8: Principal components for the yeast data indicating the cluster structure and natural outliers. Colour represents known component membership.

Figure 4.9: Scatterplot of the first two discriminant coordinates for the yeast data obtained with \texttt{tclust}. Colour and shape represent estimated cluster membership.
Olive oil data

Forina and Tiscornia (1982) analyzed data on the percentage composition of eight fatty acids (Table 4.11) found by lipid fraction of 572 Italian olive oils. The data come from three regions: Southern Italy, Sardinia, and Northern Italy. Each region is further divided into areas, giving a total of nine classes.

<table>
<thead>
<tr>
<th>Palmitic</th>
<th>Palmitoleic</th>
<th>Stearic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oleic</td>
<td>Linoleic</td>
<td>Linolenic</td>
</tr>
<tr>
<td>Arachidic</td>
<td>Eicosenoic</td>
<td></td>
</tr>
</tbody>
</table>

Our aim is to cluster the data into the nine areas, while investigating the presence of natural outliers. To better understand the data, we first ran a robust principal component analysis and noted the presence of three natural outlying observations, as depicted in Figures 4.10 and 4.11.

Model-based clustering with contaminated SAL mixtures resulted in an ARI = 0.7659 (Table 4.2.2), however no natural outliers were detected. Similarly, clustering with contaminated SN mixtures gave an ARI = 0.7467, and 3 natural outliers detected. Clustering with tclust and 0.0052 of observations trimmed, equivalent to 3 outliers, leads to ARI = 0.7015.

Table 4.13 outlines the specific estimates for $\rho$ and $\lambda$ obtained in the fitting of contaminated SAL mixtures to the real data.
Figure 4.10: Figure depicts the diagnostic plot from the ROBPCA procedure for the olive oil data, indicating possible outliers.

Figure 4.11: Figure shows the pairs plot of the data projected on the first four robust PCs. Colour indicates known cluster membership.
Table 4.12: Clustering results for the best model, in terms of BIC, fitted to the olive oil data using a contaminated SAL mixture.

<table>
<thead>
<tr>
<th>Known</th>
<th>Fitted</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>54</td>
<td>4</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>2</td>
<td>185</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>50</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>30</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>48</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>48</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>46</td>
</tr>
<tr>
<td>Outliers</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4.13: Summary of estimates for $\rho$ and $\lambda$ obtained while fitting contaminated SAL mixtures to the real data.

<table>
<thead>
<tr>
<th>Data</th>
<th>$\hat{\rho}$</th>
<th>$\hat{\lambda}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comp. 1</td>
<td>55.5937</td>
<td>0.9867</td>
</tr>
<tr>
<td>Comp. 2</td>
<td>37.4789</td>
<td>0.9778</td>
</tr>
<tr>
<td>Wine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comp. 1</td>
<td>25.0433</td>
<td>0.9718</td>
</tr>
<tr>
<td>Comp. 2</td>
<td>40.9009</td>
<td>1</td>
</tr>
<tr>
<td>Comp. 3</td>
<td>16.0155</td>
<td>0.9808</td>
</tr>
<tr>
<td>Yeast</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comp. 1</td>
<td>30.7931</td>
<td>0.9662</td>
</tr>
<tr>
<td>Comp. 2</td>
<td>17.7755</td>
<td>0.9723</td>
</tr>
<tr>
<td>Olive oil</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comp. 1</td>
<td>1.1622</td>
<td>0.9072</td>
</tr>
<tr>
<td>Comp. 2</td>
<td>1.2565</td>
<td>0.8697</td>
</tr>
<tr>
<td>Comp. 3</td>
<td>1.3253</td>
<td>0.8393</td>
</tr>
<tr>
<td>Comp. 4</td>
<td>1.7960</td>
<td>0.9990</td>
</tr>
<tr>
<td>Comp. 5</td>
<td>1.4115</td>
<td>0.8836</td>
</tr>
<tr>
<td>Comp. 6</td>
<td>1.6746</td>
<td>0.9494</td>
</tr>
<tr>
<td>Comp. 7</td>
<td>1.3027</td>
<td>0.9601</td>
</tr>
<tr>
<td>Comp. 8</td>
<td>1.4417</td>
<td>0.8906</td>
</tr>
<tr>
<td>Comp. 9</td>
<td>1.3457</td>
<td>0.9784</td>
</tr>
</tbody>
</table>
4.3 Summary

This chapter introduced a method for detecting outliers for model-based clustering using mixtures of shifted asymmetric Laplace distributions. The main idea was to introduce a contamination factor which increased the dispersion of the fitted distribution by altering the skewness and covariance parameters. The extension of the procedure to model-based classification was also outlined. The ECM algorithm was employed to classify of an observation $x_i$ by first determining its group membership, and then establishing whether it was an outlier within that group.

The method was applied to simulated and real data where it yielded good results when outliers (artificial or natural) were present. The procedure was benchmarked against trimmed clustering, and it outperformed the latter in two out of the three real data sets considered. One limitation of our current approach was the non-automated way of selecting the initial value for the contamination factors. Future work could focus on developing ways of computing suitable initial parameters and incorporating shifted asymmetric Laplace analogues of the PMCGD models.
Chapter 5

Outlier Detection for Clustering using Mixtures of Contaminated Skew Normal Distributions

In this chapter we focus on detecting outliers for model-based clustering using the class of multivariate skew normal distributions (SN) proposed by Sahu et al. (2003). SN mixtures are useful in practical applications where asymmetric data are present. In these cases, SN mixtures are more appropriate than Gaussian mixtures because they do not overfit the data by including additional components to capture skewness. We employ a notation akin to Lin (2009), where a flexible SN mixture modelling framework was derived. An approach similar to that of Punzo and McNicholas (2013) for contaminated Gaussian mixtures is developed for clustering with contaminated SN mixtures.

5.1 Methodology

A random vector $X$ that follows a $p$-dimensional skew normal distribution with a $p \times 1$ location vector $\mu$, a $p \times p$ positive definite scale covariance matrix $\Sigma$,
and a $p \times p$ skewness matrix $A$ has the density function

$$f_{SN}(x|\mu, \Sigma, A) = 2^p \phi_p(x|\mu, \Omega) \Phi_p(A^\top \Omega^{-1}(x - \mu)|\Delta), \quad (5.1)$$

with $\Omega = \Sigma + AA^\top$ and $\Delta = (I_p + A^\top \Sigma^{-1} A)^{-1} = I_p - A^\top \Omega^{-1} A$, where $I_p$ indicates a $p \times p$ identity matrix. We will assume that $A = \text{diag}(\alpha)$ is a diagonal matrix so that the covariance structure of $X$ is not affected by skewness. Also, $\phi_p(\cdot)$ denotes the probability density function of $N_p(\mu, \Sigma)$ and $\Phi_p(\cdot)$ gives the cumulative density function of $N_p(0, \Sigma)$.

Arelanno-Valle et al. (2007) showed that the random variable $X$ from (5.1) can be generated through the relationship

$$X = \mu + A\tau + U, \quad (5.2)$$

where $\tau$ is independently distributed as a standard half-normal distribution $HN_p(0, I_p)$, and $U$ is independently distributed as a standard normal distribution $N_p(0, \Sigma)$. It follows that $X|\tau \sim N_p(\mu + A\tau, \Sigma)$ and the density in (5.1) can be rewritten as

$$f_{SN}(x|\mu, \Sigma, A) = |2\pi\Sigma|^{-1/2} \exp\{-\frac{1}{2}(x - (\mu + A\tau))^\top \Sigma^{-1}(\mu + A\tau)\}. \quad (5.3)$$

Furthermore, Lin (2009) showed that $\tau|X$ can be represented via a truncated normal (TN) distribution

$$\tau|X \sim TN_p(A^\top \Omega^{-1}(x - \mu), \Delta; \mathbb{R}^p_+), \quad (5.4)$$

where $\Omega = \Sigma + AA^\top$, $\Delta = (I_p + A^\top \Sigma^{-1} A)^{-1}$, and $\mathbb{R}^p_+ = \{y = (y_1, \ldots, y_p)^\top \in \mathbb{R}^p| y_i > 0, i = 1, \ldots, p\}$. 
The TN distribution is used here because of its computational advantages, and further details on its framework and properties can be found, for example, in Tallis (1961) and Lin (2009).

We follow the approach for contaminated Gaussian mixtures from Punzo and McNicholas (2013) to develop the framework for clustering with contaminated SN mixtures. Similar to the case of contaminated SAL mixtures discussed in Chapter 4, we consider the contamination scheme. To this end, we employ the contamination scheme where $\Sigma_g$ gets scaled to $\rho_g \Sigma_g$ and $\alpha_g$ gets scaled to $\sqrt{\rho_g} \alpha_g$, with $\rho_g > 0$ denoting the contamination factor. This leads to the following contaminated SN distribution

$$C_{SN}(x|\theta) = \lambda_g f_{SN}(x_i|\mu_g, A_g, \Sigma_g) + (1 - \lambda_g) f_{SN}(x_i|\mu_g, \sqrt{\rho_g} A_g, \rho_g \Sigma_g), \quad (5.5)$$

where $\lambda_g \in [0, 1]$ denotes the proportion of outliers. The corresponding contaminated SN mixture model is defined by

$$f_{SN}(x|\theta) = \sum_{g=1}^{G} \pi_g \{ \lambda_g f_{SN}(x_i|\mu_g, \alpha_g, \Sigma_g) + (1 - \lambda_g) f_{SN}(x_i|\mu_g, \sqrt{\rho_g} \alpha_g, \rho_g \Sigma_g) \}. \quad (5.6)$$

There are two sources of missing data: the group memberships, $z_{ig}$, and the fact that we do not know whether an observation in group $g$ is “good” or “bad” (Punzo and McNicholas, 2013). We denote this second source of missing data by $v_{ig}$ so that $v_{ig} = 1$ if observation $i$ in group $g$ is “good”, and $v_{ig} = 0$ if observation $i$ in group $g$ is “bad”.

We note that mixtures of contaminated SN distributions for the univariate case have been proposed by Cabral et al. (2012). The contaminated model introduced
there has the form

\[
\text{SCN}_p(x|\rho, \lambda, \alpha, \mu, \Sigma) = 2\{\lambda \phi_p(x|\mu, \rho^{-1}\Sigma)\Phi(\rho^{1/2}\alpha^\top\Sigma^{-1/2}(x - \mu)) + (1 - \lambda) \phi_p(x|\mu, \Sigma)\Phi(\alpha^\top\Sigma^{-1/2}(x - \mu))\},
\]

(5.7)

where \(\phi_p(\cdot)\) stands for the density of the \(p\)-variate normal distribution, and \(\Phi(\cdot)\) represents the distribution function of the standard univariate normal distribution. Also, \(\lambda\) and \(\rho\) denote the proportion of outliers and contamination factor, respectively.

In contrast to Punzo and McNicholas (2013), we do not consider covariance decomposed structures for the model in (5.5). Instead, we focus on the most general case by letting the covariance vary between components, which is similar to the Gaussian mixture model VVV of Celeux and Govaert (1995). Recall that this scenario was also used for the case of contaminated SAL mixtures discussed in Chapter 4.

We can visualize the effect of our contamination scheme on a typical SN mixture. Figure 5.1 depicts this for a SN mixture with \(\alpha = (1.5, 1.5)^\top\), \(\mu = (0, 0)^\top\), \(\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\), and a contamination factor of \(\rho = 3\). By using the contamination scheme described above, the mode in the contaminated SN density is the same as in the original density. This can be seen in Figure 5.1.

The complete-data likelihood for the mixture of contaminated SN distributions defined in (5.5) is

\[
\mathcal{L}(\theta|x) = \prod_{i=1}^n \prod_{g=1}^G \pi_g \left[ \lambda_j f_{\text{SN}}(x_i|\mu_g, A_g, \Sigma_g) \right]^{z_{ig}} \times \left[ (1 - \lambda_j) f_{\text{SN}}(x_i|\mu_g, \sqrt{\rho_g A_g, \rho_g \Sigma_g}) \right]^{(1-v_{ig})}.
\]

(5.8)
Figure 5.1: Contour plot of a contaminated SN mixture superimposed on its uncontaminated SAL mixture counterpart.

Then, the complete-data log-likelihood corresponding to (5.8) becomes

\[
\ln \bigl( \mathbf{\theta} | \mathbf{x} \bigr) = \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} v_{ig} \log(\pi_g) + \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} v_{ig} \log(\lambda_g) \\
+ \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} v_{ig} \log(f_{SN}(\mathbf{x}_i | \mathbf{\mu}_g, \mathbf{A}_g, \mathbf{\Sigma}_g)) \\
+ \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} (1 - v_{ig}) \log(\pi_g) + \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} (1 - v_{ig}) \log(1 - \lambda_g) \\
+ \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} (1 - v_{ig}) \log(f_{SN}(\mathbf{x}_i | \mathbf{\mu}_g, \sqrt{\rho_g} \mathbf{A}_g, \rho_g \mathbf{\Sigma}_g)), \tag{5.9}
\]
and, ignoring additive constants, the above simplifies to

\[
l(\theta|x) = \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} v_{ig} \log(\pi_g) + \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} v_{ig} \log(\lambda_g) \\
+ \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} v_{ig} \left\{ -\frac{1}{2} \log |\Sigma_g^{-1}| \right\} \\
- \frac{1}{2} (x_i - \mu_g - A_g \tau_{ig})^\top \Sigma_g^{-1} (x_i - \mu_g - A_g \tau_{ig}) - \frac{1}{2} \tau_{ig}^\top \tau_{ig} \right\} \\
+ \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} (1 - v_{ig}) \log(\pi_g) + \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} (1 - v_{ig}) \log(1 - \lambda_g) \\
+ \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} (1 - v_{ig}) \left\{ -\frac{1}{2} \log |(\rho_g \Sigma_g)^{-1}| \right\} \\
- \frac{1}{2} (x_i - \mu_g - \sqrt{\rho_g} A_g \tilde{\tau}_{ig})^\top (\rho_g \Sigma_g)^{-1} (x_i - \mu_g - \sqrt{\rho_g} A_g \tilde{\tau}_{ig}) - \frac{1}{2} \tilde{\tau}_{ig}^\top \tilde{\tau}_{ig} \right\},
\]

(5.10)

where \( \tilde{\tau} \) is the analogue of \( \tau \) when \( \Sigma_g \longrightarrow \rho_g \Sigma_g \) and \( A_g \longrightarrow \sqrt{\rho_g} A_g \).

As described in Chapter 4, we need to address the issue of identifiability of mixtures of contaminated skew-normal distributions. Finite mixture of skew-normal distributions are identifiable in general (Castro et al., 2011; Azzalini and Dalla Valle, 1996). Proving identifiability in our context is similar to the approach of Punzo and McNicholas (2013) and follows the steps outlined in Proposition 4.1.1.

The ECM algorithm is used to find the maximum likelihood estimates in (5.10). The E-step of the algorithm relies on the computation of the Q-function, 

\[
Q(\theta|\hat{\theta}) = E[l(\theta|x)|x, \hat{\theta}].
\]

This Q-function denotes the conditional expectation of (5.10) given the observed data \( x \) and the current estimated parameters \( \hat{\theta} \). The conditional expectations of the terms \(-\frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} v_{ig} \tau_{ig}^\top \tau_{ig} \) and \(-\frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} (1 - v_{ig}) \tilde{\tau}_{ig}^\top \tilde{\tau}_{ig} \) can be omitted because they do not contain any parameters. For the com-
putations involved in the Q-function we adopt the following notation

\[
E[\tau_{ig}|x_i, z_{ig} = 1] = \eta_{ig}, \quad E[\tau_{ig}^T \tau_{ig}|x_i, z_{ig} = 1] = \psi_{ig} \quad \text{and} \quad E[\tilde{\tau}_{ig}|x_i, z_{ig} = 1] = \tilde{\eta}_{ig}, \quad E[\tilde{\tau}_{ig}^T \tilde{\tau}_{ig}|x_i, z_{ig} = 1] = \tilde{\psi}_{ig},
\]

(5.11)

where \(\eta_{ig}, \tilde{\eta}_{ig}, \psi_{ig}, \tilde{\psi}_{ig}\) are all implicit functions of the parameters \((\mu_g, \Sigma_g, A_g)\) and \(\rho_g\) where required, and can be calculated numerically using the properties of the TN distribution, as outlined in Lin (2009). R code that implements these TN calculations was kindly provided by Professor Tsung I. Lin. The Q-function for the contaminated SN mixture model is defined as

\[
Q = \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} v_{ig} \left\{ \log(\lambda_g) + \log(\pi_g) + \frac{1}{2} \log |\Sigma_g^{-1}| \right. \\
- \frac{1}{2} (x_i - \mu_g - A_g \eta_{ig})^T \Sigma_g^{-1} (x_i - \mu_g - A_g \eta_{ig}) \\
- \frac{1}{2} \text{tr}(\Sigma_g^{-1} A_g (\psi_{ig} - \eta_{ig} \eta_{ig}^T) A_g^T) \\
+ \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} (1 - v_{ig}) \left\{ \log(1 - \lambda_g) + \log(\pi_g) + \frac{1}{2} \log |(\rho_g \Sigma_g)^{-1}| \right. \\
- \frac{1}{2} (x_i - \mu_g - \sqrt{\rho_g} A_g \tilde{\eta}_{ig})^T (\rho_g \Sigma_g)^{-1} (x_i - \mu_g - \sqrt{\rho_g} A_g \tilde{\eta}_{ig}) \\
- \frac{1}{2} \text{tr}((\rho_g \Sigma_g)^{-1} \sqrt{\rho_g} A_g (\tilde{\psi}_{ig} - \tilde{\eta}_{ig} \tilde{\eta}_{ig}^T) (\sqrt{\rho_g} A_g)^T) \right\}.
\]

(5.12)

At the E-step of the ECM algorithm we also update the \(z_{ig}\) with

\[
\hat{z}_{ig} = \frac{\pi_g C_{SN}(x_i|\theta_g)}{\sum_{g=1}^{G} \pi_g C_{SN}(x_i|\theta_g)} = \frac{\pi_g [\lambda_g f_{SN}(x_i|\mu_g, \alpha_g, \Sigma_g) + (1 - \lambda_g) f_{SN}(x_i|\mu_g, \sqrt{\rho_g} \alpha_g, \rho_g \Sigma_g)]}{\sum_{g=1}^{G} \pi_g [\lambda_g f_{SN}(x_i|\mu_g, \alpha_g, \Sigma_g) + (1 - \lambda_g) f_{SN}(x_i|\mu_g, \sqrt{\rho_g} \alpha_g, \rho_g \Sigma_g)]},
\]

(5.13)
and similarly update the $v_{ig}$ with
\[\hat{v}_{ig} = \frac{\lambda_g f_{SN}(x_i | \vartheta_g)}{C_{SN}(x_i | \vartheta_g)} = \frac{\lambda_g f_{SN}(x_i | \mu_g, \alpha_g, \Sigma_g)}{\lambda_g f_{SN}(x_i | \mu_g, \alpha_g, \Sigma_g) + (1 - \lambda_g) f_{SN}(x_i | \mu_g, \sqrt{\rho_g} \alpha_g, \rho_g \Sigma_g)}. \tag{5.14}\]

At the first CM-step of the ECM algorithm we update $\pi_g$ and $\lambda_g$ with
\[\hat{\pi}_g = \sum_{i=1}^n \hat{z}_{ig}, \quad \hat{\lambda}_g = \sum_{i=1}^n \hat{z}_{ig} v_{ig}, \tag{5.14}\]
respectively. Then we use the Q-function to compute parameter updates for $\mu_g, \Sigma_g, \alpha_g$ and $\rho_g$. The mathematical derivations for obtaining the updates appear in Section A.2 of the Appendix A. The calculations for $\mu_g$ yield the update
\[\hat{\mu}_g = \frac{\sum_{i=1}^n z_{ig} v_{ig} (x_i - A_g \eta_{ig}) + \sum_{i=1}^n \hat{z}_{ig} (1 - v_{ig}) \left( x_i - \sqrt{\rho_g} A_g \tilde{\eta}_{ig} \right)}{\sum_{i=1}^n \hat{z}_{ig} v_{ig} + \sum_{i=1}^n \hat{z}_{ig} (1 - v_{ig})}, \tag{5.15}\]
and the update for $\Sigma_g$ becomes
\[\hat{\Sigma}_g = \frac{1}{\sum_{i=1}^n \hat{z}_{ig}} \left[ \sum_{i=1}^n \hat{z}_{ig} v_{ig} \left\{ (x_i - \mu_g - A_g \eta_{ig}) (x_i - \mu_g - A_g \eta_{ig})^\top \right. \right.
\\\\ \left. + A_g (\psi_{ig} - \eta_{ig} \eta_{ig}^\top) A_g^\top \right\} + \sum_{i=1}^n \hat{z}_{ig} (1 - v_{ig}) \left\{ \frac{1}{\rho_g} (x_i - \mu_g - \sqrt{\rho_g} A_g \tilde{\eta}_{ig}) \times (x_i - \mu_g - \sqrt{\rho_g} A_g \tilde{\eta}_{ig})^\top + A_g (\tilde{\psi}_{ig} - \tilde{\eta}_{ig} \tilde{\eta}_{ig}^\top) A_g^\top \right\} \right]. \tag{5.16}\]

Recalling that $A_g$ is the diagonal matrix version of $\alpha_g$, we obtain the update for skewness
\[\hat{\alpha}_g = \left[ \sum_{i=1}^n \hat{z}_{ig} v_{ig} (\Sigma_g^{-1} \odot \psi_{ig}) + \sum_{i=1}^n \hat{z}_{ig} (1 - v_{ig}) \left( \Sigma_g^{-1} \odot \tilde{\psi}_{ig} \right) \right]^{-1}
\times \left[ \sum_{i=1}^n \hat{z}_{ig} v_{ig} (\Sigma_g^{-1} \odot \eta_{ig} (x_i - \mu_g)^\top) + \sum_{i=1}^n \hat{z}_{ig} (1 - v_{ig}) \rho_g \left( \Sigma_g^{-1} \odot \tilde{\eta}_{ig} (x_i - \mu_g)^\top \right) \right] \mathbf{1}_p, \tag{5.17}\]
where the operator $\odot$ denotes the Hadamard (elementwise) product of two matrices of the same dimension.

At the second CM-step of the ECM algorithm we update the contamination factor $\rho_g$. Looking at the Q-function in (5.12) we observe that there are two possible ways of updating $\rho_g$.

1. Use the R function `optim()` to find a maximum for $Q$, as $\rho_g$ varies in the interval $(1, \rho_{g,\text{max}})$, by considering

$$Q = \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig}(1 - v_{ig}) \left\{ -\frac{p}{2} \log(\rho_g) - \frac{1}{2\rho_g} (x_i - \mu_g)^\top \Sigma_g^{-1} (x_i - \mu_g) + \frac{1}{2\sqrt{\rho_g}} \left[ (x_i - \mu_g)^\top \Sigma_g^{-1} (A_g \tilde{\eta}_{ig}) + (A_g \tilde{\eta}_{ig})^\top \Sigma_g^{-1} (x_i - \mu_g) \right] \right\} + \text{terms independent of } \rho_g.$$  

(5.18)

2. Solve the equation $\frac{\partial Q}{\partial \rho_g} = 0$ for $\rho_g$ directly,

$$0 = \rho_g \left( -p \sum_{i=1}^{n} z_{ig}(1 - v_{ig}) \right) + \sum_{i=1}^{n} z_{ig}(1 - v_{ig}) (x_i - \mu_g)^\top \Sigma_g^{-1} (x_i - \mu_g) - \rho_g^{1/2} \sum_{i=1}^{n} z_{ig}(1 - v_{ig}) \left[ (x_i - \mu_g)^\top \Sigma_g^{-1} (A_g \tilde{\eta}_{ig}) + (A_g \tilde{\eta}_{ig})^\top \Sigma_g^{-1} (x_i - \mu_g) \right],$$  

(5.19)

which is a quadratic equation of the form $S\rho_g + T\rho_g^{1/2} + U = 0$ that can be easily solved for $\rho_g$ (where $S$, $T$, $U$ are constants).

Similar to the case of contaminated SAL mixtures discussed in Chapter 4, it was worthwhile to implement both methods of computing the $\rho_g$ and alternate between the two in case of numerical errors.
Within the model-based clustering framework, we allow the number of components, $G$ to vary between $G = 2$ and $G = \text{no. known components} + 2$, and the best model is selected using the BIC (1.17). In the case of contaminated SN mixtures, similar to their contaminated SAL analogue (Table 4.1), we have $\text{BIC} = 2l(x, \hat{\theta}) - (3G - 1 + 2Gp + Gp(p + 1)/2) \log(n)$.

By using an analogous process, the model-based clustering framework for contaminated SAL mixtures can be applied to model-based classification, and the details regarding this extension and parameter updates appear in Section A.2.2 of the Appendix A.

## 5.2 Applications

### 5.2.1 Simulated data

We generated $n = 180$ data points from a SN mixture with two components of equal size ($n_1 = n_2 = 90$) using the relationship $X = \mu + A\tau + U$, as discussed in (5.2). The specific parameters used to simulate these data appear in Table 5.1. Then, to test our contaminated SN model (5.6), we incorporate artificial “outliers”.

To understand the performance of the fitted contaminated mixture, we use the values of $v_{ig}$ to determine if a point should be assigned to the “good” or “bad” categories. This resulted in 3 fitted groups, namely Component 1, Component 2, and Outliers. The observation assignments to these groups were compared against the known labels of the simulated data. Table 5.2 shows the best fit observed among the
Table 5.1: Parameters used to simulate data for a two component SN mixture.

<table>
<thead>
<tr>
<th>Component 1</th>
<th>Component 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_1 = \begin{pmatrix} 1 \ 1 \end{pmatrix}$</td>
<td>$\alpha_2 = \begin{pmatrix} -1 \ 1 \end{pmatrix}$</td>
</tr>
<tr>
<td>$\mu_1 = \begin{pmatrix} 1 \ -2 \end{pmatrix}$</td>
<td>$\mu_2 = \begin{pmatrix} 5 \ -5 \end{pmatrix}$</td>
</tr>
<tr>
<td>$\Sigma_1 = \begin{bmatrix} 1 &amp; 0.5 \ 0.5 &amp; 1 \end{bmatrix}$</td>
<td>$\Sigma_2 = \begin{bmatrix} 1 &amp; 0 \ 0 &amp; 1 \end{bmatrix}$</td>
</tr>
</tbody>
</table>

100 runs, with an ARI of 0.9590. Overall, we obtain an average ARI of 0.8955 with a standard deviation of 0.0349.

Table 5.2: Clustering results for the best model, in terms of ARI, fitted to the simulated SN data.

<table>
<thead>
<tr>
<th>Fitted</th>
<th>Known</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Comp. 1</td>
</tr>
<tr>
<td>Comp. 1</td>
<td>89</td>
</tr>
<tr>
<td>Comp. 2</td>
<td>1</td>
</tr>
<tr>
<td>Outliers</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 5.2 plots the results of the clustering with contaminated SN mixtures for the simulated data. Colour indicates the estimated group membership and shape indicates the known membership.

5.2.2 Real data

To examine the performance of the contaminated SN mixture model for real data, we focus on two possible avenues for using the algorithm.

1. Attempt to detect any natural occurring outliers in the data.
Figure 5.2: Scatterplot of simulated data for scenario one, with added contours. Color indicates the estimated component membership and shape indicates the known membership.

2. Introduce an artificial outlier and attempt to detect it. For this method, choose an observation and add to it a suitable linear combination of the first and/or second PC vectors. The PCs are determined using robust principal component analysis (ROBPCA; Hubert et al., 2005), as implemented in the R package 
\texttt{rrcov} (Todorov and Filzmoser, 2009).

Similar to the real data analysis in Chapter 4, we compare our clustering results with those obtained with the trimmed clustering approach of Fritz et al. (2012) which is implemented in the R package \texttt{tclust}. Note that the plots shown in Figure 5.4 and Figure 5.6 were produced using the functionality from \texttt{tclust}.
Swiss banknotes data

Flury and Riedwyl (1988) present six measurements taken from Swiss banknotes (Table 2.2), which we analyzed in Chapter 2 and 3. The aim here is to detect any natural outliers using the contaminated SN model (5.5). By projecting the data onto the first three PCs, as depicted in Figure 5.3, we can visualize the clustering structure and the two outliers found by the algorithm. The latter are represented by points with vertical links to the underlying plane.

The clustering results from Table 5.3 show that one outlier is detected in each class of banknotes (genuine and counterfeit). The performance, as measured by ARI = 0.9602, is higher than the ARI = 0.6711 obtained with tclust, for the clustering in Table 5.4. Figure 5.4 represents the first two discriminant coordinates for the bank data obtained with tclust, with colour and shape indicating the estimated cluster membership.

Table 5.3: Clustering results for the best model, in terms of BIC, fitted to the bank data using a contaminated SN mixture.

<table>
<thead>
<tr>
<th></th>
<th>Known</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fitted</td>
<td>Genuine</td>
</tr>
<tr>
<td>Genuine</td>
<td>98</td>
</tr>
<tr>
<td>Counterfeit</td>
<td>1</td>
</tr>
<tr>
<td>Outliers</td>
<td>1</td>
</tr>
</tbody>
</table>
Table 5.4: Clustering results for the model fitted to the bank data using \texttt{tclust} with \( \alpha = 0.01 \).

<table>
<thead>
<tr>
<th>Known</th>
<th>Fitted</th>
<th>Genuine</th>
<th>Counterfeit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genuine</td>
<td>98</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>Counterfeit</td>
<td>0</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>Trimmed</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Figure 5.3: Principal components for the bank data indicating the cluster structure and natural outliers. Colour represents known component membership.
Figure 5.4: Scatterplot of the first two discriminant coordinates for the bank data obtained with \texttt{tclust}. Colour and shape represent estimated cluster membership.
Female voles data

Flury (1997, Table 5.3.7) discuss seven measurements of female voles from two species (*Microtus californicus* and *Microtus ochrogaster*, Table 3.5) originally studied by Airoldi and Hoffmann (1984). These data were also analyzed in Chapter 3. We contaminate the first observation with a linear combination of the first two PC vectors. This artificial outlier is correctly detected by our method, while the best fitted model (Table 5.5) results in an ARI = 0.9540. This indicates a better clustering than the one obtained with *tclust* (Table 5.6) which produced an ARI = 0.7187.

Figure 5.5 depicts the voles data projected onto the first three PCs, with observations coloured according to component membership. The artificial outlier is represented by a point with a vertical link to the underlying plane. Figure 5.6 represents the first two discriminant coordinates for the voles data obtained with *tclust*, with colour and shape indicating the estimated cluster membership.

Table 5.5: Clustering results for the best model, in terms of BIC, fitted to the voles data using a contaminated SN mixture.

<table>
<thead>
<tr>
<th></th>
<th>Known</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>M. Californicus</em></td>
<td><em>M. Ochrogaster</em></td>
<td>Outliers</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>M. Californicus</em></td>
<td>39</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>M. Ochrogaster</em></td>
<td>1</td>
<td>45</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outliers</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.7 outlines the specific estimates for $\rho$ and $\lambda$ obtained in the fitting of contaminated SN mixtures to the real data.
Table 5.6: Clustering results for the model fitted to the voles data using \texttt{tclust} with $\alpha = 0.0102$.

<table>
<thead>
<tr>
<th>Fitted</th>
<th>Known</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M. \text{Californicus}$</td>
<td>40 6</td>
</tr>
<tr>
<td>$M. \text{Ochrogaster}$</td>
<td>0 39</td>
</tr>
<tr>
<td>Trimmed</td>
<td>1 0</td>
</tr>
</tbody>
</table>

Figure 5.5: Principal components for the voles data indicating the cluster structure and artificial outlier. Colour represents known component membership.

Table 5.7: Summary of estimates for $\rho$ and $\lambda$ obtained while fitting contaminated SN mixtures to the real data.

<table>
<thead>
<tr>
<th>Data</th>
<th>$\hat{\rho}$</th>
<th>$\hat{\lambda}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swiss banknotes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comp. 1</td>
<td>11.4547</td>
<td>0.9952</td>
</tr>
<tr>
<td>Comp. 2</td>
<td>8.6721</td>
<td>0.9974</td>
</tr>
</tbody>
</table>

| Female voles    |              |                  |
| Comp. 1         | 9.1557       | 0.975            |
| Comp. 2         | 15           | 1                |
Figure 5.6: Scatterplot of the first two discriminant coordinates for the voles data obtained with \texttt{tclust}. Colour and shape represent estimated cluster membership.

### 5.3 Summary

This chapter introduced a method for detecting outliers within the framework of model-based clustering by using mixtures of skew-normal distributions, and briefly outlined its extension to model-based classification. The introduction of a contamination factor allowed us to increase the dispersion of the fitted distributions. By utilizing the ECM algorithm we determined the group membership and outlier status (yes or no) of each observation.

The results were encouraging for the simulated data, where the procedure identified most of the artificial outliers. In the analysis of real data, the method performed well when compared to trimmed clustering. Future work could focus on determining robust starting parameters and whether introducing parsimony into the contaminated model would be beneficial.
Overall, we noticed that contaminated SAL mixtures produce better clustering results than contaminated SN mixtures (Table 5.8), and they are more efficient computationally.

Table 5.8: Summary of clustering results for the best model, in terms of BIC, fitted to the real data using contaminated mixtures.

<table>
<thead>
<tr>
<th>Data</th>
<th>Model</th>
<th>ARI</th>
<th>Outliers</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIS</td>
<td>Cont. SAL</td>
<td>0.8932</td>
<td>Artificial - detected correctly</td>
</tr>
<tr>
<td></td>
<td>Cont. SN</td>
<td>0.8748</td>
<td>Artificial - detected correctly</td>
</tr>
<tr>
<td>Bank</td>
<td>Cont. SAL</td>
<td>0.9205</td>
<td>Natural - detected 2</td>
</tr>
<tr>
<td></td>
<td>Cont. SN</td>
<td>0.9602</td>
<td>Natural - detected 2</td>
</tr>
<tr>
<td>Olive oil</td>
<td>Cont. SAL</td>
<td>0.7659</td>
<td>Natural - not detected</td>
</tr>
<tr>
<td></td>
<td>Cont. SN</td>
<td>0.7467</td>
<td>Natural - detected 3</td>
</tr>
<tr>
<td>Voles</td>
<td>Cont. SAL</td>
<td>0.9540</td>
<td>Artificial - detected correctly</td>
</tr>
<tr>
<td></td>
<td>Cont. SN</td>
<td>0.9540</td>
<td>Artificial - detected correctly</td>
</tr>
<tr>
<td>Wine</td>
<td>Cont. SAL</td>
<td>0.9226</td>
<td>Artificial - detected correctly</td>
</tr>
<tr>
<td></td>
<td>Cont. SN</td>
<td>0.9172</td>
<td>Artificial - not detected</td>
</tr>
</tbody>
</table>
Chapter 6

Conclusion

This thesis was an exposition of three recent pieces of work within the general framework of model-based clustering for finite mixtures.

First, we introduced the idea of dimension reduction for model-based clustering using mixtures of shifted asymmetric Laplace distributions (SALMMDR), and mixtures of generalized hyperbolic distributions (HMMDR). The methods were based on existing work for reducing dimensionality in the case of finite Gaussian mixtures (GMMDR).

Work focused on identifying the smallest subspace of the data which captured the inherent cluster structure. This information was gathered by looking at how the group means and group covariances varied, using the eigen-decomposition of the kernel matrix. The elements of the subspace consisted of linear combinations of the original data, which were ordered by importance via the associated eigenvalues. Observations were then projected onto the subspace, and the resulting set of variables captured most of the clustering structure available in the data.

Both the SALMMDR and the HMMDR methods were illustrated using simulated and real data, and their performance compared to that of the GMMDR analogue and several well-established dimension reduction procedures. The evaluation
was done using the ARI. In the case of synthetic data, the algorithms produced good results generally. For the real data, the SALMMDR and HMMMDR approaches showed higher ARI values than their GMMDR analogue, but the former selected a slightly bigger set of features.

Overall, using dimension reduction via either SAL mixtures or GH mixtures shows great potential for model-based clustering of asymmetric data. In terms of future work, the main idea would be to study whether incorporating SAL and GH analogues of the GPCM models into our methods is beneficial.

Next, we proposed the idea of detecting and modelling outliers using mixtures of contaminated shifted asymmetric Laplace distributions, as well as mixtures of contaminated skew-normal distributions. The methods were based on existing work on automatic outlier detection in the case of contaminated Gaussian mixtures (the PMCGD family).

Work focused on introducing a contamination factor which increased the dispersion of the fitted distribution by altering the skewness and covariance parameters. The ECM algorithm was employed to detect and classify outliers, by producing robust model parameter estimates. Both methods were applied to simulated and real data, and the performance compared to their contaminated Gaussian mixture analogue, and the trimmed clustering approach. Once again, this was done using the ARI. For simulated data, our contaminated SAL and contaminated SN models showed promising results. In the case of real data, both algorithms outperformed trimmed clustering.

Overall, using contaminated mixtures of SAL or SN distributions shows
good potential for outlier detection. In particular, one of the advantages of these methods is that, unlike trimmed clustering, they do not require the pre-specification of the proportion of observations to be trimmed. However, initial values for the contamination factor are required, and a robust method of selecting these would improve our method. In terms of future work, it would be interesting to incorporate parsimony into our models by considering the contaminated SAL and contaminated SN analogues of the PMCGD family.
Bibliography


Blæsild, P. (1978). The shape of the generalized inverse Gaussian and hyperbolic distributions. Research Report 37, Aarhus University, Denmark, Department of Theoretical Statistics. 32


Fraley, C., Raftery, A. E., Murphy, T. B., and Scrucca, L. (2012a). *MCLUST version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation*. Department of Statistics, University of Washington. 47


Schwender, H. (2012). siggenes: Multiple testing using SAM and Efron’s empirical Bayes approaches. R package version 1.32.0. 57


Appendix A

Appendix

A.1 Outlier detection for clustering using mixtures of contaminated shifted asymmetric Laplace distributions

A.1.1 Parameter updates

We start by differentiating the Q-function in (4.13) with respect to $\mu_g$

$$\frac{\partial Q}{\partial \mu_g} = \sum_{i=1}^{n} \hat{z}_{ig} \hat{v}_{ig} E[W_{ig}^{-1}|x_i, z_{ig} = 1] \Sigma_g^{-1}(x_i - \mu_g)$$

$$- \frac{1}{2} \sum_{i=1}^{n} \hat{z}_{ig} \hat{v}_{ig} \Sigma_g^{-1} \alpha_g - \frac{1}{2} \sum_{i=1}^{n} \hat{z}_{ig} \hat{v}_{ig} \Sigma_g^{-1} \alpha_g$$

$$+ \sum_{i=1}^{n} \hat{z}_{ig}(1 - \hat{v}_{ig}) E[\tilde{W}_{ig}^{-1}|x_i, z_{ig} = 1](\rho_g \Sigma_g)^{-1}(x_i - \mu_g)$$

$$- \frac{1}{2} \sum_{i=1}^{n} \hat{z}_{ig}(1 - \hat{v}_{ig})(\rho_g \Sigma_g)^{-1}(\sqrt{\rho_g} \alpha_g) - \frac{1}{2} \sum_{i=1}^{n} \hat{z}_{ig}(1 - \hat{v}_{ig})(\rho_g \Sigma_g)^{-1}(\sqrt{\rho_g} \alpha_g).$$

(A.1)
Then setting $\frac{\partial Q}{\partial \mu_g} = 0$ leads to

$$0 = \sum_{i=1}^{n} \hat{z}_{ig} \hat{v}_{ig} E[W_{ig}^{-1}|x_i, z_{ig} = 1](x_i - \mu_g) - \sum_{i=1}^{n} \hat{z}_{ig} \hat{v}_{ig} \alpha_g$$

$$+ \sum_{i=1}^{n} \hat{z}_{ig} \left(1 - \hat{v}_{ig}\right) \rho_g E[W_{ig}^{-1}|x_i, z_{ig} = 1](x_i - \mu_g) - \sum_{i=1}^{n} \hat{z}_{ig} \left(1 - \hat{v}_{ig}\right) \sqrt{\rho_g} \alpha_g.$$  

(A.2)

Rearranging the terms involving $\mu_g$ gives

$$\sum_{i=1}^{n} \hat{z}_{ig} \hat{v}_{ig} E[W_{ig}^{-1}|x_i, z_{ig} = 1]x_i - \sum_{i=1}^{n} \hat{z}_{ig} \hat{v}_{ig} \alpha_g$$

$$+ \sum_{i=1}^{n} \hat{z}_{ig} \left(1 - \hat{v}_{ig}\right) \rho_g E[W_{ig}^{-1}|x_i, z_{ig} = 1]x_i - \sum_{i=1}^{n} \hat{z}_{ig} \left(1 - \hat{v}_{ig}\right) \rho_g \alpha_g$$

$$= \sum_{i=1}^{n} \hat{z}_{ig} \hat{v}_{ig} E[W_{ig}^{-1}|x_i, z_{ig} = 1] \mu_g + \sum_{i=1}^{n} \hat{z}_{ig} \left(1 - \hat{v}_{ig}\right) E[W_{ig}^{-1}|x_i, z_{ig} = 1] \mu_g,$$  

(A.3)

and solving for $\mu_g$ results in

$$\mu_g = \left\{ \sum_{i=1}^{n} \hat{z}_{ig} \hat{v}_{ig} E[W_{ig}^{-1}|x_i, z_{ig} = 1] + \frac{1 - \hat{v}_{ig}}{\rho_g} E[W_{ig}^{-1}|x_i, z_{ig} = 1] \right\} x_i$$

$$- \sum_{i=1}^{n} \hat{z}_{ig} \left(\hat{v}_{ig} + \frac{1 - \hat{v}_{ig}}{\sqrt{\rho_g}}\right) \alpha_g \} \times \left\{ \sum_{i=1}^{n} \hat{z}_{ig} \hat{v}_{ig} E[W_{ig}^{-1}|x_i, z_{ig} = 1]$$

$$+ \frac{1 - \hat{v}_{ig}}{\rho_g} E[W_{ig}^{-1}|x_i, z_{ig} = 1] \right\}^{-1}. \quad \text{ (A.4)}$$
For simplicity let 

\[ E_{1ig} := E[W_{ig} | x_i, z_{ig} = 1] = \sqrt{\frac{a}{b}} K_\nu(u), \]

\[ E_{2ig} := E[W_{ig}^{-1} | x_i, z_{ig} = 1] = \sqrt{\frac{b}{a}} K_\nu(u) - \frac{2\nu}{a}, \]

\[ E_{3ig} := E[\tilde{W}_{ig} | x_i, z_{ig} = 1] = \sqrt{\frac{a}{b}} K_\nu(\tilde{u}), \]

\[ E_{4ig} := E[\tilde{W}_{ig}^{-1} | x_i, z_{ig} = 1] = \sqrt{\frac{b}{a}} K_\nu(\tilde{u}) - \frac{2\nu}{a}, \]  

(A.5)

where \( a = (x_i - \mu_g)^\top \Sigma_g^{-1} (x_i - \mu_g) \) and \( b = 2 + \alpha_g^\top \Sigma_g^{-1} \alpha_g \), as defined previously in Chapter 2, with \( \tilde{a} \) and \( \tilde{b} \) their analogues when \( \Sigma_g^{-1} \rightarrow (\rho_g \Sigma_g)^{-1} \) and \( \alpha_g \rightarrow \sqrt{\rho_g} \alpha_g \).

Here, we set 

\[ u = \sqrt{ab} \quad \text{and} \quad \tilde{u} = \sqrt{\tilde{a}\tilde{b}}, \]  

(A.6)

and rewriting (A.4) with the simplified notation (A.5) we obtain

\[ \mu_g = \frac{\sum_{i=1}^n \tilde{z}_{ig}(\tilde{v}_{ig} E_{2ig} + \frac{1-\tilde{v}_{ig}}{\rho_g} E_{4ig}) x_i - \sum_{i=1}^n \tilde{z}_{ig}(\tilde{v}_{ig} + \frac{1-\tilde{v}_{ig}}{\sqrt{\rho_g}}) \alpha_g}{\sum_{i=1}^n \tilde{z}_{ig}(\tilde{v}_{ig} E_{2ig} + \frac{1-\tilde{v}_{ig}}{\rho_g} E_{4ig})}. \]  

(A.7)

Next, we differentiate (4.13) with respect to \( \alpha_g \)

\[
\frac{\partial Q}{\partial \alpha_g} = \frac{1}{2} \sum_{i=1}^n \tilde{z}_{ig} \tilde{v}_{ig} \Sigma_g^{-1} (x_i - \mu_g) + \frac{1}{2} \sum_{i=1}^n \tilde{z}_{ig} \tilde{v}_{ig} \Sigma_g^{-1} (x_i - \mu_g) \\
- \frac{1}{2} \sum_{i=1}^n \tilde{z}_{ig} \tilde{v}_{ig} E[W_{ig} | x_i, z_{ig} = 1](\Sigma_g^{-1} + \Sigma_g^{-T}) \alpha_g \\
+ \frac{1}{2} \sum_{i=1}^n \tilde{z}_{ig} (1 - \tilde{v}_{ig})(\rho_g \Sigma_g)^{-1} (x_i - \mu_g) \sqrt{\rho_g} \\
+ \frac{1}{2} \sum_{i=1}^n \tilde{z}_{ig} (1 - \tilde{v}_{ig})(\rho_g \Sigma_g)^{-1} (x_i - \mu_g) \sqrt{\rho_g} \\
- \frac{1}{2} \sum_{i=1}^n \tilde{z}_{ig} (1 - \tilde{v}_{ig}) E[\tilde{W}_{ig} | x_i, z_{ig} = 1](\Sigma_g^{-1} + \Sigma_g^{-T}) \alpha_g, \]  

(A.8)
Letting \( \frac{\partial Q}{\partial \mu_g} \) the update for \( \mu \), we use the simplified notation (A.5) to rewrite the above as

\[
\frac{\partial (\mathbf{x}_i - \mu_g)^\top \Sigma_g^{-1} \alpha_g}{\partial \alpha_g} = \frac{\partial ((\Sigma_g^{-1})^\top (\mathbf{x}_i - \mu_g))}{\partial \alpha_g} = (\Sigma_g^{-1})^\top (\mathbf{x}_i - \mu_g) = \Sigma_g^{-1} (\mathbf{x}_i - \mu_g).
\]

Letting \( \frac{\partial Q}{\partial \alpha_g} = 0 \), one obtains

\[
\sum_{i=1}^n \hat{z}_{ig} \hat{v}_{ig} \Sigma_g^{-1} (\mathbf{x}_i - \mu_g) + \sum_{i=1}^n \hat{z}_{ig} (1 - \hat{v}_{ig}) (\rho_g \Sigma_g)^{-1} (\mathbf{x}_i - \mu_g) \sqrt{\rho_g} = \sum_{i=1}^n \hat{z}_{ig} (1 - \hat{v}_{ig}) E[W_{ig} | \mathbf{x}_i, z_{ig} = 1] \Sigma_g^{-1} \alpha_g + \sum_{i=1}^n \hat{z}_{ig} (1 - \hat{v}_{ig}) E[\tilde{W}_{ig} | \mathbf{x}_i, z_{ig} = 1] \Sigma_g^{-1} \alpha_g,
\]

and multiplying throughout by \( \Sigma_g \), then solving for \( \alpha_g \) gives

\[
\alpha_g = \frac{\sum_{i=1}^n \hat{z}_{ig} (\hat{v}_{ig} + \frac{1 - \hat{v}_{ig}}{\sqrt{\rho_g}}) (\mathbf{x}_i - \mu_g)}{\sum_{i=1}^n \hat{z}_{ig} (\hat{v}_{ig} E[1_{ig}] + (1 - \hat{v}_{ig}) E[3_{ig}]).}
\]

We use the simplified notation (A.5) to rewrite the above as

\[
\alpha_g = \frac{\sum_{i=1}^n \hat{z}_{ig} (\hat{v}_{ig} + \frac{1 - \hat{v}_{ig}}{\sqrt{\rho_g}}) (\mathbf{x}_i - \mu_g)}{\sum_{i=1}^n \hat{z}_{ig} (\hat{v}_{ig} E[1_{ig}] + (1 - \hat{v}_{ig}) E[3_{ig}]).}
\]

Now, (A.7) and (A.10) can be used as updates for \( \mu_g \) and \( \alpha_g \), respectively. However, they are amenable to being expressed solely in terms of conditional expectations for the quantities \( W_{ig} \) and \( \tilde{W}_{ig} \). After substituting (A.10) into (A.7) we obtain the update for \( \mu_g \)

\[
\mu_g = \left\{ \left( \sum_{i=1}^n \hat{z}_{ig} (\hat{v}_{ig} E[1_{ig}] + (1 - \hat{v}_{ig}) E[3_{ig}]) \left( \sum_{i=1}^n \hat{z}_{ig} (\hat{v}_{ig} E[2_{ig}] + \frac{1 - \hat{v}_{ig}}{\rho_g} E[4_{ig}]) \mathbf{x}_i \right) \right) \right.
\]

\[
- \left( \sum_{i=1}^n \hat{z}_{ig} (\hat{v}_{ig} + \frac{1 - \hat{v}_{ig}}{\rho_g}) \left( \sum_{i=1}^n \hat{z}_{ig} (\hat{v}_{ig} + \frac{1 - \hat{v}_{ig}}{\rho_g}) \mathbf{x}_i \right) \right) \left\} \times \left\{ \left( \sum_{i=1}^n \hat{z}_{ig} (\hat{v}_{ig} E[1_{ig}] + (1 - \hat{v}_{ig}) E[3_{ig}]) \left( \sum_{i=1}^n \hat{z}_{ig} (\hat{v}_{ig} E[2_{ig}] + \frac{1 - \hat{v}_{ig}}{\rho_g} E[4_{ig}]) \right. \right) \right.
\]

\[
- \left( \sum_{i=1}^n \hat{z}_{ig} (\hat{v}_{ig} + \frac{1 - \hat{v}_{ig}}{\rho_g}) \right) \left\} \right\}^{-1}.
\]
For simplicity, we introduce the following notation

\[
A = \sum_{i=1}^{n} \hat{z}_{ig}(\hat{v}_{ig} E_{2ig} + \frac{1 - \hat{v}_{ig}}{\rho} E_{4ig}),
\]

\[
B = \sum_{i=1}^{n} \hat{z}_{ig}(\hat{v}_{ig} E_{1ig} + (1 - \hat{v}_{ig}) E_{3ig}),
\]

\[
C = \sum_{i=1}^{n} \hat{z}_{ig}(\hat{v}_{ig} + \frac{1 - \hat{v}_{ig}}{\rho})
\]

\[
D = \sum_{i=1}^{n} \hat{z}_{ig}(\hat{v}_{ig} + \frac{1 - \hat{v}_{ig}}{\sqrt{\rho}}),
\]

(A.12)

which reduces the \(\mu_g\) update to

\[
\hat{\mu}_g = \frac{B\left(\sum_{i=1}^{n} \hat{z}_{ig}(\hat{v}_{ig} E_{2ig} + \frac{1 - \hat{v}_{ig}}{\rho} E_{4ig})x_i\right) - C\left(\sum_{i=1}^{n} \hat{z}_{ig}(\hat{v}_{ig} + \frac{1 - \hat{v}_{ig}}{\rho})x_i\right)}{BA - C^2}.
\]

(A.13)

Next, we can substitute (A.11) into (A.10) to find the update for \(\alpha_g\)

\[
\hat{\alpha}_g = \left\{ \left(\sum_{i=1}^{n} \hat{z}_{ig}(\hat{v}_{ig} E_{2ig} + \frac{1 - \hat{v}_{ig}}{\rho} E_{4ig})\right) \left(\sum_{i=1}^{n} \hat{z}_{ig}(\hat{v}_{ig} + \frac{1 - \hat{v}_{ig}}{\sqrt{\rho}})x_i\right) \right. \\
- \left. \left(\sum_{i=1}^{n} \hat{z}_{ig}(\hat{v}_{ig} + \frac{1 - \hat{v}_{ig}}{\sqrt{\rho}})\right) \left(\sum_{i=1}^{n} \hat{z}_{ig}(\hat{v}_{ig} E_{2ig} + \frac{1 - \hat{v}_{ig}}{\rho} E_{4ig})x_i\right) \right\} \\
\times \left\{ \left(\sum_{i=1}^{n} \hat{z}_{ig}(\hat{v}_{ig} E_{1ig} + (1 - \hat{v}_{ig}) E_{3ig})\right) \left(\sum_{i=1}^{n} \hat{z}_{ig}(\hat{v}_{ig} E_{2ig} + \frac{1 - \hat{v}_{ig}}{\rho} E_{4ig})\right) \right. \\
- \left. \left(\sum_{i=1}^{n} \hat{z}_{ig}(\hat{v}_{ig} + \frac{1 - \hat{v}_{ig}}{\sqrt{\rho}})\right)^2 \right\}^{-1},
\]

(A.14)

and it can be simplified using the notation in (A.12)

\[
\hat{\alpha}_g = \frac{A\left(\sum_{i=1}^{n} \hat{z}_{ig}(\hat{v}_{ig} + \frac{1 - \hat{v}_{ig}}{\sqrt{\rho}})x_i\right) - D\left(\sum_{i=1}^{n} \hat{z}_{ig}(\hat{v}_{ig} E_{2ig} + \frac{1 - \hat{v}_{ig}}{\rho} E_{4ig})x_i\right)}{BA - D^2},
\]

(A.15)
Next, differentiating the Q-function in (4.13) with respect to $\Sigma^{-1}_{g}$ leads to

$$\frac{\partial Q}{\partial \Sigma^{-1}_{g}} = \sum_{i=1}^{n} \frac{\hat{z}_{ig}\hat{v}_{ig}}{2} \Sigma_{g} - \frac{1}{2} \sum_{i=1}^{n} \hat{z}_{ig}\hat{v}_{ig} E[W_{ig}^{-1}|x_{i}, z_{ig} = 1](\hat{x}_{i} - \hat{\mu}_{g})(\hat{x}_{i} - \hat{\mu}_{g})^T$$

$$+ \frac{1}{2} \sum_{i=1}^{n} \hat{z}_{ig}\hat{v}_{ig}(x_{i} - \mu_{g})\alpha_{g}^T + \frac{1}{2} \sum_{i=1}^{n} \hat{z}_{ig}\hat{v}_{ig} \alpha_{g}^T$$

$$- \frac{1}{2} \sum_{i=1}^{n} \hat{z}_{ig}\hat{v}_{ig} E[W_{ig}|x_{i}, z_{ig} = 1]\alpha_{g}^T + \sum_{i=1}^{n} \frac{\hat{z}_{ig}(1 - \hat{v}_{ig})}{2} \Sigma_{g}$$

$$- \frac{1}{2} \sum_{i=1}^{n} \hat{z}_{ig}(1 - \hat{v}_{ig}) E[\hat{W}_{ig}^{-1}|x_{i}, z_{ig} = 1]\rho_{g}^{-1}(x_{i} - \mu_{g})(x_{i} - \mu_{g})^T$$

$$+ \frac{1}{2} \sum_{i=1}^{n} \hat{z}_{ig}(1 - \hat{v}_{ig}) \sqrt{\frac{\rho_{g}}{\rho_{g}}} (x_{i} - \mu_{g})\alpha_{g}^T + \frac{1}{2} \sum_{i=1}^{n} \hat{z}_{ig}(1 - \hat{v}_{ig}) \sqrt{\frac{\rho_{g}}{\rho_{g}}} \alpha_{g}^T$$

$$- \frac{1}{2} \sum_{i=1}^{n} \hat{z}_{ig}(1 - \hat{v}_{ig}) E[\hat{W}_{ig}|x_{i}, z_{ig} = 1] \sqrt{\frac{\rho_{g}}{\rho_{g}}} \alpha_{g}^T \alpha_{g}^T. \quad (A.16)$$

Then setting $\frac{\partial Q}{\partial \Sigma_{g}} = 0$ and solving for $\Sigma_{g}$ we obtain the corresponding update

$$\hat{\Sigma}_{g} = \frac{1}{\sum_{i=1}^{n} \hat{z}_{ig}} \left( \sum_{i=1}^{n} \hat{z}_{ig}(\hat{v}_{ig} E_{2ig} + \frac{1 - \hat{v}_{ig}}{\rho_{g}} E_{4ig}) \right)(x_{i} - \hat{\mu}_{g})(x_{i} - \hat{\mu}_{g})^T$$

$$- \frac{1}{\sum_{i=1}^{n} \hat{z}_{ig}} \sum_{i=1}^{n} \hat{z}_{ig} \left( \hat{v}_{ig} + \frac{1 - \hat{v}_{ig}}{\sqrt{\rho_{g}}} \right) (x_{i} - \hat{\mu}_{g})\alpha_{g}^T$$

$$- \frac{1}{\sum_{i=1}^{n} \hat{z}_{ig}} \sum_{i=1}^{n} \hat{z}_{ig} \left( \hat{v}_{ig} + \frac{1 - \hat{v}_{ig}}{\sqrt{\rho_{g}}} \right) \alpha_{g}(x_{i} - \hat{\mu}_{g})^T$$

$$+ \frac{1}{\sum_{i=1}^{n} \hat{z}_{ig}} \alpha_{g} \alpha_{g}^T \left( \sum_{i=1}^{n} \hat{z}_{ig}(\hat{v}_{ig} E_{1ig} + (1 - \hat{v}_{ig}) E_{3ig}) \right). \quad (A.17)$$
and the above can be written more simply using the notation in (A.12)

\[
\hat{\Sigma}_g = \frac{1}{\sum_{i=1}^{n} \hat{z}_{ig}} A(x_i - \hat{\mu}_g)(x_i - \hat{\mu}_g)\top
\]

\[- \frac{1}{\sum_{i=1}^{n} \hat{z}_{ig}} \sum_{i=1}^{n} \hat{z}_{ig} (\hat{v}_{ig} + 1 - \hat{v}_{ig}) (x_i - \hat{\mu}_g)\hat{\alpha}_g \top
\]

\[- \frac{1}{\sum_{i=1}^{n} \hat{z}_{ig}} \sum_{i=1}^{n} \hat{z}_{ig} \left( \hat{v}_{ig} + 1 - \hat{v}_{ig} \right) \hat{\alpha}_g (x_i - \hat{\mu}_g)\top
\]

\[+ \frac{1}{\sum_{i=1}^{n} \hat{z}_{ig}} \hat{\alpha}_g \hat{\alpha}_g\top B. \quad (A.18)
\]

At the second CM-step of the ECM algorithm we update the contamination factor \(\rho_g\) by differentiating (4.13) wrt \(\rho_g^{-1}\). Let \(\rho_g^{-1} = a_g\) and differentiate (4.13) term by term.

\[
\text{Term 1} = \frac{\partial}{\partial a_g} \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} (1 - \hat{v}_{ig}) \frac{\partial}{\partial a_g} \log(\det(a_g \Sigma_g^{-1}))
\]

\[= \sum_{i=1}^{n} \hat{z}_{ig} (1 - \hat{v}_{ig}) \frac{\partial}{\partial a_g} \log(a_g \det(\Sigma_g^{-1}))
\]

\[= \sum_{i=1}^{n} \hat{z}_{ig} (1 - \hat{v}_{ig}) \frac{\partial}{\partial a_g} (p \log(a_g) + \log(\det(\Sigma_g^{-1}))
\]

\[= \sum_{i=1}^{n} \hat{z}_{ig} (1 - \hat{v}_{ig}) \frac{p}{a_g} = \frac{\sum_{i=1}^{n} \hat{z}_{ig} (1 - \hat{v}_{ig})}{2} p \rho_g, \quad (A.19)
\]

\[
\text{Term 2} = \frac{\partial}{\partial a_g} \left[ - \frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} (1 - \hat{v}_{ig}) E[\hat{W}_{ig}^{-1} | x_i, z_{ig} = 1] \right.
\]

\[\times (x_i - \mu_g)\top a_g \Sigma_g^{-1} (x_i - \mu_g) \right]
\]

\[= - \frac{1}{2} \sum_{i=1}^{n} \hat{z}_{ig} (1 - \hat{v}_{ig}) E[\hat{W}_{ig}^{-1} | x_i, z_{ig} = 1] (x_i - \mu_g)\top \Sigma_g^{-1} (x_i - \mu_g),
\]

\[\quad (A.20)\]
Term 3 = \[ \frac{\partial}{\partial a_g} \left[ \frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig})(x_i - \mu_g)^\top \Sigma_g^{-1} \alpha_g a_g^{1/2} \right] \]

= \[ \frac{1}{2} a_g^{-1/2} \frac{1}{2} \sum_{i=1}^{n} \hat{z}_{ig}(1 - \hat{v}_{ig})(x_i - \mu_g)^\top \Sigma_g^{-1} \alpha_g \]

= \[ \frac{1}{4} \rho_y^{1/2} \sum_{i=1}^{n} \hat{z}_{ig}(1 - \hat{v}_{ig})(x_i - \mu_g)^\top \Sigma_g^{-1} \alpha_g, \quad (A.21) \]

Term 4 = \[ \frac{\partial}{\partial a_g} \left[ \frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig}) a_g^{1/2} \alpha_g^\top \Sigma_g^{-1} (x_i - \mu_g) \right] \]

= \[ \frac{1}{2} a_g^{-1/2} \frac{1}{2} \sum_{i=1}^{n} \hat{z}_{ig}(1 - \hat{v}_{ig}) \alpha_g^\top \Sigma_g^{-1} (x_i - \mu_g) \]

= \[ \frac{1}{4} \rho_y^{1/2} \sum_{i=1}^{n} \hat{z}_{ig}(1 - \hat{v}_{ig}) \alpha_g^\top \Sigma_g^{-1} (x_i - \mu_g). \quad (A.22) \]

Combining these four terms, we observe that the aim becomes solving the quadratic equation

\[ 0 = \rho_y p \sum_{i=1}^{n} \frac{\hat{z}_{ig}(1 - \hat{v}_{ig})}{2} + \rho_y^{1/2} \left[ \frac{1}{4} \sum_{i=1}^{n} \hat{z}_{ig}(1 - \hat{v}_{ig})(x_i - \mu_g)^\top \Sigma_g^{-1} \alpha_g \right] \]

+ \[ \frac{1}{4} \sum_{i=1}^{n} \hat{z}_{ig}(1 - \hat{v}_{ig}) \alpha_g^\top \Sigma_g^{-1} (x_i - \mu_g) \]

\[ - \frac{1}{2} \sum_{i=1}^{n} \hat{z}_{ig}(1 - \hat{v}_{ig}) E[\hat{W}_{ig}^{-1}|x_i, z_{ig} = 1](x_i - \mu_g)^\top \Sigma_g^{-1} (x_i - \mu_g), \quad (A.23) \]

or its equivalent simplified version

\[ 0 = \rho_y p \sum_{i=1}^{n} \frac{\hat{z}_{ig}(1 - \hat{v}_{ig})}{2} + \rho_y^{1/2} \left[ \frac{1}{4} \sum_{i=1}^{n} \hat{z}_{ig}(1 - \hat{v}_{ig})(x_i - \mu_g)^\top \Sigma_g^{-1} \alpha_g \right] \]

+ \[ \frac{1}{4} \sum_{i=1}^{n} \hat{z}_{ig}(1 - \hat{v}_{ig}) \alpha_g^\top \Sigma_g^{-1} (x_i - \mu_g) \]

\[ - \frac{1}{2} \sum_{i=1}^{n} \hat{z}_{ig}(1 - \hat{v}_{ig}) E_{4ig}(x_i - \mu_g)^\top \Sigma_g^{-1} (x_i - \mu_g), \quad (A.24) \]
which is of the form $S \rho_g + T \rho_g^{1/2} + U = 0$, for $S$, $T$ and $U$ constant. Its well-known solution is $\rho_g^{1/2} = -\frac{T}{2} + \frac{\sqrt{T^2 - 4SU}}{2S}$, and hence $\rho_g = \left(-\frac{T}{2} + \frac{\sqrt{T^2 - 4SU}}{2S}\right)^2$.

### A.1.2 Extension to model-based classification

For model-based classification with contaminated SAL mixtures, the likelihood equation corresponding to our model in (4.3) becomes

$$
L(\vartheta|x) = \prod_{i=1}^{k} \prod_{g=1}^{G} \left\{ \pi_g f_{\text{SAL}}(x_i|\mu_g, \alpha_g, \Sigma_g)[(1 - \lambda_g)f_{\text{SAL}}(x_i|\mu_g, \sqrt{\rho_g} \alpha_g, \rho_g \Sigma_g)] \right\}^{z_{ig}} \times \prod_{i=k+1}^{n} \sum_{h=1}^{G} \pi_h \left[ \lambda_h f_{\text{SAL}}(x_i|\mu_h, \alpha_h, \Sigma_h) \right] \left[ (1 - \lambda_h)f_{\text{SAL}}(x_i|\mu_h, \sqrt{\rho_h} \alpha_h, \rho_h \Sigma_h) \right].
$$

(A.25)

Then we have the complete-data likelihood as

$$
L(\vartheta|x) = \prod_{i=1}^{k} \prod_{g=1}^{G} \left\{ \pi_g f_{\text{SAL}}(x_i|\mu_g, \alpha_g, \Sigma_g) \right\}^{v_{ig}} \times \left[ (1 - \lambda_g)f_s(x_i|\mu_g, \sqrt{\rho_g} \alpha_g, \rho_g \Sigma_g) \right]^{(1-v_{ig})} \times \prod_{i=k+1}^{n} \sum_{h=1}^{G} \left\{ \pi_h \left[ \lambda_h f_{\text{SAL}}(x_i|\mu_h, \alpha_h, \Sigma_h) \right] \right\}^{v_{ig}} \times \left[ (1 - \lambda_h)f_{\text{SAL}}(x_i|\mu_h, \sqrt{\rho_h} \alpha_h, \rho_h \Sigma_h) \right]^{(1-v_{ig})},
$$

(A.26)

and its associated complete-data log-likelihood can be defined as

$$
l(\vartheta|x) = l_1(\vartheta|x) + l_2(\vartheta|x) + l_3(\vartheta|x) + l_4(\vartheta|x)
$$

(A.27)
where

\[
l_1(\theta|\mathbf{x}) = \sum_{i=1}^{k} \sum_{g=1}^{G} z_{ig} v_{ig} \log(\pi_g) + \sum_{i=1}^{k} \sum_{g=1}^{G} z_{ig} v_{ig} \log(\lambda_g) \\
+ \sum_{i=1}^{k} \sum_{g=1}^{G} z_{ig} v_{ig} \log(f_{\text{SAL}}(\mathbf{x}_i|\mu_g, \alpha_g, \Sigma_g)),
\]

\[
l_2(\theta|\mathbf{x}) = \sum_{i=k+1}^{n} \sum_{h=1}^{G} z_{ih} v_{ih} \log(\pi_h) + \sum_{i=k+1}^{n} \sum_{g=1}^{G} z_{ih} v_{ih} \log(\lambda_h) \\
+ \sum_{i=k+1}^{n} \sum_{h=1}^{G} z_{ih} v_{ih} \log(f_{\text{SAL}}(\mathbf{x}_i|\mu_h, \alpha_h, \Sigma_h)),
\]

\[
l_3(\theta|\mathbf{x}) = \sum_{i=1}^{k} \sum_{g=1}^{G} z_{ig} (1 - v_{ig}) \log(\pi_g) + \sum_{i=1}^{k} \sum_{g=1}^{G} z_{ig} (1 - v_{ig}) \log(1 - \lambda_g) \\
+ \sum_{i=1}^{k} \sum_{g=1}^{G} z_{ig} (1 - v_{ig}) \log(f_{\text{SAL}}(\mathbf{x}_i|\mu_g, \sqrt{\rho_g} \alpha_g, \rho_g \Sigma_g)),
\]

\[
l_4(\theta|\mathbf{x}) = \sum_{i=k+1}^{n} \sum_{h=1}^{G} z_{ih} (1 - v_{ih}) \log(\pi_h) + \sum_{i=k+1}^{n} \sum_{h=1}^{G} z_{ih} (1 - v_{ih}) \log(1 - \lambda_h) \\
+ \sum_{i=k+1}^{n} \sum_{h=1}^{G} z_{ih} (1 - v_{ih}) \log(f_{\text{SAL}}(\mathbf{x}_i|\mu_h, \sqrt{\rho_h} \alpha_h, \rho_h \Sigma_h)).
\]

We employ the ECM algorithm to find the maximum likelihood estimates in (A.54). At the E-step of the algorithm the Q-function, \( Q(\theta|\hat{\theta}) = E[l(\theta|\mathbf{x})|\mathbf{x}, \hat{\theta}] \) is computed. This Q-function denotes the conditional expectation of (A.54) given the observed data \( \mathbf{x} \) and the current estimated parameters \( \hat{\theta} \). For simplicity we have
written \( l(\vartheta |x) = l_1(\vartheta |x) + l_2(\vartheta |x) + l_3(\vartheta |x) + l_4(\vartheta |x) \), where each \( l_i(\vartheta |x) \) contains three related terms from (A.54). Then we can write the Q-function as
\[
Q = E[l_1(\vartheta |x)] + E[l_2(\vartheta |x)] + E[l_3(\vartheta |x)] + E[l_4(\vartheta |x)],
\]  
(A.28)
thus for the first term in the Q-function (A.28) we obtain
\[
E[l_1(\vartheta |x)] = \sum_{i=1}^{k} \sum_{g=1}^{G} \hat{z}_{ig} \hat{v}_{ig} \log(\pi_g) + \sum_{i=1}^{k} \sum_{g=1}^{G} \hat{z}_{ig} \hat{v}_{ig} \log(\lambda_g)
\]
\[
- \sum_{i=1}^{k} \sum_{g=1}^{G} \hat{z}_{ig} \hat{v}_{ig} \log(2\pi) + \sum_{i=1}^{k} \sum_{g=1}^{G} \hat{z}_{ig} \hat{v}_{ig} \frac{1}{2} \log(\Sigma_g^{-1})
\]
\[
- \frac{1}{2} \sum_{i=1}^{k} \sum_{g=1}^{G} \hat{z}_{ig} \hat{v}_{ig} E[W_{ig}^{-1} | x_i, \pi_g = 1] (x_i - \mu_g)^\top \Sigma_g^{-1} (x_i - \mu_g)
\]
\[
- \frac{p}{2} \sum_{i=1}^{k} \sum_{g=1}^{G} \hat{z}_{ig} \hat{v}_{ig} E[\log(W_{ig}) | x_i, \pi_g = 1]
\]
\[
+ \frac{1}{2} \sum_{i=1}^{k} \sum_{g=1}^{G} \hat{z}_{ig} \hat{v}_{ig} (x_i - \mu_g)^\top \Sigma_g^{-1} \alpha_g + \frac{1}{2} \sum_{i=1}^{k} \sum_{g=1}^{G} \hat{z}_{ig} \hat{v}_{ig} \alpha_g^\top \Sigma_g^{-1} (x_i - \mu_g)
\]
\[
- \frac{1}{2} \sum_{i=1}^{k} \sum_{g=1}^{G} \hat{z}_{ig} \hat{v}_{ig} E[W_{ig} | x_i, \pi_g = 1] \alpha_g^\top \Sigma_g^{-1} \alpha_g
\]
\[
- \sum_{i=1}^{k} \sum_{g=1}^{G} \hat{z}_{ig} \hat{v}_{ig} E[W_{ig} | x_i, \pi_g = 1].
\]  
(A.29)
Since the terms \(-\frac{p}{2} \sum_{i=1}^{k} \sum_{g=1}^{G} \hat{z}_{ig} \hat{v}_{ig} E[\log(W_{ig}) | x_i, \pi_g = 1] \) and \(- \sum_{i=1}^{k} \sum_{g=1}^{G} \hat{z}_{ig} \hat{v}_{ig} \times E[W_{ig} | x_i, \pi_g = 1] \) are constant with respect to the model parameters, we can omit them from our calculations. The second term in the Q-function (A.28) has the fol-
Following structure

\[
E[l_2(\theta | x)] = \sum_{i=k+1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} \hat{v}_{ig} \log(\pi_g) + \sum_{i=k+1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} \hat{v}_{ig} \log(\lambda_g) \\
- \sum_{i=k+1}^{n} \sum_{g=1}^{G} \frac{\hat{z}_{ig} \hat{v}_{ig} p}{2} \log(2\pi) + \sum_{i=k+1}^{n} \sum_{g=1}^{G} \frac{\hat{z}_{ig} \hat{v}_{ig}}{2} \log(|\Sigma_g^{-1}|) \\
- \frac{1}{2} \sum_{i=k+1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} \hat{v}_{ig} E[\log(W_{ig}) | x_i, z_{ig} = 1] \sum_{i=k+1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} \hat{v}_{ig} E[\log(W_{ig}) | x_i, z_{ig} = 1] \\
+ \frac{1}{2} \sum_{i=k+1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} \hat{v}_{ig} (x_i - \mu_g)^\top \Sigma_g^{-1} (x_i - \mu_g) + \frac{1}{2} \sum_{i=k+1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} \hat{v}_{ig} \alpha_g^\top \Sigma_g^{-1} (x_i - \mu_g) \\
- \frac{1}{2} \sum_{i=k+1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} \hat{v}_{ig} E[\log(W_{ig}) | x_i, z_{ig} = 1] \alpha_g^\top \Sigma_g^{-1} \alpha_g \\
- \sum_{i=k+1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} \hat{v}_{ig} E[W_{ig} | x_i, z_{ig} = 1].
\]

(A.30)

Since the terms \(-\frac{p}{2} \sum_{i=k+1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} \hat{v}_{ig} E[\log(W_{ig}) | x_i, z_{ig} = 1] \) and \(-\sum_{i=k+1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} \hat{v}_{ig} \times E[W_{ig} | x_i, z_{ig} = 1] \) are constant with respect to the model parameters, we can omit
them from our calculations. The third term in the Q-function (A.28) becomes

$$E[l_3(\theta | x)] = \sum_{i=1}^{k} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig}) \log(1 - \lambda_g) + \sum_{i=1}^{k} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig}) \log(\pi_g)$$

$$- \sum_{i=1}^{k} \sum_{g=1}^{G} \frac{\hat{z}_{ig}(1 - \hat{v}_{ig})p}{2} \log(2\pi) + \sum_{i=1}^{k} \sum_{g=1}^{G} \frac{\hat{z}_{ig}(1 - \hat{v}_{ig})}{2} \log(||(\rho_g \Sigma_g)^{-1}||)$$

$$- \frac{1}{2} \sum_{i=1}^{k} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig}) E[\tilde{W}_{ig}^{-1}|x_i, z_{ig} = 1]|(x_i - \mu_g)^\top (\rho_g \Sigma_g)^{-1}(x_i - \mu_g)$$

$$- \frac{p}{2} \sum_{i=1}^{k} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig}) E[\log(\tilde{W}_{ig})|x_i, z_{ig} = 1]$$

$$+ \frac{1}{2} \sum_{i=1}^{k} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig})(x_i - \mu_g)^\top (\rho_g \Sigma_g)^{-1}(\sqrt{\rho_g} \alpha_g)$$

$$+ \frac{1}{2} \sum_{i=1}^{k} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig})(\sqrt{\rho_g} \alpha_g)^\top (\rho_g \Sigma_g)^{-1}(x_i - \mu_g)$$

$$- \frac{1}{2} \sum_{i=1}^{k} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig}) E[\tilde{W}_{ig}|x_i, z_{ig} = 1]|(\sqrt{\rho_g} \alpha_g)^\top (\rho_g \Sigma_g)^{-1}(\sqrt{\rho_g} \alpha_g)$$

$$- \sum_{i=1}^{k} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig}) E[\tilde{W}_{ig}|x_i, z_{ig} = 1]. \quad (A.31)$$

Once again two terms can be omitted from the calculations, namely $-\frac{p}{2} \sum_{i=1}^{k} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig}) E[\log(\tilde{W}_{ig})|x_i, z_{ig} = 1]$ and $-\sum_{i=1}^{k} \sum_{g=1}^{G} \hat{z}_{ig} \hat{v}_{ig} E[\tilde{W}_{ig}|x_i, z_{ig} = 1]$, because they are constant with respect to the model parameters. The fourth term in the Q-function
(A.28) has the following representation

\[ E[l_2(\theta | x)] = \sum_{i=k+1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig}) \log(1 - \lambda_g) + \sum_{i=k+1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig}) \log(\pi_g) \]

\[ - \sum_{i=k+1}^{n} \sum_{g=1}^{G} \frac{\hat{z}_{ig}(1 - \hat{v}_{ig})}{2} \log(2\pi) + \sum_{i=k+1}^{n} \sum_{g=1}^{G} \frac{\hat{z}_{ig}(1 - \hat{v}_{ig})}{2} \log((\rho_g \Sigma_g)^{-1}) \]

\[ - \frac{1}{2} \sum_{i=k+1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig}) E[\hat{W}_{ig}^{-1}|x_i, z_{ig} = 1](x_i - \mu_g)^\top(\rho_g \Sigma_g)^{-1}(x_i - \mu_g) \]

\[ - \frac{p}{2} \sum_{i=k+1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig}) E[\log(\hat{W}_{ig})|x_i, z_{ig} = 1] \]

\[ + \frac{1}{2} \sum_{i=k+1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig})(x_i - \mu_g)^\top(\rho_g \Sigma_g)^{-1}(\sqrt{\rho_g} \alpha_g) \]

\[ + \frac{1}{2} \sum_{i=k+1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig})(\sqrt{\rho_g} \alpha_g)^\top(\rho_g \Sigma_g)^{-1}(x_i - \mu_g) \]

\[ - \frac{1}{2} \sum_{i=k+1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig}) E[\hat{W}_{ig}|x_i, z_{ig} = 1](\sqrt{\rho_g} \alpha_g)^\top(\rho_g \Sigma_g)^{-1}(\sqrt{\rho_g} \alpha_g) \]

\[ - \sum_{i=k+1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig}) E[\hat{W}_{ig}|x_i, z_{ig} = 1]. \quad \text{(A.32)} \]

Once again two terms can be omitted from the calculations, namely \(-\frac{p}{2} \sum_{i=k+1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig}) E[\log(\hat{W}_{ig})|x_i, z_{ig} = 1]\) and \(-\sum_{i=k+1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} \hat{v}_{ig} E[\hat{W}_{ig}|x_i, z_{ig} = 1]\), because they are constant with respect to the model parameters.

In what follows, if the \(i\)-th observation is labelled then we denote its component membership by \(\hat{z}_{ig}\). The calculations of the updates follow the same procedure taken in the model-based clustering framework. Thus, we only summarize the updates for the parameters in the contaminated SAL mixture, without providing all the details pertaining to their derivation. To simplify the calculation steps used in the
updates we introduce the following notation.

\[ \tilde{A} = \sum_{i=1}^{k} \tilde{z}_{ig}(v_{ig}E_{2ig} + \frac{1-v_{ig}}{\rho_g}E_{4ig}), \]
\[ A = \sum_{i=k+1}^{n} \tilde{z}_{ig}(v_{ig}E_{2ig} + \frac{1-v_{ig}}{\rho_g}E_{4ig}), \]
\[ \tilde{B} = \sum_{i=1}^{k} \tilde{z}_{ig}(v_{ig}E_{1ig} + (1-v_{ig})E_{3ig}), \]
\[ B = \sum_{i=k+1}^{n} z_{ig}(v_{ig}E_{1ig} + (1-v_{ig})E_{3ig}), \]
\[ \tilde{C} = \sum_{i=1}^{k} \tilde{z}_{ig}(v_{ig} + \frac{1-v_{ig}}{\rho_g}), \]
\[ C = \sum_{i=1}^{k} z_{ig}(v_{ig} + \frac{1-v_{ig}}{\rho_g}), \]
\[ \tilde{D} = \sum_{i=1}^{k} \tilde{z}_{ig}(v_{ig} + \frac{1-v_{ig}}{\sqrt{\rho_g}}), \]
\[ D = \sum_{i=k+1}^{n} z_{ig}(v_{ig} + \frac{1-v_{ig}}{\sqrt{\rho_g}}). \]  

(A.33)

Then the updates for \( \mu_g \) and \( \alpha_g \) are

\[ \tilde{\mu}_g = \frac{\tilde{B}\left(\sum_{i=1}^{k} \tilde{z}_{ig}(v_{ig}E_{2ig} + \frac{1-v_{ig}}{\rho_g}E_{4ig})\mathbf{x}_i\right) - \tilde{C}\left(\sum_{i=1}^{k} \tilde{z}_{ig}(v_{ig} + \frac{1-v_{ig}}{\rho_g})\mathbf{x}_i\right)}{BA - \tilde{C}^2} \]
\[ + \frac{B\left(\sum_{i=k+1}^{n} z_{ig}(v_{ig}E_{2ig} + \frac{1-v_{ig}}{\rho_g}E_{4ig})\mathbf{x}_i\right) - C\left(\sum_{i=k+1}^{n} z_{ig}(v_{ig} + \frac{1-v_{ig}}{\rho_g})\mathbf{x}_i\right)}{BA - \tilde{C}^2}, \]

(A.34)

\[ \tilde{\alpha}_g = \frac{\tilde{A}\left(\sum_{i=1}^{k} \tilde{z}_{ig}(v_{ig} + \frac{1-v_{ig}}{\sqrt{\rho_g}})\mathbf{x}_i\right) - \tilde{D}\left(\sum_{i=1}^{k} \tilde{z}_{ig}(v_{ig}E_{2ig} + \frac{1-v_{ig}}{\rho_g}E_{4ig})\mathbf{x}_i\right)}{BA - \tilde{D}^2} \]
\[ + \frac{A\left(\sum_{i=k+1}^{n} z_{ig}(v_{ig} + \frac{1-v_{ig}}{\sqrt{\rho_g}})\mathbf{x}_i\right) - D\left(\sum_{i=k+1}^{n} z_{ig}(v_{ig}E_{2ig} + \frac{1-v_{ig}}{\rho_g}E_{4ig})\mathbf{x}_i\right)}{BA - \tilde{D}^2}. \]

(A.35)
For the update of $\Sigma_g$, we set $\hat{\Sigma}_g = \hat{\Sigma}_{1g} + \hat{\Sigma}_{2g}$, where

\[
\hat{\Sigma}_{1g} = \frac{1}{\sum_{i=1}^{k} \tilde{z}_{ig}} \tilde{A}(x_i - \mu_g)(x_i - \mu_g)^\top
\]
\[
- \frac{1}{\sum_{i=1}^{k} \tilde{z}_{ig}} \sum_{i=1}^{k} \tilde{z}_{ig}(v_{ig} + \frac{1 - v_{ig}}{\sqrt{\rho_g}})(x_i - \mu_g)^\top \alpha_g^\top (x_i - \mu_g)
\]
\[
- \frac{1}{\sum_{i=1}^{k} \tilde{z}_{ig}} \sum_{i=1}^{k} \tilde{z}_{ig} \left( v_{ig} + \frac{1 - v_{ig}}{\sqrt{\rho_g}} \right) \alpha_g (x_i - \mu_g)^\top
\]
\[
+ \frac{1}{\sum_{i=1}^{k} \tilde{z}_{ig}} \alpha_g \alpha_g^\top \tilde{B}. \tag{A.36}
\]

\[
\hat{\Sigma}_{2g} = \frac{1}{\sum_{i=k+1}^{n} \tilde{z}_{ig}} \tilde{A}(x_i - \mu_g)(x_i - \mu_g)^\top
\]
\[
- \frac{1}{\sum_{i=k+1}^{n} \tilde{z}_{ig}} \sum_{i=k+1}^{n} \tilde{z}_{ig}(v_{ig} + \frac{1 - v_{ig}}{\sqrt{\rho_g}})(x_i - \mu_g)^\top \alpha_g^\top (x_i - \mu_g)
\]
\[
- \frac{1}{\sum_{i=k+1}^{n} \tilde{z}_{ig}} \sum_{i=k+1}^{n} \tilde{z}_{ig} \left( v_{ig} + \frac{1 - v_{ig}}{\sqrt{\rho_g}} \right) \alpha_g (x_i - \mu_g)^\top
\]
\[
+ \frac{1}{\sum_{i=k+1}^{n} \tilde{z}_{ig}} \alpha_g \alpha_g^\top \tilde{B}. \tag{A.37}
\]

At the second CM-step of the ECM algorithm we update the contamination factor $\rho_g$. Looking at the dependence of the Q-function in (A.28) on $\rho_g$ we have $Q = Q_1 + Q_2$ where

\[
Q_1 = \sum_{i=1}^{k} \sum_{g=1}^{G} \tilde{z}_{ig}(1 - v_{ig}) \left\{ - \frac{p}{2} \log(\rho_g) - \frac{1}{2} \left( x_i - \mu_g \right)^\top (\rho_g \Sigma_g)^{-1} (x_i - \mu_g) \right\}
\]
\[
+ \frac{1}{2} \sum_{i=1}^{k} \sum_{g=1}^{G} \tilde{z}_{ig}(1 - v_{ig}) E_{4ig}(x_i - \mu_g)^\top (\rho_g \Sigma_g)^{-1} (x_i - \mu_g)
\]
\[
+ \frac{1}{2} \sum_{i=1}^{k} \sum_{g=1}^{G} \tilde{z}_{ig}(1 - v_{ig})(x_i - \mu_g)^\top (\rho_g \Sigma_g)^{-1} (\sqrt{\rho_g} \alpha_g)
\]
\[
+ \frac{1}{2} \sum_{i=1}^{k} \sum_{g=1}^{G} \tilde{z}_{ig}(1 - v_{ig})(\sqrt{\rho_g} \alpha_g)^\top (\rho_g \Sigma_g)^{-1} (x_i - \mu_g)
\]
\[
+ \text{terms independent of } \rho_g, \tag{A.38}
\]
and

\[ Q_2 = \sum_{i=k+1}^{n} \sum_{g=1}^{G} z_{ig}(1 - \hat{v}_{ig}) \left\{ -\frac{p}{2} \log(\rho_g) ight\} 
- \frac{1}{2} \sum_{i=k+1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig}) E_{ig}(x_i - \mu_g)^T (\rho_g \Sigma_g)^{-1}(x_i - \mu_g) 
+ \frac{1}{2} \sum_{i=k+1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig})(x_i - \mu_g)^T (\rho_g \Sigma_g)^{-1}(\sqrt{\rho_g} \alpha_g) 
+ \frac{1}{2} \sum_{i=k+1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}(1 - \hat{v}_{ig})(\sqrt{\rho_g} \alpha_g)^T (\rho_g \Sigma_g)^{-1}(x_i - \mu_g) \right\} 
+ \text{terms independent of } \rho_g. \tag{A.39}

Then one can use the R function `optim()` to find the \( \rho_g \) which maximizes the function \( Q = Q_1 + Q_2 \), over the interval (1, \( \rho^{\text{max}}_g \)).
A.2 Outlier detection for clustering using mixtures of contaminated skew normal distributions

A.2.1 Parameter updates

We start by differentiating the Q-function in (5.12) with respect to $\mu_g$

$$\frac{\partial Q}{\partial \mu_g} = \sum_{i=1}^{n} z_{ig} v_{ig} \left\{ \Sigma^{-1}_g (x_i - \mu_g - A_g \eta_{ig}) \right\} + \sum_{i=1}^{n} z_{ig} (1 - v_{ig}) \left\{ \rho_g^{-1} \Sigma^{-1}_g (x_i - \mu_g - \sqrt{\rho_g} A_g \tilde{\eta}_{ig}) \right\}, \tag{A.40}$$

and setting $\frac{\partial Q}{\partial \mu_g} = 0$, then solving for $\mu_g$ gives

$$\sum_{i=1}^{n} z_{ig} v_{ig} \left\{ \Sigma^{-1}_g (x_i - A_g \eta_{ig}) \right\} + \sum_{i=1}^{n} z_{ig} (1 - v_{ig}) \left\{ \rho_g^{-1} \Sigma^{-1}_g (x_i - \sqrt{\rho_g} A_g \tilde{\eta}_{ig}) \right\} = \sum_{i=1}^{n} z_{ig} v_{ig} \Sigma^{-1}_g \mu_g + \sum_{i=1}^{n} z_{ig} (1 - v_{ig}) \rho_g^{-1} \Sigma^{-1}_g \mu_g. \tag{A.41}$$

Then, multiplying throughout by $\Sigma_g$ in the above equation, we obtain the update for $\mu_g$

$$\hat{\mu}_g = \frac{\sum_{i=1}^{n} z_{ig} v_{ig} (x_i - A_g \eta_{ig}) + \sum_{i=1}^{n} z_{ig} (1 - v_{ig}) (x_i - \sqrt{\rho_g} A_g \tilde{\eta}_{ig})}{\sum_{i=1}^{n} z_{ig} v_{ig} + \sum_{i=1}^{n} z_{ig} (1 - v_{ig})}. \tag{A.42}$$

Next, differentiating the Q-function in (5.12) with respect to $\Sigma_g$ we obtain

$$\frac{\partial Q}{\partial \Sigma_g} = -\frac{1}{2} \sum_{i=1}^{n} z_{ig} \Sigma^{-1}_g \frac{1}{2} \sum_{i=1}^{n} z_{ig} v_{ig} \left\{ -\Sigma^{-1}_g (x_i - \mu_g - A_g \eta_{ig}) \right\} \times (x_i - \mu_g - A_g \eta_{ig})^\top \Sigma^{-1}_g - (\Sigma^{-1}_g A_g (\psi_{ig} - \eta_{ig} \tilde{\eta}_{ig}) A_g^\top \Sigma^{-1}_g) \right\}$$

$$- \frac{1}{2} \sum_{i=1}^{n} z_{ig} (1 - v_{ig}) \left\{ -\rho_g^{-1} \Sigma^{-1}_g (x_i - \mu_g - \sqrt{\rho_g} A_g \tilde{\eta}_{ig}) \right\} \times (x_i - \mu_g - \sqrt{\rho_g} A_g \tilde{\eta}_{ig})^\top \Sigma^{-1}_g - (\Sigma^{-1}_g A_g (\tilde{\psi}_{ig} - \tilde{\eta}_{ig} \tilde{\eta}_{ig} \tilde{\eta}_{ig}) A_g^\top \Sigma^{-1}_g) \right\}, \tag{A.43}$$
and letting \( \frac{\partial Q}{\partial \Sigma_g} = 0 \), then multiplying the left and right sides of the equation above by \( \Sigma_g \) leads to the update for \( \Sigma_g \)

\[
\hat{\Sigma}_g = \frac{1}{\sum_{i=1}^{n} z_{ig}} \left[ \sum_{i=1}^{n} z_{ig} v_{ig} \left( (x_i - \mu_g - A_g \eta_{ig}) (x_i - \mu_g - A_g \eta_{ig})^\top \right) + A_g (\psi_{ig} - \eta_{ig} \eta_{ig}^\top ) A_g^\top \right] + \sum_{i=1}^{n} z_{ig} (1 - v_{ig}) \left[ \frac{1}{\rho_g} (x_i - \mu_g - A_g \tilde{\eta}_{ig}) (x_i - \mu_g - \sqrt{\rho_g} A_g \tilde{\eta}_{ig})^\top \right].
\]

(A.44)

Now we turn our attention to computing the update for \( A_g \). First we differentiate (5.12) with respect to \( A_g \)

\[
\frac{\partial Q}{\partial A_g} = \sum_{i=1}^{n} z_{ig} v_{ig} \left\{ \text{diag}(\Sigma_g^{-1} (x_i - \mu_g) \eta_{ig}^\top ) - \text{diag}(\Sigma_g^{-1} A_g \psi_{ig}) \right\} + \sum_{i=1}^{n} z_{ig} (1 - v_{ig}) \left\{ \text{diag}(\rho_g^{-1} \Sigma_g^{-1} (x_i - \mu_g) \tilde{\eta}_{ig}^\top ) - \text{diag}(\rho_g^{-1} \Sigma_g^{-1} \sqrt{\rho_g} A_g \tilde{\psi}_{ig}) \right\}.
\]

(A.45)

and then rearranging the terms we obtain

\[
\sum_{i=1}^{n} z_{ig} v_{ig} (\Sigma_g^{-1} A_g \psi_{ig}) + \sum_{i=1}^{n} z_{ig} (1 - v_{ig}) (\Sigma_g^{-1} A_g \tilde{\psi}_{ig}) = \sum_{i=1}^{n} z_{ig} v_{ig} (\Sigma_g^{-1} (x_i - \mu_g) \eta_{ig}^\top ) + \sum_{i=1}^{n} z_{ig} (1 - v_{ig}) (\Sigma_g^{-1} (x_i - \mu_g) \tilde{\eta}_{ig}^\top ).
\]

(A.46)

Recalling that \( A_g \) was defined as a \( p \times p \) skewness matrix, we wish to express
it in terms of a skewness vector, \( \boldsymbol{\alpha}_g \). One can write
\[
\sum_{i=1}^{n} z_{ig} v_{ig} (\Sigma_g^{-1} \odot \psi_{ig}) + \sum_{i=1}^{n} z_{ig} \frac{(1 - v_{ig})}{\sqrt{\rho_g}} (\Sigma_g^{-1} \odot \tilde{\psi}_{ig}) \] 
\[
\sum_{i=1}^{n} z_{ig} v_{ig} (\Sigma_g^{-1} \odot \eta_{ig} (x_i - \mu_g)^\top) + \sum_{i=1}^{n} z_{ig} \frac{(1 - v_{ig})}{\rho_g} (\Sigma_g^{-1} \odot \tilde{\eta}_{ig} (x_i - \mu_g)^\top) \]
\[
\alpha_g = \left[ \sum_{i=1}^{n} z_{ig} v_{ig} (\Sigma_g^{-1} \odot \psi_{ig}) + \sum_{i=1}^{n} z_{ig} \frac{(1 - v_{ig})}{\sqrt{\rho_g}} (\Sigma_g^{-1} \odot \tilde{\psi}_{ig}) \right] \mathbf{1}_p,
\]
(A.47)

since \( A_g \) is the diagonal version of \( \alpha_g \), and solving for \( \alpha_g \) gives the desired update for skewness
\[
\hat{\alpha}_g = \left[ \sum_{i=1}^{n} z_{ig} v_{ig} (\Sigma_g^{-1} \odot \eta_{ig} (x_i - \mu_g)^\top) + \sum_{i=1}^{n} z_{ig} \frac{(1 - v_{ig})}{\rho_g} (\Sigma_g^{-1} \odot \tilde{\eta}_{ig} (x_i - \mu_g)^\top) \right] \mathbf{1}_p,
\]
(A.48)

where the operator \( \odot \) denotes the Hadamard (elementwise) product of two matrices of the same dimension.

At the second CM-step of the ECM algorithm we update the contamination factor \( \rho_g \). Looking at the Q-function in (5.12) we observe that
\[
Q = \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} (1 - v_{ig}) \left\{ \frac{1}{2} \log \rho_g^{-1} \Sigma_g^{-1} - \frac{1}{2} (x_i - \mu_g - (\sqrt{\rho_g} A_g) \tilde{\eta}_{ig})^\top \rho_g^{-1} \Sigma_g^{-1} \right\}
\times \left\{ x_i - \mu_g - (\sqrt{\rho_g} A_g) \tilde{\eta}_{ig} \right\} + \text{terms independent of } \rho_g
\]
\[
= \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} (1 - v_{ig}) \left\{ - \frac{p}{2} \log(\rho_g) - \frac{1}{2\rho_g} (x_i - \mu_g)^\top \Sigma_g^{-1} (x_i - \mu_g) \right\}
\]
\[
+ \frac{1}{\sqrt{\rho_g}} \left[ (x_i - \mu_g)^\top \Sigma_g^{-1} (A_g \tilde{\eta}_{ig}) + (A_g \tilde{\eta}_{ig})^\top \Sigma_g^{-1} (x_i - \mu_g) \right] + \text{terms independent of } \rho_g.
\]
(A.49)
The above equation is differentiated with respect to $\rho_g$

$$\frac{\partial Q}{\partial \rho_g} = \sum_{i=1}^{n} z_{ig}(1 - v_{ig}) \left\{ -\frac{p}{2\rho_g} + \frac{1}{2\rho_g^3} (x_i - \mu_g)^\top \Sigma_g^{-1} (x_i - \mu_g) \right. \\
- \frac{1}{2\rho_g^{3/2}} \left[ (x_i - \mu_g)^\top \Sigma_g^{-1} (A_g \tilde{\eta}_{ig}) + (A_g \tilde{\eta}_{ig})^\top \Sigma_g^{-1} (x_i - \mu_g) \right\},$$

(A.50)

and setting $\frac{\partial Q}{\partial \rho_g} = 0$ then multiplying throughout by $2\rho_g^2$ results in the following

$$0 = \rho_g \left( -p \sum_{i=1}^{n} z_{ig}(1 - v_{ig}) \right) + \sum_{i=1}^{n} z_{ig}(1 - v_{ig})(x_i - \mu_g)^\top \Sigma_g^{-1} (x_i - \mu_g) \\
- \rho_g^{1/2} \sum_{i=1}^{n} z_{ig}(1 - v_{ig}) \left[ (x_i - \mu_g)^\top \Sigma_g^{-1} (A_g \tilde{\eta}_{ig}) + (A_g \tilde{\eta}_{ig})^\top \Sigma_g^{-1} (x_i - \mu_g) \right],$$

(A.51)

which is a quadratic equation of the form $S\rho_g + T\rho_g^{1/2} + U = 0$ that can be easily solved for $\rho_g$ (where $S$, $T$, $U$ are constants).

**A.2.2 Extension to model-based classification**

For model-based classification with contaminated SN mixtures, the likelihood equation corresponding to our model in (5.5) becomes

$$L(\vartheta|x) = \prod_{g=1}^{G} \prod_{i=1}^{n} \pi_g \left\{ \lambda_g f_{SN}(x_i|\mu_g, A_g, \Sigma_g) \right\} \left\{ (1 - \lambda_g) f_{SN}(x_i|\mu_g, \sqrt{\rho_g} A_g, \rho_g \Sigma_g) \right\}^{z_{ig}} \\
\times \prod_{h=1}^{H} \sum_{i=k+1}^{n} \pi_h \left\{ \lambda_h f_s(x_i|\mu_h, A_h, \Sigma_h) \right\} \left\{ (1 - \lambda_h) f_{SN}(x_i|\mu_h, \sqrt{\rho_h} A_h, \rho_h \Sigma_h) \right\}.$$

(A.52)
Then we have the complete-data likelihood as

\[
\mathcal{L}(\theta | x) = \prod_{i=1}^{k} \prod_{g=1}^{G} \left\{ \pi_g \left[ \lambda_g f_{SN}(x_i | \mu_g, \alpha_g, \Sigma_g) \right]^{z_{ig}} \right\} \\
\times \left[ (1 - \lambda_g) f_{SN}(x_i | \mu_g, \sqrt{\rho_g} \alpha_g, \rho_g \Sigma_g) \right]^{(1 - v_{ig})} \\
\times \prod_{i=k+1}^{n} \sum_{h=1}^{G} \left\{ \pi_h \left[ \lambda_h f_{SN}(x_i | \mu_h, \alpha_h, \Sigma_h) \right]^{v_{ig}} \right\}^{z_{ig}},
\]

(A.53)

and its associated complete-data log-likelihood can be defined as

\[
l(\theta | x) = \sum_{i=1}^{k} \sum_{g=1}^{G} z_{ig} v_{ig} \log(\pi_g) + \sum_{i=1}^{k} \sum_{g=1}^{G} z_{ig} v_{ig} \log(\lambda_g) \\
+ \sum_{i=1}^{k} \sum_{g=1}^{G} z_{ig} v_{ig} \log(f_{SN}(x_i | \mu_g, A_g, \Sigma_g)) + \sum_{i=1}^{k} \sum_{g=1}^{G} z_{ig} (1 - v_{ig}) \log(1 - \lambda_g) \\
+ \sum_{i=1}^{k} \sum_{g=1}^{G} z_{ig} (1 - v_{ig}) \log(\pi_g) + \sum_{i=1}^{k} \sum_{g=1}^{G} z_{ig} (1 - v_{ig}) \log(f_{SN}(x_i | \mu_g, \sqrt{\rho_g} A_g, \rho_g \Sigma_g)) \\
+ \sum_{i=k+1}^{n} \sum_{g=1}^{G} z_{ig} v_{ig} \log(\pi_g) + \sum_{i=k+1}^{n} \sum_{g=1}^{G} z_{ig} v_{ig} \log(\lambda_g) \\
+ \sum_{i=k+1}^{n} \sum_{g=1}^{G} z_{ig} v_{ig} \log(f_{SN}(x_i | \mu_g, A_g, \Sigma_g)) + \sum_{i=k+1}^{n} \sum_{g=1}^{G} z_{ig} (1 - v_{ig}) \log(1 - \lambda_g) \\
+ \sum_{i=k+1}^{n} \sum_{g=1}^{G} z_{ig} (1 - v_{ig}) \log(\pi_g) + \sum_{i=k+1}^{n} \sum_{g=1}^{G} z_{ig} (1 - v_{ig}) \log(f_{SN}(x_i | \mu_g, \sqrt{\rho_g} A_g, \rho_g \Sigma_g)).
\]

(A.54)
Ignoring additive constants, the above simplifies to $l(\vartheta|x) = l_1(\vartheta|x) + l_1(\vartheta|x)$ where

\[
l_1(\vartheta|x) = \sum_{i=1}^{k} \sum_{g=1}^{G} z_{ig} v_{ig} \log(\pi_g) + \sum_{i=1}^{k} \sum_{g=1}^{G} z_{ig} v_{ig} \log(\lambda_g) \\
+ \sum_{i=1}^{k} \sum_{g=1}^{G} z_{ig} v_{ig} \left\{ -\frac{1}{2} \log |\Sigma_g^{-1}| \right\} \\
- \frac{1}{2} (x_i - \mu_g - A_g \tau_{ig})^\top \Sigma_g^{-1} (x_i - \mu_g - A_g \tau_{ig}) - \frac{1}{2} \tau_{ig}^\top \tau_{ig} \right\} \\
+ \sum_{i=1}^{k} \sum_{g=1}^{G} z_{ig} (1 - v_{ig}) \log(\pi_g) + \sum_{i=1}^{k} \sum_{g=1}^{G} z_{ig} (1 - v_{ig}) \log(1 - \lambda_g) \\
+ \sum_{i=1}^{k} \sum_{g=1}^{G} z_{ig} (1 - v_{ig}) \left\{ -\frac{1}{2} \log |(\rho_g \Sigma_g)^{-1}| \right\} \\
- \frac{1}{2} (x_i - \mu_g - \sqrt{\rho_g A_g} \tilde{\tau}_{ig})^\top (\rho_g \Sigma_g)^{-1} (x_i - \mu_g - \sqrt{\rho_g A_g} \tilde{\tau}_{ig}) - \frac{1}{2} \tilde{\tau}_{ig}^\top \tilde{\tau}_{ig} \right\},
\]

(A.55)

and

\[
l_2(\vartheta|x) = \sum_{i=k+1}^{n} \sum_{g=1}^{G} z_{ig} v_{ig} \log(\pi_g) + \sum_{i=k+1}^{n} \sum_{g=1}^{G} z_{ig} v_{ig} \log(\lambda_g) \\
+ \sum_{i=k+1}^{n} \sum_{g=1}^{G} z_{ig} v_{ig} \left\{ -\frac{1}{2} \log |\Sigma_g^{-1}| \right\} \\
- \frac{1}{2} (x_i - \mu_g - A_g \tau_{ig})^\top \Sigma_g^{-1} (x_i - \mu_g - A_g \tau_{ig}) - \frac{1}{2} \tau_{ig}^\top \tau_{ig} \right\} \\
+ \sum_{i=k+1}^{n} \sum_{g=1}^{G} z_{ig} (1 - v_{ig}) \log(\pi_g) + \sum_{i=k+1}^{n} \sum_{g=1}^{G} z_{ig} (1 - v_{ig}) \log(1 - \lambda_g) \\
+ \sum_{i=k+1}^{n} \sum_{g=1}^{G} z_{ig} (1 - v_{ig}) \left\{ -\frac{1}{2} \log |(\rho_g \Sigma_g)^{-1}| \right\} \\
- \frac{1}{2} (x_i - \mu_g - \sqrt{\rho_g A_g} \tilde{\tau}_{ig})^\top (\rho_g \Sigma_g)^{-1} (x_i - \mu_g - \sqrt{\rho_g A_g} \tilde{\tau}_{ig}) - \frac{1}{2} \tilde{\tau}_{ig}^\top \tilde{\tau}_{ig} \right\},
\]

(A.56)

where $\tilde{\tau}$ is the analogue of $\tau$ when $\Sigma_g \rightarrow \rho_g \Sigma_g$ and $A_g \rightarrow \sqrt{\rho_g} A_g$. 
The ECM algorithm is used to find the maximum likelihood estimates in (5.10). The E-step of the algorithm relies on the computation of the Q-function, $Q(\vartheta|\hat{\vartheta}) = E[l(\vartheta|x)|x, \hat{\vartheta}]$. This Q-function denotes the conditional expectation of (5.10) given the observed data $x$ and the current estimated parameters $\hat{\vartheta}$. The conditional expectations of the terms $-\frac{1}{2} \sum_{i=1}^{k} \sum_{g=1}^{G} z_{ig} v_{ig} \tau_{ig}^T \tau_{ig}$, $-\frac{1}{2} \sum_{i=1}^{k} \sum_{g=1}^{G} z_{ig}(1 - v_{ig}) \tilde{\tau}_{ig}^T \tilde{\tau}_{ig}$, and $-\frac{1}{2} \sum_{i=k+1}^{n} \sum_{g=1}^{G} z_{ig}(1 - v_{ig}) \tilde{\tau}_{ig}^T \tilde{\tau}_{ig}$ can be omitted because they do not contain any parameters. For the computations involved in the Q-function we adopt the following notation

$$E[\tau_{ig}|x_i, z_{ig} = 1] = \eta_{ig}, \quad E[\tau_{ig}^T \tau_{ig}^T|x_i, z_{ig} = 1] = \psi_{ig} \quad \text{and}$$

$$E[\tilde{\tau}_{ig}|x_i, z_{ig} = 1] = \tilde{\eta}_{ig}, \quad E[\tilde{\tau}_{ig}^T \tilde{\tau}_{ig}^T|x_i, z_{ig} = 1] = \tilde{\psi}_{ig},$$

(A.57)

where $\eta_{ig}$, $\tilde{\eta}_{ig}$, $\psi_{ig}$, $\tilde{\psi}_{ig}$ are all implicit functions of the parameters $(\mu_g, \Sigma_g, A_g)$ and $\rho_g$ where required, and can be calculated numerically using the properties of the TN distribution, as outlined in Lin (2009). The Q-function for the contaminated SN mixture model is defined as $Q = Q_1 + Q_2$ where

$$Q_1 = \sum_{i=1}^{k} \sum_{g=1}^{G} z_{ig} v_{ig} \left\{ \log(\lambda_g) + \log(\pi_g) + \frac{1}{2} \log |\Sigma_g^{-1}| \right\}$$

$$- \frac{1}{2} (x_i - \mu_g - A_g \eta_{ig})^T \Sigma_g^{-1} (x_i - \mu_g - A_g \eta_{ig})$$

$$- \frac{1}{2} \text{tr}(\Sigma_g^{-1} A_g (\psi_{ig} - \eta_{ig} \eta_{ig}^T) A_g^T) \}$$

$$+ \sum_{i=1}^{k} \sum_{g=1}^{G} z_{ig}(1 - v_{ig}) \left\{ \log(1 - \lambda_g) + \log(\pi_g) + \frac{1}{2} \log |(\rho_g \Sigma_g)^{-1}| \right\}$$

$$- \frac{1}{2} (x_i - \mu_g - \sqrt{\rho_g} A_g \tilde{\eta}_{ig})^T (\rho_g \Sigma_g)^{-1} (x_i - \mu_g - \sqrt{\rho_g} A_g \tilde{\eta}_{ig})$$

$$- \frac{1}{2} \text{tr}((\rho_g \Sigma_g)^{-1} \sqrt{\rho_g} A_g (\tilde{\psi}_{ig} - \tilde{\eta}_{ig} \tilde{\eta}_{ig}^T) (\sqrt{\rho_g} A_g)^T),$$

(A.58)
and

\[
Q_2 = \sum_{i=k+1}^{n} \sum_{g=1}^{G} z_{ig} v_{ig} \left\{ \log(\lambda_g) + \log(\pi_g) + \frac{1}{2} \log |\Sigma_g^{-1}| \right.
\]

\[
- \frac{1}{2} (x_i - \mu_g - A_g \eta_{ig})^\top \Sigma_g^{-1} (x_i - \mu_g - A_g \eta_{ig})
\]

\[
- \frac{1}{2} \text{tr}(\Sigma_g^{-1} A_g (\psi_{ig} - \eta_{ig} \eta_{ig}^\top) A_g^\top)
\}

\[
+ \sum_{i=k+1}^{n} \sum_{g=1}^{G} z_{ig} (1 - v_{ig}) \left\{ \log(1 - \lambda_g) + \log(\pi_g) + \frac{1}{2} \log |(\rho_g \Sigma_g)^{-1}| \right.
\]

\[
- \frac{1}{2} (x_i - \mu_g - \sqrt{\rho_g} A_g \tilde{\eta}_{ig})^\top (\rho_g \Sigma_g)^{-1} (x_i - \mu_g - \sqrt{\rho_g} A_g \tilde{\eta}_{ig})
\]

\[
- \frac{1}{2} \text{tr}((\rho_g \Sigma_g)^{-1} \sqrt{\rho_g} A_g (\tilde{\psi}_{ig} - \tilde{\eta}_{ig} \tilde{\eta}_{ig}^\top)(\sqrt{\rho_g} A_g)^\top)
\}. \quad (A.59)
\]

In what follows, if the \(i\)-th observation is labelled then we denote its component membership by \(\tilde{z}_{ig}\). The calculations of the updates follow the same procedure taken in the model-based clustering framework. Thus, we only summarize the updates for the parameters in the contaminated SN mixture, without providing all the details pertaining to their derivation. The update for \(\mu_g\) is

\[
\hat{\mu}_g = \frac{\sum_{i=1}^{k} z_{ig} v_{ig} (x_i - A_g \eta_{ig}) + \sum_{i=1}^{k} z_{ig} \frac{(1 - v_{ig})}{\rho} (x_i - \sqrt{\rho_g} A_g \tilde{\eta}_{ig})}{\sum_{i=1}^{k} z_{ig} v_{ig} + \sum_{i=1}^{k} z_{ig} \frac{(1 - v_{ig})}{\rho}}
\]

\[
+ \frac{\sum_{i=k+1}^{n} z_{ig} v_{ig} (x_i - A_g \eta_{ig}) + \sum_{i=k+1}^{n} z_{ig} \frac{(1 - v_{ig})}{\rho} (x_i - \sqrt{\rho_g} A_g \tilde{\eta}_{ig})}{\sum_{i=k+1}^{n} z_{ig} v_{ig} + \sum_{i=k+1}^{n} z_{ig} \frac{(1 - v_{ig})}{\rho}}.
\] \quad (A.60)

Then the update for \(\alpha_g\) becomes
\[ \hat{\alpha}_g = \left[ \sum_{i=1}^{k} z_{ig} \nu_{ig} (\Sigma_g^{-1} \odot \psi_{ig}) + \sum_{i=1}^{k} z_{ig} \frac{(1 - \nu_{ig})}{\sqrt{\rho_g}} (\Sigma_g^{-1} \odot \tilde{\psi}_{ig}) \right]^{-1} \]
\[
\times \left[ \sum_{i=1}^{k} z_{ig} \nu_{ig} (\Sigma_g^{-1} \odot \eta_{ig} (x_i - \mu_g)^\top) + \sum_{i=1}^{k} z_{ig} \frac{(1 - \nu_{ig})}{\rho_g} (\Sigma_g^{-1} \odot \tilde{\eta}_{ig} (x_i - \mu_g)^\top) \right] 1_p \]
\[
+ \left[ \sum_{i=k+1}^{n} z_{ig} \nu_{ig} (\Sigma_g^{-1} \odot \psi_{ig}) + \sum_{i=k+1}^{n} z_{ig} \frac{(1 - \nu_{ig})}{\sqrt{\rho_g}} (\Sigma_g^{-1} \odot \tilde{\psi}_{ig}) \right]^{-1} \]
\[
\times \left[ \sum_{i=k+1}^{n} z_{ig} \nu_{ig} (\Sigma_g^{-1} \odot \eta_{ig} (x_i - \mu_g)^\top) + \sum_{i=k+1}^{n} z_{ig} \frac{(1 - \nu_{ig})}{\rho_g} (\Sigma_g^{-1} \odot \tilde{\eta}_{ig} (x_i - \mu_g)^\top) \right] 1_p. \]
\[ (A.61) \]

The update for \( \Sigma_g \) has the following form

\[ \hat{\Sigma}_g = \frac{1}{\sum_{i=1}^{k} z_{ig}} \left[ \sum_{i=1}^{n} z_{ig} \nu_{ig} \left\{ (x_i - \mu_g - A_g \eta_{ig}) (x_i - \mu_g - A_g \eta_{ig})^\top \right\} \right. \]
\[
+ A_g (\psi_{ig} - \eta_{ig} \eta_{ig})^\top A_g^\top \left\{ \sum_{i=1}^{n} z_{ig} (1 - \nu_{ig}) \left\{ \frac{1}{\rho_g} (x_i - \mu_g - \sqrt{\rho_g} A_g \tilde{\eta}_{ig}) \right\} \right\} \]
\[
\times \left( x_i - \mu_g - \sqrt{\rho_g} A_g \tilde{\eta}_{ig} \right)^\top + A_g (\tilde{\psi}_{ig} - \tilde{\eta}_{ig} \tilde{\eta}_{ig}^\top A_g^\top) \right\} \]
\[
+ \frac{1}{\sum_{i=k+1}^{n} z_{ig}} \left[ \sum_{i=1}^{n} z_{ig} \nu_{ig} \left\{ (x_i - \mu_g - A_g \eta_{ig}) (x_i - \mu_g - A_g \eta_{ig})^\top \right\} \right. \]
\[
+ A_g (\psi_{ig} - \eta_{ig} \eta_{ig})^\top A_g^\top \left\{ \sum_{i=1}^{n} z_{ig} (1 - \nu_{ig}) \left\{ \frac{1}{\rho_g} (x_i - \mu_g - \sqrt{\rho_g} A_g \tilde{\eta}_{ig}) \right\} \right\} \]
\[
\times \left( x_i - \mu_g - \sqrt{\rho_g} A_g \tilde{\eta}_{ig} \right)^\top + A_g (\tilde{\psi}_{ig} - \tilde{\eta}_{ig} \tilde{\eta}_{ig}^\top A_g^\top) \right\} \].
\[ (A.62) \]

At the second CM-step of the ECM algorithm we update the contamination factor \( \rho_g \).

Looking at the dependence of the Q-function in (A.28) on \( \rho_g \) we have \( Q = Q_1 + Q_2 \)
where

\[
Q_1 = \sum_{i=1}^{k} \sum_{g=1}^{G} z_{ig} (1 - v_{ig}) \left\{ \frac{1}{2} \log |\rho_g^{-1} \Sigma_g^{-1}| - \frac{1}{2} (x_i - \mu_g - (\sqrt{\rho_g} A_g) \bar{\eta}_g) \top \rho_g^{-1} \Sigma_g^{-1} \right. \\
\times \left. (x_i - \mu_g - (\sqrt{\rho_g} A_g) \bar{\eta}_g) \right\} + \text{terms independent of } \rho_g \\
= \sum_{i=1}^{k} \sum_{g=1}^{G} z_{ig} (1 - v_{ig}) \left\{ - \frac{p}{2} \log(\rho_g) - \frac{1}{2 \rho_g} (x_i - \mu_g) \top \Sigma_g^{-1} (x_i - \mu_g) \right. \\
+ \frac{1}{\sqrt{\rho_g}} \left[ (x_i - \mu_g) \top \Sigma_g^{-1} (A_g \bar{\eta}_g) + (A_g \bar{\eta}_g) \top \Sigma_g^{-1} (x_i - \mu_g) \right] \right\} \\
+ \text{terms independent of } \rho_g,
\] (A.63)

and

\[
Q_2 = \sum_{i=k+1}^{n} \sum_{g=1}^{G} z_{ig} (1 - v_{ig}) \left\{ \frac{1}{2} \log |\rho_g^{-1} \Sigma_g^{-1}| - \frac{1}{2} (x_i - \mu_g - (\sqrt{\rho_g} A_g) \bar{\eta}_g) \top \rho_g^{-1} \Sigma_g^{-1} \right. \\
\times \left. (x_i - \mu_g - (\sqrt{\rho_g} A_g) \bar{\eta}_g) \right\} + \text{terms independent of } \rho_g \\
= \sum_{i=k+1}^{n} \sum_{g=1}^{G} z_{ig} (1 - v_{ig}) \left\{ - \frac{p}{2} \log(\rho_g) - \frac{1}{2 \rho_g} (x_i - \mu_g) \top \Sigma_g^{-1} (x_i - \mu_g) \right. \\
+ \frac{1}{\sqrt{\rho_g}} \left[ (x_i - \mu_g) \top \Sigma_g^{-1} (A_g \bar{\eta}_g) + (A_g \bar{\eta}_g) \top \Sigma_g^{-1} (x_i - \mu_g) \right] \right\} \\
+ \text{terms independent of } \rho_g.
\] (A.64)

Then one can use the R function \texttt{optim()} to find the $\rho_g$ which maximizes the function $Q = Q_1 + Q_2$, over the interval $(1, \rho_g^{\text{max}})$. 

\[
Q = Q_1 + Q_2, 
\]
A.3 Dimension Reduction and Clustering using Generalized Hyperbolic Mixtures

A.3.1 Selected genes for the colon data.

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hsa.462</td>
<td>Human serine kinase mRNA</td>
</tr>
<tr>
<td>Hsa.549</td>
<td>Transcription factor IIIA</td>
</tr>
<tr>
<td>Hsa.601</td>
<td>Human aspartyl-tRNA synthetase alpha-2 subunit mRNA</td>
</tr>
<tr>
<td>Hsa.627</td>
<td>Human monocyte-derived neutrophil-activating protein (MONAP) mRNA</td>
</tr>
<tr>
<td>Hsa.773</td>
<td>Macrophage migration inhibitory factor (Human)</td>
</tr>
<tr>
<td>Hsa.821</td>
<td>Human hmgI mRNA for high mobility group protein Y</td>
</tr>
<tr>
<td>Hsa.831</td>
<td>Mitochondrial matrix protein P1 precursor (Human)</td>
</tr>
<tr>
<td>Hsa.957</td>
<td>Human nucleolar protein (B23) mRNA</td>
</tr>
<tr>
<td>Hsa.1832</td>
<td>Myosin regulatory light chain 2, smooth muscle isoform (Human)</td>
</tr>
<tr>
<td>Hsa.2097</td>
<td>Human vasoactive intestinal peptide (VIP) mRNA</td>
</tr>
<tr>
<td>Hsa.2645</td>
<td>H.sapiens ckshs2 mRNA for Cks1 protein homologue</td>
</tr>
<tr>
<td>Hsa.2928</td>
<td>H.sapiens mRNA for p cadherin</td>
</tr>
<tr>
<td>Hsa.3016</td>
<td>S-100P protein (Human)</td>
</tr>
<tr>
<td>Hsa.3306</td>
<td>Human gene for heterogeneous nuclear ribonucleoprotein (hnRNP) core protein A1</td>
</tr>
<tr>
<td>Hsa.3331</td>
<td>Nucleoside diphosphate kinase A (Human)</td>
</tr>
<tr>
<td>Hsa.5971</td>
<td>Human splicing factor SRp30c mRNA</td>
</tr>
<tr>
<td>Hsa.6472</td>
<td>Tubulin beta chain (Haliotis discus)</td>
</tr>
<tr>
<td>Hsa.6814</td>
<td>Collagen alpha 2(XI) chain (Homo sapiens)</td>
</tr>
<tr>
<td>Hsa.8125</td>
<td>Human</td>
</tr>
<tr>
<td>Hsa.8147</td>
<td>Human desmin gene</td>
</tr>
<tr>
<td>Hsa.36689</td>
<td>H.sapiens mRNA for GCAP-II/uroguanylin precursor</td>
</tr>
<tr>
<td>Hsa.36952</td>
<td>Complement factor D precursor (Homo sapiens)</td>
</tr>
<tr>
<td>Hsa.37937</td>
<td>Myosyn heavy chain, nonmuscle (Gallus gallus)</td>
</tr>
</tbody>
</table>