Mixtures of Shifted Asymmetric Laplace Distributions

by

Brian C. Franczak

A thesis
presented to
The University of Guelph

In partial fulfilment of requirements
for the degree of
Doctor of Philosophy
in
Statistics

Guelph, Ontario, Canada

© Brian C. Franczak, March, 2014
ABSTRACT

Mixtures of Shifted Asymmetric Laplace Distributions

Brian C. Franczak
University of Guelph, 2014

Advisors:
Professor Paul D. McNicholas
Professor Ryan P. Browne

In this thesis we introduce a mixture of shifted asymmetric Laplace (SAL) distributions for model-based clustering and classification. The mixture of SAL distributions allows for the parameterization of skewness as well as location and scale. Furthermore, we extend the general SAL mixture by decomposing the component scale matrices; this results in two families of SAL mixture models and a generalization of the multivariate SAL density. In developing these models we review and utilize several facets of model-based clustering. Specifically, to estimate the parameters of our mixture models we use the well-known expectation-maximization algorithm, to choose the best fitting mixture model we consider both the Bayesian information criterion and integrated completed likelihood, and to evaluate classification performance we use the adjusted Rand index. Both simulated and real data are used to demonstrate our models.
For my uncle Ed (1938 - 2010).
I wish to extend my sincerest gratitude to my advisors, Dr. Paul McNicholas and Dr. Ryan Browne. To Dr. McNicholas, thank you for your encouraging words, support and patience. You taught me the value of a well-written paper and gave me the opportunity to become a part of the statistical community, both at home and abroad. To Dr. Browne, thank you for your patience, your “open-door” policy and for demonstrating many invaluable coding and problem-solving techniques.

I would also like to thank Dr. Zeny Feng and Dr. Steven Rothstein for agreeing to be a part of my advisory committee, Dr. Sanjeena Dang and Dr. Rob Deardon for agreeing to be a part of my examining committee and to Dr. Doug Steinley, thank you for agreeing to serve as the external examiner on this work and for travelling to Guelph for my examination.

To my mother (Lee), father (Rick), brother (Andrew) and grandparents (Lorraine and Lloyd; 1930 - 2007), thank you for your unwavering support, patience and understanding. Without you guys I am certain this wouldn’t have been possible. To the McNicholas research group, thank you for making this experience so enjoyable. I wish to thank specifically two collaborators, Dr. Cristina Tortora and Paula Murray, the soon-to-be-Dr. Irene Vrbik and the “t-clust” guy, Dr. Jeffrey Andrews, for all the helpful conversations. To Jeff, thanks for all the quotable material.

I would also like to acknowledge the financial support I have received through an
Ontario Graduate Scholarship in Science and Technology (OGSST), an Ontario Graduate Scholarship (OGS) and NSERC Discovery Grants from the Natural Sciences and Engineering Research Council of Canada and Ontario Ministry of Research and Innovation’s Early Researcher Award.
Publications

The following articles based on the work in this thesis have been accepted for publication, submitted for publication or are in preparation.


# Contents

1 Introduction ................................................. 1
   1.1 Finite Mixture Models .................................. 1
   1.2 Outline .................................................. 3
       1.2.1 Chapter 2 ........................................ 3
       1.2.2 Chapter 3 ........................................ 4
       1.2.3 Chapter 4 ........................................ 4
       1.2.4 Chapter 5 ........................................ 5
       1.2.5 Chapter 6 ........................................ 5
       1.2.6 Chapter 7 ........................................ 5
   1.3 Impact .................................................. 6

2 Background .................................................. 7
   2.1 Normal Variance-Mean Mixtures .......................... 7
   2.2 Generalized Inverse Gaussian Distribution ............. 8
   2.3 Generalized Hyperbolic Distribution ..................... 9
   2.4 Families of Finite Mixture Models ........................ 10
       2.4.1 Gaussian Parsimonious Clustering Models ...... 10
       2.4.2 Parsimonious Gaussian Mixture Models ......... 12
   2.5 Asymmetric Laplace Distributions ........................ 14
       2.5.1 Univariate Representation ....................... 15
       2.5.2 Shifted Asymmetric Laplace ..................... 15
   2.6 The Expectation Maximization Algorithm .................. 18
       2.6.1 MM Algorithms .................................. 18
       2.6.2 The EM Algorithm ................................ 19
       2.6.3 Extensions ....................................... 21
       2.6.4 Deterministic Annealing ......................... 22
   2.7 Convergence Criterion ................................... 24
       2.7.1 Aitken acceleration ................................ 24
   2.8 Model Selection and Performance ........................ 25
       2.8.1 Bayesian Information Criterion .................. 26
2.8.2 Integrated Completed Likelihood ........................................... 27
2.9 Performance Assessment ......................................................... 27
2.9.1 Adjusted Rand Index .......................................................... 27
2.10 Model-Based Classification ..................................................... 28

3 Mixtures of Shifted Asymmetric Laplace Distributions .............. 29
3.1 Introduction .................................................................................. 29
3.2 Parameter Estimation .................................................................. 30
   3.2.1 EM Algorithm ........................................................................ 30
   3.2.2 Application to Mixture of SAL Distributions ......................... 31
   3.2.3 Initialization ......................................................................... 34
   3.2.4 Dealing with Infinite Log-Likelihood Values ......................... 35
   3.2.5 Model-Based Classification .................................................. 36
3.3 Data Analyses ............................................................................. 36
   3.3.1 Introduction ......................................................................... 36
   3.3.2 Simulation Study .................................................................. 37
   3.3.3 Old Faithful Geyser Data ..................................................... 40
   3.3.4 Yeast Data ........................................................................... 42
3.4 Summary and Discussion ............................................................ 46

4 Parsimonious Variants ................................................................. 49
4.1 Introduction ................................................................................. 49
4.2 The ParSAL Family of Models .................................................. 50
4.3 Parameter Estimation ................................................................. 51
   4.3.1 EM Algorithm ........................................................................ 51
   4.3.2 Application to the ParSAL Family ......................................... 51
   4.3.3 Fitting an EVV Model .......................................................... 53
   4.3.4 A “Push-Back” Procedure ................................................... 55
4.4 Data Analyses ............................................................................. 56
   4.4.1 Introduction ......................................................................... 56
   4.4.2 The ‘Pizza’ Data ................................................................. 56
   4.4.3 Hematopoietic Stem Cell Transplant Data .............................. 58
4.5 Summary and Discussion ............................................................ 60

5 Extending the Mixture of Factor Analyzers .............................. 62
5.1 Introduction ................................................................................. 62
5.2 The Model ................................................................................... 63
5.3 Parameter Estimation ................................................................. 64
   5.3.1 A Two-Phase Approach ...................................................... 64
   5.3.2 Woodbury Identity .............................................................. 68
   5.3.3 Model-Based Classification ................................................ 69
5.4 Data Analyses ................................................................. 69
  5.4.1 Introduction ............................................................... 69
  5.4.2 Simulation Study ........................................................ 69
  5.4.3 Leptograpsus Crabs ...................................................... 70
  5.4.4 Gaussian Cluster Merging & Swiss Bank Notes .................... 73
  5.4.5 Yeast Data ................................................................. 74
  5.4.6 Model-Based Classification .......................................... 74

5.5 Summary and Discussion .................................................. 76

6 Multiple Scaled Shifted Asymmetric Laplace Distributions ............ 78
  6.1 Introduction ................................................................. 78
  6.2 Multiple Scaled Distributions .......................................... 79
  6.3 Mixtures of Multiple Scaled Shifted Asymmetric Laplace distributions ......................................................... 81
  6.4 Parameter Estimation ...................................................... 83
    6.4.1 E-step ................................................................. 85
    6.4.2 M-step ................................................................. 86
  6.5 Applications ............................................................... 89
    6.5.1 Introduction ........................................................... 89
    6.5.2 Leptograpsus Crabs .................................................. 90
    6.5.3 Italian Wines Data .................................................... 90
  6.6 Summary and Discussion ................................................ 92

7 Conclusion ............................................................................ 94
  7.1 Summary ................................................................. 94
    7.1.1 Mixtures of shifted asymmetric Laplace distributions ........ 94
    7.1.2 Extensions .............................................................. 95
    7.1.3 Multiple scaled shifted asymmetric Laplace distributions ....... 96
  7.2 Future Work ............................................................... 96
    7.2.1 Infinite Likelihoods .................................................... 96
    7.2.2 Mixtures of Common SAL Factor Analyzers ....................... 97
    7.2.3 Model Selection ........................................................ 97

Bibliography ............................................................................ 98

A Calculations for the Mixture of SALs ...................................... 107
  A.1 The distribution of \( W \mid X = x \) ..................................... 107
  A.2 Maximum Likelihood Estimates ........................................ 109

B Calculations for the MSSAL .................................................. 112
  B.1 Maximum Likelihood Estimates ........................................ 112
List of Tables

2.1 Nomenclature, covariance decomposition and number of free covariance parameters for the Gaussian parsimonious clustering models. .................................................................................. 12
2.2 Nomenclature, component scale matrix structure, and number of free covariance parameters for each parsimonious Gaussian mixture models. .................................................. 14
3.1 Proportion of times $G = 2$ components were selected as well as average ARI values (with std. dev.) for the SAL and Gaussian mixtures, respectively, applied to the simulated data. ................................................................. 39
3.2 Clustering results for the chosen SAL and Gaussian mixture models for the yeast data. The localization sites are cross-tabulated against our predicted classifications (A, B, C) in each case. .......................................................... 44
3.3 Clustering results for the two-component Gaussian mixture model for the yeast data. The localization sites are cross-tabulated against our predicted classifications (A, B). .................................................................................. 44
3.4 Aggregate classification results for the SAL and Gaussian mixture models for the yeast data with 70% of the labels taken as known. In each case, the localization sites are cross-tabulated against our predicted classifications (A, B) for the observations with unknown labels. .................................................. 46
4.1 Clustering results for the chosen ParSAL, GPCM and tEIGEN models. The fluorescent dyes are cross-tabulated against our predicted classifications (A, B, C, D) in each case. .................................................. 60
5.1 Clustering results for the chosen PSALM and PGMM for the simulated data. .................................................................................................................. 70
5.2 Clustering results for the chosen PSALM and PGMM for the Swiss bank notes data. The bill types are cross-tabulated against our predicted classifications (A, B) in each case. .................................................. 73
5.3 Clustering results for the chosen PSALM and PGMM for the yeast data. The cellular localization sites are cross-tabulated against our predicted classifications (A, B, C, D, E, F) in each case. .................................................. 74
5.4 Aggregated classification results for the PSALM CCCU (ARI = 0.853) and PGMM UCUU (ARI = 0.737) for the Leptograpsus crabs data. The colour and gender combinations are cross-tabulated against our predicted classifications (A, B, C, D) in each case.

5.5 Clustering results for the chosen PSALM and PGMM for the yeast data. The localization sites are cross-tabulated against our predicted classifications (A, B) in each case.

6.1 The thirteen chemical and physical properties of the Italian wines.

6.2 Classification results for the mixture of MSSAL distribution (ARI = 0.829) and GMM (ARI = 0.461) for the Italian wines data. The types of wine are cross-tabulated against our predicted classifications (A, B, C) in each case.
# List of Figures

2.1 Bivariate SAL contours with $\mu = (0, 0)$ and varying $\alpha$ and $\zeta$. ............... 17

3.1 Example of a simulated data set ($n = 500$), coloured by true component. .... 38
3.2 One of the $G = 5$ component Gaussian solutions for the simulated data, coloured by MAP classification results. ................................. 39
3.3 Model-based clustering results with contours for the SAL and Gaussian mixture models (GMMs) on the Old Faithful data. ................................. 41
3.4 3D contours for the SAL and Gaussian mixture models (GMMs) on the Old Faithful data. ................................................................. 42
3.5 The yeast data with the CYT and ME3 location sites highlighted. ............. 43
3.6 Clustering results for the SAL (left) and Gaussian (right) mixture models on the yeast data; the colours highlight the predicted component memberships. 45

4.1 Clustering results for the best fitting ParSAL (Panel 1), GPCM (Panel 2) and tEIGEN (Panel 3) models on the ‘pizza’ data; the colours highlight the predicted component memberships. ................................. 57
4.2 The true cell groupings (represented by colour) and clustering results for the ParSAL VEV (represented by symbol). ................................. 59

5.1 An example of the predicted group memberships for the chosen PSALM and PGMM models on a simulated data set with $p = 10$ dimensions. ................. 71
5.2 The first and third principal components of the crabs data with contours reprinting the fit of the chosen PSALM and PGMM models. ................. 72
5.3 The yeast data with predicted group memberships highlighted for the chosen PSALM and PGMM models. ............................................ 75

6.1 Bivariate contour plots of the multiple scaled SAL density with $\mu = (0, 0)'$, $A = \text{diag}(1, 1)$ and varying $\beta$ and $\zeta$. ............................................ 84
6.2 The first and third principal components of the crabs data with contours reprinting the fit of the chosen MSSAL and GMM models. ....................... 91
Chapter 1

Introduction

1.1 Finite Mixture Models

Finite mixture models are based on the underlying assumption that a population is a convex combination of a finite number of densities. They therefore lend themselves quite naturally to classification and clustering problems. Formally, a random vector $X$ arises from a parametric finite mixture distribution if, for all $x \subset X$, we can write its density as

$$f(x \mid \vartheta) = \sum_{g=1}^{G} \pi_g f_g(x \mid \theta_g), \quad (1.1)$$

where $\pi_g > 0$, such that $\sum_{g=1}^{G} \pi_g = 1$, are the mixing proportions, $f_1(x \mid \theta_1), \ldots, f_G(x \mid \theta_G)$ are called component densities, and $\vartheta = (\pi, \theta_1, \ldots, \theta_G)$ is the vector of parameters with $\pi = (\pi_1, \ldots, \pi_G)$. 
The component densities are usually taken to be of the same type, most often multivariate Gaussian. In the event that the component densities are multivariate Gaussian, the density of the mixture model is

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

\[ = \sum_{g=1}^{G} \frac{\pi_g}{(2\pi)^{p/2}\sqrt{\Sigma_g}} \exp \left\{ -\frac{1}{2} (x_i - \mu_g) \Sigma_g^{-1} (x_i - \mu_g) \right\}, \tag{1.2} \]

where \( \pi_g \) are defined for (1.1), \( \mu_g \) is the component location parameter and \( \Sigma_g \) is the component covariance matrix. The popularity of the multivariate Gaussian distribution is due to its mathematical tractability and its flexibility in capturing densities; we will return to this latter point in Section 3.3.1. Herein, we shall follow convention and use the term model-based clustering to mean clustering using mixture models. Model-based classification (e.g., McNicholas, 2010), or partial classification (cf. McLachlan, 1992, Section 2.10), can be regarded as a semi-supervised version of model-based clustering.

At the time of the review paper of Fraley and Raftery (2002), almost all work on clustering and classification using mixture models had been based on Gaussian mixture models (e.g., Banfield and Raftery, 1993; Celeux and Govaert, 1995; Ghahramani and Hinton, 1997; McLachlan and Peel, 2000; Tipping and Bishop, 1999a, amongst others). An important example of non-Gaussian work from this time is the early work on clustering using mixtures of multivariate \( t \)-distributions carried out by McLachlan and Peel (1998) and Peel and McLachlan (2000). This work was the forerunner to several papers on clustering using mixtures of multivariate \( t \)-distributions, including those by McLachlan et al. (2007), Andrews and McNicholas (2011b), Baek and McLachlan (2011), and McNicholas and Subedi (2012). Work has also burgeoned on skew-normal distributions (e.g., Lin, 2009), skew-\( t \) distri-
butions (e.g., Lin, 2010; Lee and McLachlan, 2011; Vrbik and McNicholas, 2012), and other non-elliptically contoured distributions (e.g., Karlis and Santourian, 2009; Browne et al., 2012).

The increase of non-Gaussian approaches to model-based clustering and classification has coincided with yet more papers on Gaussian approaches. These include work on extensions of mixtures of factor analyzers (McNicholas and Murphy, 2008, 2010; Baek et al., 2010) as well as developments on variable selection and dimension reduction for Gaussian model-based clustering (Raftery and Dean, 2006; Maugis et al., 2009; Scrucca, 2010). Nevertheless, the fecundity of non-Gaussian approaches has certainly been more potent than that of Gaussian approaches over the last few years.

This thesis focuses on the development of a mixture of non-Gaussian distributions that allow for the parameterization of skewness, in addition to location and scale. Specifically, we introduce a mixture of multivariate shifted asymmetric Laplace (SAL) distributions for clustering and classification.

1.2 Outline

1.2.1 Chapter 2

Normal variance-mean mixtures are defined and a review of distributions related to the asymmetric Laplace distribution is given. Families of Gaussian mixture models are reviewed and the shifted asymmetric Laplace density is formally defined. In addition, the Expectation-Maximization (EM) algorithm and its extensions, popular model selection criterion, and
techniques for evaluating clustering performance are discussed.

1.2.2 Chapter 3

A mixture of multivariate shifted asymmetric Laplace distributions is introduced and a parameter estimation scheme is outlined using an EM algorithm. To avoid common pitfalls in EM estimation a deterministic annealing algorithm is presented to obtain reliable starting values. The mixture of SAL distributions is applied to both simulated and real data, including the famous Old Faithful Geyser Data. Compared to a mixture of multivariate Gaussian distributions our SAL mixtures give superior clustering results.

1.2.3 Chapter 4

A parsimonious family of the shifted asymmetric Laplace mixture models are developed using the eigen-decomposition of the component scale matrices. The nomenclature and number of free covariance parameters of each family member are analogous to the famous Gaussian parsimonious clustering models \((\text{Celeux and Govaert, 1995})\). In total fourteen parsimonious SAL (ParSAL) mixture models are developed for model-based clustering and classification. We compare our Parsimonious SAL models to both their multivariate Gaussian and multivariate-\(t\) analogues using simulated and real data.
1.2.4 Chapter 5

In Chapter 5 the methodology surrounding the development of a mixture of modified shifted asymmetric Laplace factor analyzers is presented. Similarly to Chapter 4 the component scale matrices of the SAL mixture model are decomposed and the elements of the decomposition constrained. In total 12 parsimonious shifted asymmetric Laplace mixtures (PSALM) are created, each having component scale parameters that are linear in $p$. The PSALM models are compared to the parsimonious Gaussian mixture models (PGMM) using both simulated and real data and give excellent classification results.

1.2.5 Chapter 6

In Chapter 6 the eigen-decomposition of the SALs component scale matrices is utilized to incorporate a multidimensional weight function. This creates a representation of the multivariate SAL distribution whose marginal distributions are also asymmetric Laplace. Contour plots showing the unique shapes of this representation are presented. Parameter estimates are derived and an EM algorithm for these models is outlined. We demonstrate our model using real data, where it is shown to out-preform a mixture of multivariate Gaussian distributions.

1.2.6 Chapter 7

In the final chapter the ideas and methods demonstrated in this work are summarized. Suggestions for future study arising from this body of work are presented.
1.3 Impact

The impact of the ideas proposed in this thesis on the current model-based clustering literature can be summarized into the following points:

- A mixture of shifted asymmetric Laplace distributions is introduced. This model extends the current literature on non-Gaussian mixture models for clustering and classification. The parameter estimation procedure for the SAL mixtures is shown to be both mathematically elegant and computationally appealing. These models show greater flexibility and the ability to recover certain data sets true group structure compared to a mixture of Gaussian distributions.

- Two families of SAL distributions, with parsimonious covariance structures, are introduced. The first family utilizes the well-known eigen-decomposition and has the potential for development into R software. The second family extends the mixture of factor analyzers model and are characterized by a covariance structure whose number of free parameters grows linearly with the dimension of the data. Both of these families are demonstrated using simulated and real data and perform favourably compared to the available symmetric alternatives.

- An alternative parameterization of the mixture of SALs density is considered. This allows for the incorporation of a multidimensional weight function. Only one other skewed distribution with this property has been proposed in the literature however, to our knowledge, no work describing a parameter estimation scheme or a demonstration of its abilities using real data have been given.
Chapter 2

Background

2.1 Normal Variance-Mean Mixtures

Consider a random variable $w \in \mathbb{R}^+$ from a probability distribution $F$ on $[0, \infty)$. According to Barndorff-Nielsen et al. (1982) the distribution of a $p$-dimensional random variable $X$ is a normal variance-mean mixture if $X \mid W = w$ follows a multivariate Gaussian distribution with mean $\mu + w\alpha$ and covariance matrix $w\Sigma$ where $\mu$ is a location parameter, $\alpha$ is a skewness or drift parameter and $\Sigma$ is a $p \times p$ covariance matrix. If $\alpha = 0$ then a normal variance mixture is obtained.

There are several well-known multivariate distributions that arise as mixtures of Normal distributions. For example, the multivariate-$t$ distribution (Kotz and Nadarajah, 2004), the generalized hyperbolic distribution (McNeil et al., 2005) and the asymmetric Laplace distribution (Kotz et al., 2001) can all be expressed in terms of a multivariate Gaussian random
variable (cf. Section 2.5). A thorough review of the properties of normal variance-mean mixtures is given in Barndorff-Nielsen et al. (1982).

2.2 Generalized Inverse Gaussian Distribution

The first proposition of a generalized inverse Gaussian distribution was in Good (1953). Formally, the density of a random variable $X$ following a generalized inverse Gaussian (GIG) distribution is given by

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

(2.1)

for $x > 0$, where $a, b \in \mathbb{R}^+, \nu \in \mathbb{R}$, and $K_\nu$ is the modified Bessel function of the third kind with index $\nu$. There are several special cases of the GIG distribution, such as the gamma distribution ($b = 0, \nu > 0$), the inverse Gaussian distribution ($\nu = -1/2$), the reciprocal gamma variate ($a = 0, \nu < 0$) and the reciprocal inverse Gaussian variate ($\nu = 1/2$).

The GIG distribution has been extensively studied within the literature. Blæsild (1978) computed moments, cumulants and studied the shape of the density, Halgreen (1979) investigated probabilistic properties, and Barndorff-Nielsen (1977, 1978) came across the distribution while studying a representation of the hyperbolic distributions as a mixture of normal distributions. Following this discovery Jørgensen (1982) outlined the properties of the GIG distribution in great detail.

For our purposes the most attractive properties of the GIG distribution are the tractability
of the following expected values:

\[
\mathbb{E}[X] = \sqrt{\frac{b}{a}} R_{\nu} \left( \sqrt{ab} \right), \\
\mathbb{E}[1/X] = \sqrt{\frac{a}{b}} R_{\nu} \left( \sqrt{ab} \right) - \frac{2\nu}{b},
\]

where \( R_{\nu}(z) := \frac{K_{\nu+1}(z)}{K_{\nu}(z)} \) and all other parameters are previously defined.

### 2.3 Generalized Hyperbolic Distribution

As mentioned in the previous section Barndorff-Nielsen and Halgreen (1977) came across the GIG distribution while studying representations of a hyperbolic distribution. Nearly 30 years after this paper McNeil et al. (2005) wrote that the probability density function of a random variable \( Y \) following the multivariate generalized hyperbolic distribution can be written

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

where \( \delta(x, \mu \mid \Sigma) = (x - \mu)' \Sigma^{-1} (x - \mu) \) is the square Mahalanobis distance between \( x \) and the location parameter \( \mu \in \mathbb{R}^p, \alpha \in \mathbb{R}^p \) is a skewness parameter, \( \Sigma \) is a \( p \times p \) scale matrix, \( \lambda \) is an index parameter, and \( \chi \) and \( \psi \) are concentration parameters.

The multivariate generalized hyperbolic distribution has a number of limiting cases. For example, the multivariate normal-inverse Gaussian distribution (Karlis and Santourian, 2009),...
the skew-t and multivariate skew-t distributions, the skew-normal distribution and the Gaussian distribution. Specifically, if $\lambda = 1$ the multivariate generalized hyperbolic with univariate marginals following the univariate hyperbolic distribution is obtained, if $\lambda = (p + 1)/2$ the $p$-dimensional hyperbolic distribution is obtained, if $\lambda = -1/2$ the inverse Gaussian distribution is obtained and if $\lambda = 1$, $\psi = 2$, and $\chi \to 0$ the multivariate asymmetric Laplace distribution (Kotz et al., 2001) is obtained.

### 2.4 Families of Finite Mixture Models

Families of finite mixture models are created by imposing a combination of constraints on the component densities; most-often on the component covariance or scale matrices. In the following subsections two famous families of parsimonious Gaussian mixture models are reviewed. Note: The general Gaussian mixture model, given in (1.2), has

$$(G - 1) + Gp + Gp(p + 1)/2$$

free parameters, of which $Gp(p + 1)/2$ are from the group covariance matrices $\Sigma_1, \ldots, \Sigma_G$.

#### 2.4.1 Gaussian Parsimonious Clustering Models

The most well-known family of mixture models are the Gaussian parsimonious clustering models (GPCM) introduced in Celeux and Govaert (1995). The GPCMs result from decomposing the component covariance matrix $\Sigma_g$ in (1.2) using the eigen (or spectral) decomposition (cf. Banfield and Raftery, 1993; Flury et al., 1994). The eigen-decomposition of the
component covariance matrix is given by,

\[ \Sigma_g = \lambda_g D_g A_g D_g', \]  

(2.4)

where \( \lambda_g = |\Sigma_g|^{1/p} \) are the associated constants of proportionality, \( D_g \) is an orthogonal matrix of eigenvectors, and \( A_g \) is a diagonal matrix with entries proportional to the eigenvalues, such that \( |A_g| = 1 \).

By imposing a combination of the constraints \( D_g = D, \, D = I_d, \, A_g = A, \, A_g = I_d, \) and \( \lambda_g = \lambda \) on (2.4) a family of fourteen models arise (Table 2.1). All fourteen GPCMs are available in the R (R Core Team, 2013) packages Rmixmod (Lebret et al., 2012) and mixture (Browne and McNicholas, 2012, 2014, 2013), and ten are available in the R package mclust (Fraley and Raftery, 2002). Recently, analogues of the GPCMs have been developed for model-based clustering and classification. Specifically, Andrews and McNicholas (2012, 2013) introduced a family of multivariate t-analogues, available as the teigen package for R and Vrbik and McNicholas (2014) introduced skew-normal and skew-t analogues. These families offer robust and flexible alternatives compared to the well-established family of Gaussian mixture models. Note: For the tEIGEN models the nomenclature uses the letter ‘C’ to denote that a constraint is imposed, the letter ‘U’ to denote that a constraint is not imposed, and the letter ‘I’ to denote that the identity matrix of suitable dimension is taken to be the matrix in question. For each member of the tEIGEN family, the relevant letters indicate the constraints on \( \lambda_g; \, D_g; \, A_g, \) and \( \nu_g \), respectively.

Table 2.1 shows that some of the GPCMs significantly reduce the number of free parameters in the component covariance matrices however, eight of the fourteen models still have a number of component covariance parameters that are quadratic in \( p \). To this end an alternative
Table 2.1: Nomenclature, covariance decomposition and number of free covariance parameters for the Gaussian parsimonious clustering models.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\lambda_g = \lambda$</th>
<th>$A_g = A$</th>
<th>$D_g = D$</th>
<th>Covariance Decomposition</th>
<th>Number of Covariance Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>EII</td>
<td>Equal</td>
<td>Identity</td>
<td>Identity</td>
<td>$\lambda I$</td>
<td>1</td>
</tr>
<tr>
<td>VII</td>
<td>Variable</td>
<td>Identity</td>
<td>Identity</td>
<td>$\lambda_g I$</td>
<td>$G$</td>
</tr>
<tr>
<td>EEI</td>
<td>Equal</td>
<td>Equal</td>
<td>Identity</td>
<td>$\lambda A$</td>
<td>$p$</td>
</tr>
<tr>
<td>VVI</td>
<td>Variable</td>
<td>Equal</td>
<td>Identity</td>
<td>$\lambda A$</td>
<td>$G + (p - 1)$</td>
</tr>
<tr>
<td>EVI</td>
<td>Equal</td>
<td>Variable</td>
<td>Identity</td>
<td>$\lambda A_g$</td>
<td>$Gp - (G + 1)$</td>
</tr>
<tr>
<td>VVI</td>
<td>Variable</td>
<td>Variable</td>
<td>Identity</td>
<td>$\lambda A_g$</td>
<td>$Gp$</td>
</tr>
<tr>
<td>EEE</td>
<td>Equal</td>
<td>Equal</td>
<td>Equal</td>
<td>$\lambda DAD'$</td>
<td>$p(p + 1)/2$</td>
</tr>
<tr>
<td>VEE</td>
<td>Variable</td>
<td>Equal</td>
<td>Equal</td>
<td>$\lambda DAD'$</td>
<td>$p(p + 1)/2 + (G - 1)$</td>
</tr>
<tr>
<td>EVE</td>
<td>Equal</td>
<td>Variable</td>
<td>Equal</td>
<td>$\lambda DAD'$</td>
<td>$p(p + 1)/2 - (G - 1)(p - 1)$</td>
</tr>
<tr>
<td>VEE</td>
<td>Variable</td>
<td>Variable</td>
<td>Equal</td>
<td>$\lambda DAD'_g$</td>
<td>$Gp(p + 1)/2 - (G - 1)p$</td>
</tr>
<tr>
<td>VEV</td>
<td>Variable</td>
<td>Equal</td>
<td>Variable</td>
<td>$\lambda DAD'_g$</td>
<td>$Gp(p + 1)/2 - (G - 1)(p - 1)$</td>
</tr>
<tr>
<td>EVV</td>
<td>Equal</td>
<td>Variable</td>
<td>Variable</td>
<td>$\lambda DAD'_g$</td>
<td>$Gp(p + 1)/2 - (G - 1)$</td>
</tr>
<tr>
<td>VVV</td>
<td>Variable</td>
<td>Variable</td>
<td>Variable</td>
<td>$\lambda DAD'_g$</td>
<td>$Gp(p + 1)/2$</td>
</tr>
</tbody>
</table>

decomposition of the component covariance matrix is discussed in the following section. This decomposition significantly reduces the number of free covariance parameters by assuming there exist underlying latent variables of (much) lower dimension.

### 2.4.2 Parsimonious Gaussian Mixture Models

Factor analysis (Spearman, 1904) is a dimension reduction technique that assumes a $p$-dimensional random vector $X$ can be modelled using a $q$-dimensional vector of unobserved (or latent) factors $U$, where $q \ll p$. The factor analysis model can be written,

$$X = \mu + \Lambda U + \epsilon,$$

where $\Lambda$ is a $p \times q$ matrix of factor loadings, the random vector $u \sim \mathcal{N}(0, I_q)$ is a vector of factors, and $\epsilon \sim \mathcal{N}(0, \Psi)$ is a vector of error terms with $\Psi = \text{diag}(\psi_1, \psi_2, \ldots, \psi_p)$. It follows
that the marginal distribution of $X$ is multivariate Gaussian with mean $\mu$ and covariance matrix $\Sigma = \Lambda \Lambda' + \Psi$.

The mixture of factor analyzers model is given by

$$f(x | \vartheta) = \sum_{g=1}^{G} \frac{\pi_g}{\sqrt{(2\pi)^p|\Lambda_g \Lambda'_g + \Psi_g|}} \exp \left\{ -\frac{1}{2} (x_i - \mu_g)' (\Lambda_g \Lambda'_g + \Psi_g)^{-1} (x_i - \mu_g) \right\}, \quad (2.5)$$

where $\pi_g$ and $\mu_g$ are defined as before, $\Lambda_g$ is a $p \times q$ matrix of factor loadings and $\Psi_g$ is a $p \times p$ diagonal matrix with positive entries. The mixture of factor analyzers model was further developed by Ghahramani and Hinton (1997), who studied the equal noise (UCCU) model, and McLachlan and Peel (2000), who studied the unequal noise (UUUU) model. Tipping and Bishop (1999b) studied a closely related mixture of probabilistic principal component analyzers model, where $\Psi_g = \psi_g I_p$.

Subsequently, McNicholas and Murphy (2008) created a family of eight parsimonious models by imposing constraints across the groups on $\Lambda_g$ and $\Psi_g$. McNicholas and Murphy (2010) then modified the factor analysis component covariance structure by setting $\Psi_g = \omega_g \Delta_g$, where $\Delta_g$ is a $p \times p$ diagonal matrix with $|\Delta_g| = 1$, and $\omega_g \in \mathbb{R}^+$. The mixture of modified factor analyzers model is given by

$$f(x | \vartheta) = \sum_{g=1}^{G} \frac{\pi_g}{\sqrt{(2\pi)^p|\Lambda_g \Lambda'_g + \omega_g \Delta_g|}} \times \exp \left\{ -\frac{1}{2} (x_i - \mu_g)' (\Lambda_g \Lambda'_g + \omega_g \Delta_g)^{-1} (x_i - \mu_g) \right\}, \quad (2.6)$$

where $\pi_g$, $\mu_g$, $\Lambda_g$ and $\Psi_g$ are defined for (2.5).

By imposing valid combinations of the constraints $\Lambda_g = \Lambda$, $\Delta_g = I_p$, $\Delta_g = \Delta$, and $\omega_g = \omega$,
a family of 12 mixtures of modified factor analyzers emerges (cf. McNicholas and Murphy, 2010). This family of models are called parsimonious Gaussian mixture models (PGMM; Table 2.2). Note: The PGMMs are available for clustering and classification applications as the R package pgmm (McNicholas et al., 2011).

Table 2.2: Nomenclature, component scale matrix structure, and number of free covariance parameters for each parsimonious Gaussian mixture models.

<table>
<thead>
<tr>
<th>PGMM Nomenclature</th>
<th>Component Scale Matrix</th>
<th>Number of Free Scale Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>C C C C</td>
<td>$\Sigma_g = \Lambda \Lambda' + \omega I_p$</td>
<td>$[pq - q(q - 1)/2] + 1$</td>
</tr>
<tr>
<td>C C U C</td>
<td>$\Sigma_g = \Lambda \Lambda' + \omega_g I_p$</td>
<td>$[pq - q(q - 1)/2] + G$</td>
</tr>
<tr>
<td>U C C C</td>
<td>$\Sigma_g = \Lambda_g \Lambda_g' + \omega I_p$</td>
<td>$G[pq - q(q - 1)/2] + 1$</td>
</tr>
<tr>
<td>U C C C</td>
<td>$\Sigma_g = \Lambda_g \Lambda_g' + \omega_g I_p$</td>
<td>$G[pq - q(q - 1)/2] + G$</td>
</tr>
<tr>
<td>C C C U</td>
<td>$\Sigma_g = \Lambda \Lambda' + \omega \Delta$</td>
<td>$[pq - q(q - 1)/2] + p$</td>
</tr>
<tr>
<td>C C U U</td>
<td>$\Sigma_g = \Lambda \Lambda' + \omega \Delta$</td>
<td>$[pq - q(q - 1)/2] + [G + (p - 1)]$</td>
</tr>
<tr>
<td>U C C C</td>
<td>$\Sigma_g = \Lambda_g \Lambda_g' + \omega \Delta$</td>
<td>$G[pq - q(q - 1)/2] + [G + (p - 1)]$</td>
</tr>
<tr>
<td>U C U U</td>
<td>$\Sigma_g = \Lambda \Lambda' + \omega \Delta_g$</td>
<td>$[pq - q(q - 1)/2] + [1 + G(p - 1)]$</td>
</tr>
<tr>
<td>C U U U</td>
<td>$\Sigma_g = \Lambda_g \Lambda_g' + \omega \Delta_g$</td>
<td>$[pq - q(q - 1)/2] + p$</td>
</tr>
<tr>
<td>U U C U</td>
<td>$\Sigma_g = \Lambda \Lambda' + \omega \Delta_g$</td>
<td>$G[pq - q(q - 1)/2] + [1 + G(p - 1)]$</td>
</tr>
<tr>
<td>U U U U</td>
<td>$\Sigma_g = \Lambda \Lambda' + \omega \Delta_g$</td>
<td>$G[pq - q(q - 1)/2] + Gp$</td>
</tr>
<tr>
<td>U U U U</td>
<td>$\Sigma_g = \Lambda \Lambda' + \omega \Delta_g$</td>
<td>$G[pq - q(q - 1)/2] + Gp$</td>
</tr>
</tbody>
</table>

2.5 Asymmetric Laplace Distributions

In the following section we pay special attention to asymmetric Laplace distributions. Specifically, we define useful representations of both a univariate and multivariate density.
2.5.1 Univariate Representation

A random variable $X$ following an univariate asymmetric Laplace distribution has density

$$g(x) = \frac{1}{\gamma} \exp \left\{ -\frac{|x - \mu|}{\sigma^2} \left( \gamma - \alpha \text{sign}(x - \mu) \right) \right\},$$

(2.7)

where $\gamma = \sqrt{\alpha^2 + 2\sigma^2}$, $\alpha$ is a skewness parameter, $\sigma^2$ is a scale parameter and $\mu$ is a location parameter. When $\mu = 0$ and $\sigma^2 = 1$ the distribution is said to be standard and when $\sigma^2 = \mu = 0$ and $\alpha \neq 0$ an exponential distribution with mean $\alpha$ is obtained.

2.5.2 Shifted Asymmetric Laplace

Consider a $p$-dimensional random vector $V$ from an asymmetric Laplace distribution (see Kotz et al., 2001); for clarity, we shall refer to this distribution as the centralized asymmetric Laplace (CAL) distribution hereafter. The density of $V$ is given by

$$f(v | \alpha, \Sigma) = \frac{2 \exp \{v' \Sigma^{-1} \alpha\}}{(2\pi)^{p/2} |\Sigma|^{1/2}} \left( \frac{v' \Sigma^{-1} v}{2 + \alpha' \Sigma^{-1} \alpha} \right)^{\nu/2} K_\nu(u),$$

(2.8)

where $\nu = (2 - p)/2$, $u = \sqrt{(2 + \alpha' \Sigma^{-1} \alpha) (v' \Sigma^{-1} v)}$, $\Sigma$ is a scale matrix, and $\alpha \in \mathbb{R}^p$ represents the skewness in each dimension. Kotz et al. (2001) use the notation $V \sim \mathcal{AL}_p(\alpha, \Sigma)$ to indicate that the random variable $V$ follows a $p$-dimensional CAL distribution, and they provide extensive details.

The CAL density (2.8) is prohibitive for model-based clustering and classification applications because it would force each component density to be joined at the same origin. To
address this problem, consider $V \sim \mathcal{AL}_p(\alpha, \Sigma)$ and introduce a shift parameter $\mu \in \mathbb{R}^p$ by considering a random vector $X = (V + \mu) \sim \mathcal{SAL}_p(\alpha, \Sigma, \mu)$, where $\mathcal{SAL}_p(\alpha, \Sigma, \mu)$ denotes a $p$-dimensional shifted (non-centralized) asymmetric Laplace (SAL) distribution with density given by

$$
\xi(x | \alpha, \Sigma, \mu) = \frac{2 \exp\{(x - \mu)^t \Sigma^{-1} \alpha\}}{(2\pi)^{p/2} |\Sigma|^{1/2}} \left(\frac{\delta(x, \mu | \Sigma)}{2 + \alpha^t \Sigma^{-1} \alpha}\right)^{\nu/2} K_\nu(u), \tag{2.9}
$$

where $u = \sqrt{(2 + \alpha^t \Sigma^{-1} \alpha)} \delta(x, \mu | \Sigma)$, $\delta(x, \mu | \Sigma) = (x - \mu)^t \Sigma^{-1} (x - \mu)$, and $\nu, \alpha,$ and $\Sigma$ are defined as for (2.8). Figure 2.1 presents contour plots depicting some of the shapes that are obtainable using this distribution. For each row we vary the amount of correlation such that $\Sigma = [1, \zeta, \zeta, 1]$ and for all plots we vary the amount of skewness.

Kotz et al. (2001) note that the random variable $V \sim \mathcal{AL}_p(\alpha, \Sigma)$ can be generated through the relationship

$$
V = W \alpha + \sqrt{W} Y,
$$

where $W$ is a random variable from an exponential distribution with rate 1 and $Y \sim \mathcal{N}(0, \Sigma)$ is generated independent of $W$. Therefore, the random variable $X \sim \mathcal{SAL}_p(\alpha, \Sigma, \mu)$ can be generated through the relationship

$$
X = \mu + W \alpha + \sqrt{W} Y
$$

and so $X | W = w \sim \mathcal{N}(\mu + w \alpha, w \Sigma)$.

The distribution of $W$ conditional on the data can be computed using Bayes’ theorem, i.e.,

$$
f_W(w | X = x) = \frac{f_X(x | W = w) h(w)}{f_X(x)},
$$

16
Figure 2.1: Bivariate SAL contours with $\mu = (0, 0)$ and varying $\alpha$ and $\zeta$. 

(a) $\alpha = (-4, 4)', \zeta = 0.7$  
(b) $\alpha = (0, 5)', \zeta = 0.7$  
(c) $\alpha = (3, 2)', \zeta = 0.7$  
(d) $\alpha = (-5, 0)', \zeta = 0$  
(e) $\alpha = (0, 0)', \zeta = 0$  
(f) $\alpha = (6, 0)', \zeta = 0$  
(g) $\alpha = (-5, -7)', \zeta = 0.25$  
(h) $\alpha = (0, -2)', \zeta = 0.25$  
(i) $\alpha = (4, -8)', \zeta = 0.25$
where $X \mid W = w \sim \mathcal{N}(\mu + w\alpha, w\Sigma)$, $W \sim \text{Exp}(1)$, and $f_X(x)$ is the density of the shifted asymmetric Laplace distribution given in (2.9). It follows that

$$f_W(w \mid X = x) = \frac{w^{\nu-1}}{2} \left( \frac{2 + \alpha'\Sigma^{-1}\alpha}{\delta(x, \mu \mid \Sigma)} \right)^{\nu/2} \times \exp \left\{ -\frac{1}{2w}\delta(x, \mu \mid \Sigma) - \frac{w}{2} \left( 2 + \alpha'\Sigma^{-1}\alpha \right) \delta(x, \mu \mid \Sigma) \right\},$$

where $\nu$, $\alpha$, $\mu$, $\Sigma$, and $\delta(x, \mu \mid \Sigma)$ are as defined for (2.9). Recalling the density of a GIG random variable (2.1), it follows from (2.10) that $f_W(w \mid X = x)$ is the GIG density with $a \equiv 2 + \alpha'\Sigma^{-1}\alpha$ and $b \equiv \delta(x, \mu \mid \Sigma)$, cf. Barndorff-Nielsen (1997).

### 2.6 The Expectation Maximization Algorithm

The expectation-maximization (EM) algorithm is an iterative procedure often used when trying to estimate model parameters in the presence of missing or incomplete data. The EM algorithm belongs to a larger class of algorithms known as MM algorithms (defined in Section 2.6.1). Subsequently, the EM algorithm, some useful extensions, and a deterministic annealing approach are introduced.

#### 2.6.1 MM Algorithms

‘MM’ stands for ‘minorize-maximize’ or ‘majorize-minimize,’ depending on the purpose of the algorithm. Formally, if we let $\psi^{(m)}$ be a fixed value of a parameter $\psi$ and let $g(.)$ and
f(.) be real-value functions then $g(\psi | \psi^{(m)})$ is said to majorize $f(\psi)$ at the point $\psi^{(m)}$ if

$$g(\psi | \psi^{(m)}) \geq f(\psi) \text{ for all } \psi, \text{ and}$$

$$g(\psi^{(m)} | \psi^{(m)}) = f(\psi^{(m)}).$$

Alternatively, if $-g(\psi | \psi^{(m)})$ majorizes $-f(\psi)$ at $\psi^{(m)}$ then $g(\psi | \psi^{(m)})$ is said to minorize $f(\psi)$ at $\psi^{(m)}$.

Applications of the MM principle date back at least as far as Ortega and Rheinboldt (1970) and a thorough review of MM algorithms is given by Hunter and Lange (2004). In an EM context the minorizing function is the expected value of the complete-data log-likelihood.

### 2.6.2 The EM Algorithm

The EM algorithm was formulated in the seminal paper of Dempster et al. (1977) and due to its construction is a natural choice when preforming model-based clustering and classification. On each iteration of the EM algorithm two steps are preformed: an expectation (E-) step and a maximization (M-) step. On each E-step the expected value(s) of the complete-data log-likelihood, $Q$, are calculated, and on each M-step, the maximum likelihood estimates of the model parameters are computed. Note: Titterington et al. (1985) also cite similar approaches used by Baum et al. (1970), Orchard and Woodbury (1972), and Sundberg (1974).

In general, the term complete-data refers to the combination of the unobserved (i.e., the missing or incomplete data) and the observed data. For illustrative purposes consider an EM-algorithm designed for model-based clustering using a finite mixture of multivariate Gaussian distributions. Explicitly, the probability density function of a mixture of multivariate
Gaussian distributions is given by (1.2). Therefore, for \( x = (x_1, \ldots, x_n)' \) the complete-data likelihood is given by

\[
L(\vartheta) = \prod_{i=1}^{n} \prod_{g=1}^{G} \left[ \frac{\pi_g}{(2\pi)^{p/2}|\Sigma_g|^{1/2}} \exp \left\{ -\frac{1}{2} (x_i - \mu_g)' \Sigma_g (x_i - \mu_g) \right\} \right]^{z_{ig}},
\]

and the complete-data log-likelihood is given by

\[
Q(\vartheta) = \sum_{g=1}^{G} n_g \log \pi_g - \frac{np}{2} \log 2\pi - \sum_{g=1}^{G} \frac{n_g}{2} |\Sigma_g| - \sum_{g=1}^{G} \frac{n_g}{2} (x_i - \mu_g)' \Sigma_g^{-1} (x_i - \mu_g),
\]

where \( z_{ig} \) is an indicator variable introduced to represent whether observation \( x_i \) belongs to group \( g \) and \( n_g = \sum_{i=1}^{n} \hat{z}_{ig} \).

Formally,

\[
z_{ig} = \begin{cases} 
1 & \text{if } x_i \text{ belongs to component } g, \\
0 & \text{otherwise,}
\end{cases}
\]

for \( i = 1, \ldots, n \) and \( g = 1, \ldots, G \). In this example the \( z_{ig} \) constitutes the missing data and the complete-data is composed of it and the observations \( x_i \). On each \( E \)-step the missing \( z_{ig} \) are replaced by their expected value

\[
E[Z_{ig} \mid x_i] = \frac{\pi_g \phi_g(x \mid \mu_g, \Sigma_g)}{\sum_{h=1}^{G} \pi_h \phi_h(x \mid \mu_h, \Sigma_h)} := \hat{z}_{ig},
\]

since the random variable \( Z_i \mid x_i \) follows the multinomial distribution with a single trial such that

\[
P(Z_{ig} = 1 \mid x_i) = \frac{\pi_g \phi_g(x \mid \theta_g)}{\sum_{h=1}^{G} \pi_h \phi_h(x \mid \theta_h)}.
\]

On each \( M \)-step the expected complete-data log-likelihood is maximized with respect to the
model parameters: $\pi_g$, $\mu_g$ and $\Sigma_g$. The updates for $\pi_g$, $\mu_g$ and $\Sigma_g$ are given by $\hat{\pi}_g = n_g/n$, $\hat{\mu}_g = \frac{\sum_{i=1}^{n} \hat{z}_{ig} x_i}{\sum_{i=1}^{n} \hat{z}_{ig}}$, and $\hat{\Sigma}_g = \frac{1}{n_g} \sum_{i=1}^{n} (x_i - \mu_g) (x_i - \mu_g)'$ respectively.

On each successive iteration the EM algorithm increases the likelihood until convergence to a global maximizer is reached. Its monotonic nature is only one of the attractive features of the EM algorithm (see McLachlan and Krishnan, 2008, for complete details).

2.6.3 Extensions

Expectation-Conditional Maximization Algorithm

In situations where the maximum likelihood estimates are complicated the M-step of an EM algorithm can become computationally unattractive. However, if estimation is conducted while conditioning on some of the other model parameters, the computational ease can be regained. As such Meng and Rubin (1993) developed an expectation-conditional maximization (ECM) algorithm that replaces each M-step of the EM algorithm with a number of computationally efficient, simpler, faster and more stable conditional maximization (CM-) steps. Like a traditional M-step each CM-step will maximize the conditional expectation of the complete-date log-likelihood.
Meng and Rubin (1993) show that the ECM algorithm will converge to a stationary point under essentially the same conditions as a standard EM algorithm and preserves the appealing convergence properties however, it usually uses more iterations than a standard EM. Meng and Rubin (1993) also discuss in detail other issues related to the convergence of this algorithm and a full summary of the ECM algorithm’s properties as well as illustrative examples of this technique are given in McLachlan and Krishnan (2008).

**Alternating Expectation-Conditional Maximization Algorithm**

The alternating expectation-conditional maximization (AECM) algorithm was proposed by Meng and van Dyk (1997) as an extension of the EM algorithm although, it is really an extension of the ECM algorithm (see McLachlan and Krishnan, 2008, Section 5.13). The AECM algorithm allows for the specification of different complete-data at each stage of the algorithm. As expected the number of iterations between initialization and convergences tends to increase compared to a traditional EM algorithm however, this is can be coupled with a decrease in total computational time (McLachlan and Krishnan, 2008). Like the ECM and EM algorithms the AECM algorithm maintains the attractive properties of EM algorithm; specifically the monotone convergence of the sequence of likelihood values.

### 2.6.4 Deterministic Annealing

A common criticism of the use of the EM algorithm in model-based clustering applications is that the singularity-riddled likelihood surface makes parameter estimation unreliable and heavily dependent on the starting values (cf. Titterington et al., 1985). To help overcome this
problem Zhou and Lange (2010) developed a deterministic annealing algorithm by slightly modifying the EM algorithm presented in Section 2.6. Their annealing algorithm uses a tuning parameter to warp the likelihood surface, enhancing the chances of finding the dominant mode. They demonstrate their approach using a Gaussian mixture model with two components.

Recall the EM algorithm for Gaussian mixture model presented in Section 2.6. For the observed data \( x_1, \ldots, x_n \), we defined the indicator variables \( z_1, \ldots, z_n \), where \( z_i = (z_{i1}, \ldots, z_{iG}) \) to be equal to 1 if observation \( i \) is in group \( g \) and equal to 0 otherwise. On the \( E \)-step of the EM algorithm we replace each \( z_{ig} \) with its expected value

\[
\mathbb{E}[Z_{ig} \mid x_i] = \frac{\pi_g \phi_g(x \mid \mu_g, \Sigma_g)}{\sum_{h=1}^G \pi_h \phi_h(x \mid \mu_h, \Sigma_h)} := \hat{z}_{ig}.
\]

In the deterministic annealing algorithm the \( E \)-step is slightly modified such that

\[
\mathbb{E}[Z_{ig} \mid x_i] = \frac{[\pi_g \phi_g(x \mid \mu_g, \Sigma_g)]^v}{\sum_{h=1}^G [\pi_h \phi_h(x \mid \mu_h, \Sigma_h)]^v} := \hat{z}^*_{ig},
\]

where \( v \in [0, 1] \) is the tuning parameter introduced to control the transformation of the likelihood surface. Note that \( v \) comes from a sequence of user-specified length and as the deterministic annealing algorithm progresses, \( v \) increases from 0 to 1.
2.7 Convergence Criterion

2.7.1 Aitken acceleration

A popular approach for determining whether or not an EM algorithm has converged is to use Aitken’s acceleration (Aitken, 1926). Aitken’s acceleration uses a measurement of the asymptotic maximum of the log-likelihood at each iteration to make a decision about whether the algorithm has converged or not.

At iteration \(k\) the aitken’s acceleration has value

\[
a^{(k)} = \frac{l^{(k+1)} - l^{(k)}}{l^{(k)} - l^{(k-1)}}
\]

where \(l^{(k+1)}, l^{(k)}\) and \(l^{(k-1)}\) are the log-likelihood values from iterations \(k + 1, k,\) and \(k - 1\) respectively. The asymptotic estimate (Böhning et al., 1994) of the log-likelihood at iteration \(k + 1\) is given by

\[
l^{(t+1)} = l^{(t)} + \frac{l^{(t+1)} - l^{(t)}}{1 - a^{(t)}},
\]

(2.13)

where each value is previously defined.

Lindsay (1995) suggest that an EM algorithm can be considered to have converged when

\[
l^{(k+1)} - l^{(k+1)} < \epsilon,
\]

(2.14)

where \(\epsilon\) is some small number. McNicholas et al. (2010) propose an alternative stopping
criteria, stating that an EM algorithm can be considered to have converged when

\[ l_{\infty}^{(k+1)} - l^{(k)} < \epsilon, \quad (2.15) \]

for some small \( \epsilon \). McNicholas et al. (2010) show that the criteria in (2.15) is just as strict as (2.14) since \( l^{(k+1)} \geq l^{(k)} \). Furthermore, McNicholas et al. (2010) show that the criteria in (2.15) is also at least as strict as the lack of progress criterion

\[ l^{(k+1)} - l^{(k)} < \epsilon, \quad (2.16) \]

used in Fraley and Raftery (1998).

### 2.8 Model Selection and Performance

When fitting finite mixture models to data one of the main objectives is to choose the number of components and covariance structure (if applicable). There are several well-known model selection criteria for example, the Akaike information criteria (AIC; Akaike, 1974), the deviance information criterion (Zhu and Carlin, 2000), consistent AIC (CAIC), CAICF\(_{Corrected}\) (Bozdogan, 1987), and the more recent information complexity criterion (ICOMP; Bozdogan, 2000). The most popular model selection criterion is the Bayesian Information Criterion (BIC; Schwarz, 1978) whose use is justified by a number of authors (see Kass and Raftery, 1995; Keribin, 2000, for example). We consider both the BIC as well as the integrated complete Likelihood measure herein.
2.8.1 Bayesian Information Criterion

The BIC is derived using a Laplace approximation and its precision is influenced by the specific form of the prior density of the parameters and the correlation structure between observations. The BIC is given by

$$
\text{BIC} = 2l(x | \hat{\theta}) - \rho \log n,
$$

(2.17)

where $l(x | \hat{\theta})$ is the maximized log-likelihood, $\hat{\theta}$ is the maximum likelihood estimate of $\theta$, $\rho$ is the number of free parameters in the model, and $n$ is the number of observations. The BIC is commonly used for Gaussian mixture model selection and has some useful asymptotic properties, for example, as $n \to \infty$ the BIC is shown to consistently chooses the correct model (see Dasgupta and Raftery, 1998; Leroux et al., 1992, for example).

As the dimension of the data increases the penalty term in the BIC $\rho \log n$ dominates the likelihood, this causes the BIC to choose models that under fit the data. To rectify these issues Fraley and Raftery (2007) proposed Bayesian regularizations for Gaussian mixtures and Pan and Shen (2007) introduced a penalty to the likelihood basesd on the infamous LASSO (Tibshirani, 1996). Unfortunately the penalty introduced in Pan and Shen (2007) only works for a common diagonal component covariance matrix. To this end Bhattacharya and McNicholas (2013) propose a LASSO penalized BIC (LPBIC) and demonstrate their criterion using the PGMMs (see Section 2.4.2).
2.8.2 Integrated Completed Likelihood

The integrated completed likelihood (ICL; Biernacki et al., 2000) is given by

\[
\text{ICL} \approx \text{BIC} + \sum_{i=1}^{n} \sum_{g=1}^{G} \text{MAP}\{\hat{z}_{ig}\} \log \hat{z}_{ig},
\]  

(2.18)

where \(\sum_{i=1}^{n} \sum_{g=1}^{G} \text{MAP}\{\hat{z}_{ig}\} \log \hat{z}_{ig}\), the estimated mean entropy, reflects the uncertainty in the classification of observations into components. Essentially, the ICL penalizes the BIC for uncertainty in classification, for this reason it is considered to be more appropriate for model-based clustering and classification.

2.9 Performance Assessment

2.9.1 Adjusted Rand Index

The Rand index (Rand, 1971) was introduced to compare partitions and is given by the pair agreements:

\[
\frac{\text{number of agreements}}{\text{number of agreements} + \text{number of disagreements}}.
\]

The Rand index takes a value between 0 and 1, where 1 indicates perfect agreement. An unattractive feature of the Rand index is that it has a positive expected value under random classification.

To correct this undesirable property, Hubert and Arabie (1985) introduced the ARI to account for chance agreement. The ARI also takes a value of 1 when classification agreement...
is perfect but has an expected value of 0 under random classification. The ARI can also take negative values and this happens for classifications that are worse than would be expected by chance.

Steinley (2004) provides a set of general properties of the ARI. Specifically, the author shows that the mean and standard deviation of the ARI remain unaffected given changes to the number of clusters, the total number of observations and the percent of observations in each cluster. In addition, the author re-iterates a point made in Steinley (2003); that the ARI is a better measure of classification performance when compared to simply computing the misclassification rate.

2.10 Model-Based Classification

Suppose we have \( n \) observations and \( k \) of these observations have known group memberships. We can use the group memberships of these \( k \) observations to estimate memberships for the remaining \( n - k \) observations within a joint likelihood framework. This approach, known as model-based classification, is a semi-supervised version of model-based clustering. Without loss of generality, we order the observations \( x_1, \ldots, x_k, x_{k+1}, \ldots, x_n \) so that the first \( k \) have known group memberships. Therefore, the values of \( z_{ig} \) are known for \( i = 1, \ldots, k \) and the general model-based classification likelihood is given by

\[
\mathcal{L} (x_1, \ldots, x_n, \tau_1, \ldots, \tau_n \mid \vartheta) = \prod_{i=1}^{k} \prod_{g=1}^{G} [\pi_g f_g(x \mid \theta_g)]^{z_{ig}} \prod_{j=k+1}^{n} \sum_{h=1}^{H} \pi_h f_h(x \mid \theta_h), \tag{2.19}
\]

where \( H \geq G \). Parameter estimation is carried out in an analogous fashion to model-based clustering.
Chapter 3

Mixtures of Shifted Asymmetric Laplace Distributions

3.1 Introduction

This chapter details the methodological development of a novel mixture model designed for clustering skewed data and is split into 3 sections. Section 3.2 outlines parameter estimation via an expectation-maximization algorithm with deterministic annealing, in Section 3.3 both simulated and real data analyses are used to illustrate our approach and in Section 3.4 the chapter concludes with a discussion of the work.

As mentioned in the Chapter 1 the majority of work on using mixture models for clustering and classification focus on the Gaussian mixture model (1.2). We introduce a non-Gaussian approach that allows for skewness while also parameterizing location and scale. This ap-
proach is effective while also being mathematically elegant and relatively computationally straightforward. To this end a finite mixture of SAL distributions so that the gth component density is \( S\mathcal{AL}_p(\alpha_g, \Sigma_g, \mu_g) \), where the parameters are as defined for (2.9), is introduced. The density of a mixture of SAL distributions is

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

(3.1)

where \( \vartheta \) is the vector of all model parameters and

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

is the multivariate SAL density with \( \nu = (2 - p)/2 \), \( \alpha_g \) is a skewness parameter, \( \mu_g \) is a location parameter, \( \Sigma_g \) is a scale matrix and \( K_{\nu}(u) \) is the modified Bessel function of the third kind with \( u = \sqrt{(2 + \alpha_g' \Sigma_g^{-1} \alpha_g)} \delta(x, \mu_g | \Sigma_g) \) and \( \delta(x, \mu_g | \Sigma_g) \) is the squared Mahalanobis distance between \( x \) and \( \mu_g \).

### 3.2 Parameter Estimation

#### 3.2.1 EM Algorithm

To fit our SAL mixture model (cf. Section 3.2.3) we use a version of the deterministic annealing algorithm (cf. Section 2.6.4) in conjunction with an EM algorithm. Note that Eltoft et al. (2006a) give examples of an EM-type algorithm for fitting a multivariate Laplace distribution, and there are several examples of applications of Laplacian mixture models in a variety
of applied settings (e.g., Cord et al., 2006; Mitianoudis and Stathaki, 2005; Bhowmick et al., 2006; Shi and Selesnick, 2006). Bolin (2011) and Scallan (1992) discuss fitting these mixtures more generally.

### 3.2.2 Application to Mixture of SAL Distributions

For our SAL mixture models, the complete-data comprise the observed $x_1, \ldots, x_n$, the missing component membership labels $z_1, \ldots, z_n$, and the latent $w_{ig}$. For each $i$, we have $z_i = (z_{i1}, \ldots, z_{iG})$, where $z_{ig} = 1$ if observation $i$ is in component $g$ and $z_{ig} = 0$ otherwise, for $i = 1, \ldots, n$ and $g = 1, \ldots, G$.

Now, $X_i \mid w_{ig}, z_{ig} = 1 \sim N(\mu_g + w_{ig}\alpha_g, w_{ig}\Sigma_g)$ independently for $i = 1, \ldots, n$, $W_{ig} \mid z_{ig} = 1 \sim \text{Exp}(1)$, and $W_{ig} \mid x_i, z_{ig} = 1 \sim \text{GIG}(a_g, b_{ig})$ with $a_g := 2 + \alpha^T g \Sigma^{-1} \alpha g$ and $b_{ig} := \delta(x_i, \mu_g \mid \Sigma_g)$. Note that $Z_i \mid x_i$ follows the multinomial distribution with a single trial and

$$P(Z_{ig} = 1 \mid x_i) = \frac{\pi_g \xi(x_i \mid \alpha_g, \Sigma_g, \mu_g)}{\sum_{j=1}^{G} \pi_j \xi(x_i \mid \alpha_j, \Sigma_j, \mu_j)}.$$

Therefore, the complete-data likelihood is given by

$$L_c = \prod_{i=1}^{n} \prod_{g=1}^{G} \left[ \pi_g \phi(x_i \mid \mu_g + w_{ig}\alpha_g, w_{ig}\Sigma_g) h(w_{ig}) \right]^{z_{ig}},$$

with the same notation used previously and where $\phi(x_i \mid \mu_g + w_{ig}\alpha_g, w_{ig}\Sigma_g)$ is the density of a multivariate Gaussian distribution with mean $\mu_g + w_{ig}\alpha_g$ and covariance matrix $w_{ig}\Sigma_g$, and $h(w_{ig})$ is the density of an exponential distribution with rate 1, i.e., $h(z) = e^{-1}$, for real
The expected-value of the complete-data log-likelihood is given by

\[
Q = \sum_{g=1}^{G} n_g \log \pi_g - \frac{np}{2} \log 2\pi + \sum_{g=1}^{G} \frac{n_g}{2} \log |\Sigma_g^{-1}| - \frac{p}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} E_{3ig} \\
- \frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} E_{2ig} (x_i - \mu_g)' \Sigma_g^{-1} (x_i - \mu_g) + \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} (x_i - \mu_g)' \Sigma_g^{-1} \alpha_g \\
- \frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} E_{1ig} \alpha_g \Sigma_g^{-1} \alpha_g - \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} E_{1ig},
\]

(3.3)

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

\[
\mathbb{E} [Z_{ig} | x_i] = P (Z_{ig} = 1 | x_i) := \hat{z}_{ig},
\]

\[
\mathbb{E} [W_{ig} | x_i, z_{ig} = 1] = \sqrt{\frac{b_{ig}}{a_g} R_{\nu} \left( \sqrt{a_g b_{ig}} \right) } := E_{1ig},
\]

\[
\mathbb{E} [W_{ig}^{-1} | x_i, z_{ig} = 1] = \sqrt{\frac{a_g}{b_{ig}} R_{\nu} \left( \sqrt{a_g b_{ig}} \right)} - \frac{2\nu}{b_{ig}} := E_{2ig},
\]

and \( \nu, a_g, \) and \( b_{ig} \) are as previously defined. These attractive closed forms for \( E_{1ig} \) and \( E_{2ig} \) exist because \( W_{ig} | x_i, z_{ig} = 1 \sim GIG(a_g, b_{ig}) \) and so we can use the formulae in (2.2). Note that the expected value

\[
\mathbb{E} [\log W_{ig} | x_i, z_{ig} = 1] := E_{3ig}
\]

also has a nice closed form; however, it is not required in our calculations because the term containing \( E_{3ig} \) in \( Q \) is constant with respect to the model parameters, i.e., \( \pi_g, \alpha_g, \mu_g, \) and \( \Sigma_g \) \((g = 1, \ldots, G)\).

We maximize \( Q \) with respect to the model parameters to get the maximum likelihood estimates (details are given in Appendix A). Specifically, the mixing proportions, skewness
parameter, and shift parameter are updated by \( \hat{\pi}_g = n_g / n \),

\[
\hat{\alpha}_g = \frac{\left( \sum_{i=1}^{n} \hat{z}_{ig} E_{2ig} \right) \left( \sum_{j=1}^{n} \hat{z}_{jg} x_j \right) - n_g \sum_{i=1}^{n} \hat{z}_{ig} E_{2ig} x_i}{\left( \sum_{i=1}^{n} \hat{z}_{ig} E_{1ig} \right) \left( \sum_{j=1}^{n} \hat{z}_{jg} E_{2jg} \right) - n_g^2},
\]

and

\[
\hat{\mu}_g = \frac{\left( \sum_{i=1}^{n} \hat{z}_{ig} E_{1ig} \right) \left( \sum_{j=1}^{n} \hat{z}_{jg} E_{2jg} x_j \right) - n_g \sum_{i=1}^{n} \hat{z}_{ig} x_i}{\left( \sum_{i=1}^{n} \hat{z}_{ig} E_{1ig} \right) \left( \sum_{j=1}^{n} \hat{z}_{jg} E_{2jg} \right) - n_g^2},
\]

respectively. Each component scale matrix \( \hat{\Sigma}_g \) is updated by

\[
\hat{\Sigma}_g = S_g - \hat{\alpha}_g r_g' - r_g \hat{\alpha}_g' + \frac{1}{n_g} \hat{\alpha}_g \hat{\alpha}_g' \sum_{i=1}^{n} \hat{z}_{ig} E_{1ig}, \tag{3.4}
\]

where

\[
S_g = \frac{1}{n_g} \sum_{i=1}^{n} \hat{z}_{ig} E_{2ig} \left( x_i - \hat{\mu}_g \right) \left( x_i - \hat{\mu}_g \right)',
\]

and

\[
r_g := \frac{1}{n_g} \sum_{i=1}^{n} \hat{z}_{ig} \left( x_i - \hat{\mu}_g \right) .
\]

Now, it is possible to summarize our EM algorithm for the mixture of shifted asymmetric Laplace distributions:

initialize \( \hat{\pi}_g, \hat{\alpha}_g, \hat{\mu}_g, \) and \( \hat{\Sigma}_g \) using deterministic annealing

while not converged:

E-step

\[
\begin{align*}
\text{update } \hat{z}_{ig} \\
\text{update } \mathbb{E}[W_{ig} \mid x_i, z_{ig} = 1] \\
\text{update } \mathbb{E}[W_{ig}^{-1} \mid x_i, z_{ig} = 1]
\end{align*}
\]
M-step
\[
\begin{align*}
&\text{update } \hat{\pi}_g \\
&\text{update } \hat{\alpha}_g \\
&\text{update } \hat{\mu}_g \\
&\text{update } \hat{\Sigma}_g \\
\end{align*}
\]
compute log-likelihood
check for convergence
end while

At convergence, the \( \hat{z}_{ig} \) are the \emph{a posteriori} probabilities of component membership for each observation and can be used to cluster the observations into groups. Predicted classifications are typically obtained via maximum \emph{a posteriori} (MAP) probabilities, where MAP\{\( \hat{z}_{ig} \)\} = 1 if \( \max_g \{\hat{z}_{ig}\} \) occurs in component \( g \) and MAP\{\( \hat{z}_{ig} \)\} = 0 otherwise. To assess clustering and classification performance, we use a cross-tabulation of our MAP classifications against the true group memberships. We then compute the ARI (Hubert and Arabie, 1985).

### 3.2.3 Initialization

The deterministic annealing algorithm (Zhou and Lange, 2010) is the same as the EM algorithm described in Sections 3.2.1 and 3.2.2, except that now

\[
E[Z_{ig} \mid x_i] = \frac{[\pi_g \xi (x_i \mid \alpha_g, \Sigma_g, \mu_g)]^v}{\sum_{h=1}^G [\pi_h \xi (x_i \mid \alpha_h, \Sigma_h, \mu_h)]^v} \tag{3.5}
\]

in each E-step. Here, the auxiliary parameter \( v \in (0, 1) \), which is drawn from an increasing sequence of user-specified length, transforms the likelihood surface to improve the chances
of finding the dominant mode. The user-specified sequence runs from 0 to 1 and its length determines how many iterations of the deterministic annealing algorithm will be preformed. The annealing algorithm itself is initialized using random starting values for \(\hat{\pi}_g, \hat{\alpha}_g, \hat{\Sigma}_g,\) and \(\hat{\mu}_g\). In our analyses (Section 3.3), we use deterministic annealing to obtain starting values for our EM algorithms.

### 3.2.4 Dealing with Infinite Log-Likelihood Values

As our EM algorithm iterates, we must handle the complications that arise when updating \(\hat{\mu}_g\). Specifically, as the algorithm iterates towards convergence, the value of \(\hat{\mu}_g\) can tend to an observation \(x_i\). Although such estimates of \(\hat{\mu}_g\) maximize the likelihood, they create computational issues when trying to determine the remaining parameter values and, specifically, the expected value \(E_{2ig}\). To overcome this problem, we stop updating \(\hat{\mu}_g\) when its value equals some \(x_i\). We proceed by taking the value of \(\hat{\mu}_g\) at the iteration before it becomes equal to any \(x_i\) (we denote this value \(\hat{\mu}_g^*\)) as the estimate for \(\mu_g\). Then, we update \(\hat{\alpha}_g\) using

\[
\hat{\alpha}_g^* = \left[ \frac{\sum_{i=1}^n \hat{z}_{ig} \left( x_i - \hat{\mu}_g^* \right)' \hat{\Sigma}_{1ig}^{-1}}{\sum_{i=1}^n \hat{z}_{ig} E_{1ig}} \right]'
\]

and update \(\hat{\Sigma}_g\) as in (3.4). We use a different approach to overcome this problem in the deterministic annealing algorithm, simply restricting \(E_{2ig}\) from exceeding a value of \(-\log(1-v)\) at each iteration. We acknowledge that our solution to this problem is a simple-minded one; however, we have found it to be quite effective. A more thorough exploration of this problem in general is the subject of ongoing work.
3.2.5 Model-Based Classification

The SAL model-based classification likelihood is given by

\[
\mathcal{L} (x_1, \ldots, x_n, \tau_1, \ldots, \tau_n \mid \vartheta) = \prod_{i=1}^{k} \prod_{g=1}^{G} \left[ \pi_g \xi (x_i \mid \alpha_g, \Sigma_g, \mu_g) \right]^{z_{ig}} \times \prod_{j=k+1}^{n} \sum_{h=1}^{H} \pi_h \xi (x_j \mid \alpha_h, \Sigma_h, \mu_h),
\]

(3.6)

where \( H \geq G \). Parameter estimation is carried out in an analogous fashion to model-based clustering.

In the analyses herein, we prefer to use the integrated classification likelihood (ICL; Biernacki et al. (2000)) for model selection as it is specifically designed for clustering and classification applications.

3.3 Data Analyses

3.3.1 Introduction

Our SAL mixture models are applied to simulated data (Section 3.3.2) and to two real data sets: the famous Old Faithful geyser data (Section 3.3.3) and data on cellular localization sites for proteins in yeast (Section 3.3.4). The simulation study is conducted to compare our SAL mixtures with the Gaussian mixture model approach on data generated from SAL mixtures. In the real analyses, we illustrate the SAL approach, judging its performance alongside that of Gaussian mixtures.
On the face of it, a comparison of our SAL approach to Gaussian mixtures might be considered a little empty because one would expect Gaussian mixtures to use more than one component to model a cluster with skewness. Indeed, there has been a lot of work within the literature on merging Gaussian components (e.g., Baudry et al., 2010; Hennig, 2010). However, our examples illustrate that when predicted classifications differ for the SAL and Gaussian approaches, merging Gaussian components cannot always be used to rectify shortcomings in the Gaussian results (cf. Sections 3.3.2 and 3.3.4). One may argue that the mixture of multivariate skew-$t$ distributions could also be used to accommodate skewness. This is true. However, different representations of the mixture of skew-$t$ distributions have appeared within the literature (e.g., Sahu et al., 2003; Pyne et al., 2009) and they do not have the elegance or relative computational simplicity of our SAL mixtures. This is most apparent in the intractability of the E-step for the skew-$t$ approach (cf. Lin, 2010; Lee and McLachlan, 2011; Vrbik and McNicholas, 2012).

Note that, for all analyses, we use the same deterministic annealing starting values for the mixtures of Gaussian distributions as for our SAL mixtures. Note also that we treat each analysis as a genuine clustering problem by removing all labels before our analyses, or as a genuine classification problem by removing relevant labels.

### 3.3.2 Simulation Study

We use the relationship between the SAL and Gaussian distributions (cf. Section 2.5) to generate multivariate SAL data. Specifically, we simulate 25 data sets for $n = 500$ with $p = 2$ dimensions and $G = 2$ components (e.g., Figure 3.1). The data are generated using skewness parameters $\alpha_1 = (2, 1)'$ and $\alpha_2 = (2, 2)'$, shift parameters $\mu_1 = (0, -2)'$ and
Figure 3.1: Example of a simulated data set \((n = 500)\), coloured by true component.

\[
\mathbf{\mu}_2 = (0, 5)', \text{ and scale matrices} \\
\Sigma_1 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \quad \text{and} \quad \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.
\]

SAL and Gaussian mixtures, with \(G = 1, \ldots, 5\) components, are fitted to the data. The clustering results (Table 3.1) show that the SAL mixtures give almost perfect clustering performance over all 25 runs (average ARI = 0.9968). The Gaussian approach, however, gives relatively poor clustering performance (average ARI = 0.4988) and never returns the
correct number of clusters. The most common Gaussian solution (chosen 21 times) has five components, e.g., Figure 3.2, and each of the other four solutions has four components.

Table 3.1: Proportion of times $G = 2$ components were selected as well as average ARI values (with std. dev.) for the SAL and Gaussian mixtures, respectively, applied to the simulated data.

<table>
<thead>
<tr>
<th></th>
<th>$G = 2$ selected</th>
<th>Average ARI (std. dev.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAL</td>
<td>100%</td>
<td>0.9968 (0.00516)</td>
</tr>
<tr>
<td>Gaussian</td>
<td>0%</td>
<td>0.4988 (0.06052)</td>
</tr>
</tbody>
</table>

Figure 3.2: One of the $G = 5$ component Gaussian solutions for the simulated data, coloured by MAP classification results.
The five-component Gaussian solution depicted in Figure 3.2 is typical of 21 of the cases observed in this simulation. Particularly striking here is that the components cannot be combined to give the correct component memberships. Instead, there is a clear ‘noise’ group comprising points that do not fit neatly within any of the other four Gaussian components. Although one might argue that this problem is obvious by inspection in two dimensions, and might be resolved in some post hoc fashion, it would be virtually impossible to detect or resolve in all but very low dimensional cases.

### 3.3.3 Old Faithful Geyser Data

The famous Old Faithful geyser data, which are available as `faithful` in R (R Core Team, 2013), comprise a two-variable data set measuring the waiting time between and duration of 272 eruptions of the Old Faithful geyser in Yellowstone National Park. These data are well-known as an example of skewness and they have been used many times to illustrate approaches to analyzing skewed data (e.g., Ali et al., 2010).

Mixture of SAL distributions and mixture of Gaussian distributions models are fitted to the geyser data. There are no ‘true’ classifications for these data but both approaches select sensible groups with identical classifications. The associated contour plots (Figure 3.3 and Figure 3.4) illustrate that the fit to the data is better for the mixture of SAL distributions. Note that applying a mixture of skew-$t$ distributions to these data will also give nicely fitting contours (cf. Vrbik and McNicholas, 2012). However, as mentioned in Section 3.3.1, our mixture of SAL distributions is less computationally cumbersome.
Figure 3.3: Model-based clustering results with contours for the SAL and Gaussian mixture models (GMMs) on the Old Faithful data.
Figure 3.4: 3D contours for the SAL and Gaussian mixture models (GMMs) on the Old Faithful data.

3.3.4 Yeast Data

The Data

The yeast data, which are available from the UCI machine learning repository, contain cellular localization sites of 1,484 proteins. The development of these data, as well as classification results using a ‘rule-based expert system’, are discussed by Nakai and Kanehisa (1991) and Nakai and Kanehisa (1992). For illustration, we consider three variables: McGeoch’s method for signal sequence recognition (mcg), the score of the ALOM membrane spanning region prediction program (alm), and the score of discriminant analysis of the amino acid content of vacuolar and extracellular proteins (vac). We want to see whether our cluster
analysis can distinguish between the two localization sites, CYT (cytosolic or cytoskeletal) and ME3 (membrane protein, no N-terminal signal), cf. Figure 3.5.

Figure 3.5: The yeast data with the CYT and ME3 location sites highlighted.
Model-Based Clustering Results

SAL and Gaussian mixture models are fitted to these data for $G = 1, \ldots, 5$ components and the best fitting model is chosen using the ICL. The chosen SAL mixture model has two components ($ICL = -5173.133$) and the chosen Gaussian mixture model has three components ($ICL = -5161.878$). The classification performance of the SAL mixture model (ARI = 0.81) is superior to that of the Gaussian mixture model (ARI = 0.56), as illustrated in Table 3.2 and Figure 3.6. Again, this superiority cannot be negated by merging Gaussian components.

Table 3.2: Clustering results for the chosen SAL and Gaussian mixture models for the yeast data. The localization sites are cross-tabulated against our predicted classifications (A, B, C) in each case.

<table>
<thead>
<tr>
<th></th>
<th>SAL</th>
<th>GMM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>CYT</td>
<td>448</td>
<td>15</td>
</tr>
<tr>
<td>ME3</td>
<td>14</td>
<td>149</td>
</tr>
</tbody>
</table>

Now, one may argue that the Gaussian mixture model would perform better under a different model selection criterion. Therefore, we also investigated the Gaussian mixture model with $G = 2$ components (Table 3.3). Surprisingly, this Gaussian mixture model gives very poor clustering performance, producing classifications that are a little worse than would be expected by guessing (i.e., $ARI = -0.088 < 0$).

Table 3.3: Clustering results for the two-component Gaussian mixture model for the yeast data. The localization sites are cross-tabulated against our predicted classifications (A, B).

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYT</td>
<td>106</td>
<td>357</td>
</tr>
<tr>
<td>ME3</td>
<td>1</td>
<td>162</td>
</tr>
</tbody>
</table>
Model-Based Classification Results

To compare model-based classification within the SAL and Gaussian mixture modelling frameworks, we analyze the yeast data with 70% of the group memberships taken to be

Figure 3.6: Clustering results for the SAL (left) and Gaussian (right) mixture models on the yeast data; the colours highlight the predicted component memberships.
known. We set $G = H = 2$ and each model was fitted using 25 different random 70/30 partitions of the data. The aggregate classification results (Table 3.4) and ARIs (0.86 and $-0.080$, respectively) indicate that the SAL mixture models outperform their Gaussian counterparts by some margin.

Table 3.4: Aggregate classification results for the SAL and Gaussian mixture models for the yeast data with 70% of the labels taken as known. In each case, the localization sites are cross-tabulated against our predicted classifications (A, B) for the observations with unknown labels.

<table>
<thead>
<tr>
<th></th>
<th>SAL</th>
<th></th>
<th>GMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYT</td>
<td>3403</td>
<td>71</td>
<td>964</td>
</tr>
<tr>
<td></td>
<td>2502</td>
<td></td>
<td>2502</td>
</tr>
<tr>
<td>ME3</td>
<td>87</td>
<td>1139</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>1205</td>
<td></td>
<td>1205</td>
</tr>
</tbody>
</table>

The poor performance of the Gaussian mixture modelling approach here is surprising because the parameter estimates are computed with 70% of the location sites taken as known. This result raises questions around the efficacy of the semi-supervised Gaussian model-based classification approach, as well as further reinforcing the need for more flexible non-Gaussian approaches.

### 3.4 Summary and Discussion

A mixture of SAL distributions model was introduced and applied for both clustering and classification. An EM algorithm was used for parameter estimation, with starting values obtained using the deterministic annealing approach of Zhou and Lange (2010). To account for both parsimony and entropy, the ICL is used to select the number of mixture components. Our model-based clustering approach was illustrated on both real and simulated data.
In the simulation, data were generated from a SAL mixture model with two components that were very well separated. The SAL mixtures gave near-perfect results on these data whereas the Gaussian mixture models consistently overestimated the number of components. Furthermore, it is particularly noteworthy that merging Gaussian components did not capture the true structure of the data and still resulted in inferior performance when compared to SAL mixtures.

We also analyzed two real data sets. For the Old Faithful data, we considered model-based clustering with the SAL and Gaussian mixture models and both gave the same predicted group memberships. However, a contour plot revealed that the SAL mixture model captured the shape of the data far more effectively than its Gaussian counterpart. The yeast protein location data presented a much more difficult clustering problem, and so we also used these data to illustrate model-based classification. For model-based clustering, the chosen SAL model gave very good clustering performance ($G = 2$, $\text{ARI} = 0.81$) and outperformed its Gaussian counterpart ($G = 3$, $\text{ARI} = 0.56$). Furthermore, when we forced $G = 2$ components, the Gaussian mixture modelling approach gave worse clustering performance than would be expected by guessing ($\text{ARI} < 0$). Again, as with the simulation study, the poor clustering performance of Gaussian mixtures on the yeast data could not be mitigated by merging components. This reinforces a point made by McNicholas and Browne (2013), i.e., that merging Gaussian components is not a ‘get out of jail free card’.

In the model-based classification applications using the yeast protein location data, we set $G = H = 2$ and considered 25 random subsets of the data for which 70% of the locations were taken as known. The SAL mixtures again gave excellent performance ($\text{ARI} = 0.86$) but the Gaussian mixtures had negative ARIs; it is surprising that the Gaussian approach gave very poor classification performance in the semi-supervised case when 70% of the locations
were taken as known. This result calls into question the efficacy of the Gaussian model-based classification approach. One should note that a Gaussian model-based discriminant analysis (cf. Hastie and Tibshirani, 1996) might give better classification performance on these data because multiple components could be used for each known class.
Chapter 4

Parsimonious Variants

4.1 Introduction

In this chapter we introduce the first family of shifted asymmetric Laplace distributions. This development of this family is motivated by the work of Celeux and Govaert (1995); Andrews and McNicholas (2012, 2013) and Vrbik and McNicholas (2014). A parameter estimation procedure is outlined following an expectation-maximization framework (Section 4.3). Initialization of the models is described for each application; typically we choose to follow the default settings of the available model-based clustering software. Our family of models are demonstrated on both simulated and real data (Section 4.4).
4.2 The ParSAL Family of Models

As discussed in Section 2.4.1 writing the component covariance matrix in terms of its eigenvalue decomposition gives

\[ \Sigma_g = \lambda_g D_g A_g D'_g, \]  

(4.1)

where \( \lambda_g = |\Sigma_g|^{1/p} \) are the associated constants of proportionality, \( D_g \) is an orthogonal matrix of eigenvectors, and \( A_g \) is a diagonal matrix with entries proportional to the eigenvalues, such that \( |A_g| = 1 \). Following Celeux and Govaert (1995) we apply a combination of the constraints \( D_g = D, D = I_d, A_g = A, A_g = I_d \) to the eigen-decomposed component scale matrices of a SAL mixture model. This creates a family of fourteen parsimonious SAL mixtures, herein referred to as ParSAL, whose density is given by

\[ f(x | \vartheta) = \sum_{g=1}^{G} \pi_g \xi(x | \alpha_g, \lambda_g D_g A_g D'_g, \mu_g). \]  

(4.2)

Note: The nomenclature and scale matrix structure of the ParSAL family are analogous to the Gaussian parsimonious clustering models given in Table 2.1.

For the mixtures of Gaussian distributions the parameter \( \lambda_g \) determines the volume, \( D_g \) the orientation and \( A_g \) the shape of the \( g^{th} \) component. Unfortunately, this geometric interpretation of the component shapes does not follow for our parsimonious SAL mixtures, unless skewness is equal to zero (cf. Wang et al., 2009). However, because the constraints are being imposed on the component scale matrices only to introduce parsimony this lack of geometric interpretability does not depreciate their value.
4.3 Parameter Estimation

4.3.1 EM Algorithm

To fit our parsimonious models we use an EM algorithm. The details regarding our parameter estimation procedure are very similar to those described in Section 3.2.2 with two exceptions. First, after we compute $\hat{\Sigma}_g$ we need to estimate the required elements of its eigen-decomposition. Second, an alternative method for dealing with infinite log-likelihood values is suggested.

4.3.2 Application to the ParSAL Family

For the ParSAL family the complete-data are composed of the observed data, $x_1, \ldots, x_n$, the latent value $w_{ig}$ and the component membership labels, $z_1, \ldots, z_n$. For each $i$, we have $z_i = z_{i1}, \ldots, z_{iG}$ where $z_{ig} = 1$ if observation $i$ is in group $g$ and $z_{ig} = 0$ otherwise (cf. Section 3.2.1). For the most general ParSAL family member (i.e., the VVV model) the completed-data likelihood is given by

$$L_c = \prod_{i=1}^{n} \prod_{g=1}^{G} \left[ \pi_g \phi \left( x_i \mid \mu_g + w_{ig} \alpha_g, w_{ig} \lambda_g D_g A_g D'_g \right) h \left( w_{ig} \right)^{z_{ig}} \right], \quad (4.3)$$

where $\phi \left( x_i \mid \mu_g + w_{ig} \alpha_g, w_{ig} \lambda_g D_g A_g D'_g \right)$ is the density of a multivariate Gaussian distribution with mean $\mu_g + w_{ig} \alpha_g$ and covariance matrix $w_{ig} \lambda_g D_g A_g D'_g$, and $h(w_{ig})$ is the density of a random exponential variable with rate 1.
It follows that the expected-value of the complete-data log-likelihood is given by

\[ Q = \sum_{g=1}^{G} n_g \log \pi_g - \frac{np}{2} \log 2\pi + \sum_{g=1}^{G} \frac{n_g}{2} \log \left| \left( \lambda_g D_g A_g D_g' \right)^{-1} \right| - \frac{p}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} E_{3ig} \]

\[ + \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} (x_i - \mu_g)' \left( \lambda_g D_g A_g D_g' \right)^{-1} \alpha_g - \frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} E_{1ig} \alpha_g \left( \lambda_g D_g A_g D_g' \right)^{-1} \alpha_g \]

\[ - \frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} E_{2ig} (x_i - \mu_g)' \left( \lambda_g D_g A_g D_g' \right)^{-1} (x_i - \mu_g) - \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} E_{1ig}, \]

where \( n_g = \sum_{i=1}^{n} \hat{z}_{ig}, \) \( E \left[ W_{ig} \mid x_i, z_{ig} = 1 \right] := E_{1ig}, \) \( E \left[ W_{ig}^{-1} \mid x_i, z_{ig} = 1 \right] := E_{2ig} \) and \( E \left[ \log W_{ig} \mid x_i, z_{ig} = 1 \right] := E_{3ig}. \) Therefore, the E-step of each iteration will consist of replacing the component membership labels with their conditional expected values

\[ \mathbb{E} \left[ \tau_{ig} \mid x_i \right] = \frac{\pi_g \xi (x_i \mid \alpha_g, \lambda_g D_g A_g D_g', \mu_g)}{\sum_{h=1}^{G} \pi_h \xi (x_i \mid \alpha_h, \lambda_h D_h A_h D_h', \mu_h)} := \hat{z}_{ig} \]

and replacing the necessary missing values, \( w_{ig} \) and \( w_{ig}^{-1}, \) with their expected values:

\[ \mathbb{E} \left[ W_{ig} \mid x_i, z_{ig} = 1 \right] = \sqrt{\frac{b_{ig}}{a_g} R_{\nu} \left( \sqrt{a_g b_{ig}} \right)} := E_{1ig}, \]

and

\[ \mathbb{E} \left[ W_{ig}^{-1} \mid x_i, z_{ig} = 1 \right] = \sqrt{\frac{a_g}{b_{ig}} R_{\nu} \left( \sqrt{a_g b_{ig}} \right)} - \frac{2\nu}{b_{ig}} := E_{2ig}. \]

On the M-step of each iteration, the updates for the mixing proportions \( \pi_g, \) skewness parameter \( \alpha_g, \) and shift parameter \( \mu_g \) are given by \( \hat{\pi}_g = n_g / n, \)

\[ \hat{\alpha}_g = \frac{\left( \sum_{i=1}^{n} \hat{z}_{ig} E_{2ig} \right) \left( \sum_{j=1}^{n} \hat{z}_{jg} x_j \right) - n_g \sum_{i=1}^{n} \hat{z}_{ig} E_{2ig} x_i}{\left( \sum_{i=1}^{n} \hat{z}_{ig} E_{1ig} \right) \left( \sum_{j=1}^{n} \hat{z}_{jg} E_{2ig} \right) - n_g^2}, \quad (4.4) \]
and

\[ \hat{\mu}_g = \frac{\left( \sum_{i=1}^{n} \hat{z}_{ig} E_{i1g} \right) \left( \sum_{j=1}^{n} \hat{z}_{jg} E_{2jg} x_j \right) - n_g \sum_{i=1}^{n} \hat{z}_{ig} x_i}{\left( \sum_{i=1}^{n} \hat{z}_{ig} E_{1ig} \right) \left( \sum_{j=1}^{n} \hat{z}_{jg} E_{2jg} \right) - n_g^2}, \] (4.5)

respectively, where \( n_g = \sum_{i=1}^{n} \hat{z}_{ig} \). Each component scale matrix \( \Sigma_g \) is then updated using

\[ \hat{\Sigma}_g = S_g - \hat{\alpha}_g r'_g - r'_g \hat{\alpha}_g' + \frac{1}{n_g} \hat{\alpha}_g \hat{\alpha}_g' n_g \sum_{i=1}^{n} \hat{z}_{ig} E_{1ig}, \] (4.6)

where

\[ S_g = \frac{1}{n_g} \sum_{i=1}^{n} \hat{z}_{ig} E_{2ig} \left( x_i - \hat{\mu}_g \right) \left( x_i - \hat{\mu}_g \right)' \]

and

\[ r_g := \frac{1}{n_g} \sum_{i=1}^{n} \hat{z}_{ig} \left( x_i - \hat{\mu}_g \right). \]

After we obtain an estimate for \( \Sigma_g \) we decompose the scale matrix based on the ParSAL model of interest. To preform the decomposition we follow the procedures outlined in Celeux and Govaert (1995) for twelve of the fourteen family members. For the other two family members, specifically the VVE and EVE models (see Table 2.1) we use a combination of two MM algorithms (see Browne and McNicholas, 2014, for details). The use of these MM algorithms is motivated by Browne and McNicholas (2014) who show that the average computational times of their procedures are far less than the Flury method (Flury, 1984) suggested by Celeux and Govaert (1995).

4.3.3 Fitting an EVV Model

As an illustrative example of our approach consider the ParSAL family member with \( \lambda_g = \lambda \), i.e., the EVV model.
The complete-data log-likelihood for the EVV model is given by

\[
l(\lambda, D_g, A_g) = \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \log \pi_g + \sum_{i=1}^{n} \sum_{g=1}^{G} \log \left( \frac{z_{ig}}{\sqrt{2\pi} |w_{ig}\lambda D_g A_g D'_g|} \right)
- \frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} (x_i - \mu_g - w_{ig} \alpha_g)' (\lambda D_g A_g D'_g)^{-1} (x_i - \mu_g - w_{ig} \alpha_g)
= \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \log \pi_g + \sum_{i=1}^{n} \sum_{g=1}^{G} \log \left( \frac{z_{ig}}{\sqrt{2\pi} |w_{ig}|} \right) + \sum_{i=1}^{n} \sum_{g=1}^{G} \frac{z_{ig}}{2} \log |(\lambda D_g A_g D'_g)|^{-1}
- \frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} (x_i - \mu_g - w_{ig} \alpha_g)' (\lambda D_g A_g D'_g)^{-1} (x_i - \mu_g - w_{ig} \alpha_g)
= K + \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} p \log (\lambda)
- \frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} \frac{z_{ig}}{w_{ig} \lambda} (x_i - \mu_g - w_{ig} \alpha_g)' D_g A_g^{-1} D'_g (x_i - \mu_g - w_{ig} \alpha_g)
= K + \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} p \log (\lambda) - \frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} \frac{z_{ig}}{w_{ig} \lambda} \text{tr} \{ W_g D_g A_g^{-1} D'_g \}
\]

where \( K \) is constant with respect to the model parameters: \( \mu_g, \alpha_g \lambda, A_g \) and \( D_g \) and \( W_g = (x_i - \mu_g - w_{ig} \alpha_g) (x_i - \mu_g - w_{ig} \alpha_g)' \).

Let \( C_g = D_g A_g D'_g \) and \( \Sigma_g = \lambda C_g \). To maximize the expected value of the complete-data log-likelihood

\[
\mathcal{Q}(\lambda, D_g, A_g) = K + np \log (\lambda) + \frac{1}{2\lambda} \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} E_{2ig} \text{tr} \{ W_g D_g A_g^{-1} D'_g \},
\]

with respect to \( C_g \) and \( \lambda \) we use

\[
\hat{C}_g = \frac{W_g}{|W_g|^{1/p}} \quad \text{and} \quad \hat{\lambda} = \frac{\sum_{g=1}^{G} W_g |^{1/p}}{n}
\]

54
as outlined in Celeux and Govaert (1995). Note: This procedure constitutes updating the scale matrix $\Sigma_g$ on the M-step of one EM iteration.

To determine if our EM algorithm has converged we use Aitken’s acceleration and stop our EM algorithm when $l_{(t+1)} - l_{(t)} < \epsilon$ (McNicholas et al., 2010). The resulting component membership estimates, $\hat{z}_{ig}$, are then used to cluster the observations in groups via maximum a posteriori (MAP) probabilities, where MAP{$\hat{z}_{ig}$} = 1 if $\max_g{\hat{z}_{ig}}$ occurs in component $g$ and MAP{$\hat{z}_{ig}$} = 0 otherwise. To assess clustering and classification performance we use a cross-tabulation of our MAP classifications against the true group memberships. We then compute the ARI (Hubert and Arabie, 1985).

### 4.3.4 A “Push-Back” Procedure

When estimating the parameter values of our ParSAL mixture models we still have to deal with the computational issues that arise if infinite log-likelihood values are calculated (see Section 3.2.4 for details). For the ParSAL family we adopt a method that preemptively strikes against infinite log-likelihood values by checking to see if the value of any $E[W_{ig}^{-1} \mid x_i, z_{ig} = 1]$ exceeds a given $\epsilon^*$. If any $E[W_{ig}^{-1} \mid x_i, z_{ig} = 1]$ is found to be greater than $\epsilon^*$ than we set $\hat{\mu}_g$ equal to its value on the previous iteration and proceed to update the remaining parameters, $\alpha_g$ and $\Sigma_g$, using (4.4) and (4.6). Again, we acknowledge that this approach is simple-minded but it is both effective and easy to implement.
4.4 Data Analyses

4.4.1 Introduction

We use one simulated data set (Section 4.4.2) and one real data set from a flow cytometry study (Section 4.4.3) to illustrate the classification performance of our ParSAL mixture models. In each situation we compare our Parsimonious SAL models to their multivariate-t and Gaussian analogues using the R packages teigen and mixture.

4.4.2 The ‘Pizza’ Data

The ‘pizza’ data was created by generating and slicing bivariate-t data. Specifically, data was generated from a bivariate-t distribution with 6 degrees of freedom. Making 4 cuts resulted in 8 ‘slices’ or clusters of data. We fitted the ParSAL family and their multivariate-t and Gaussian analogues to the pizza data using k-means starting values for $G = 1, \ldots, 9$ groups. In the event a $G = 9$ component model was selected, the models were rerun for $G = 1, \ldots, 12$. The clustering results are illustrated in Figure 4.1.

The chosen ParSAL VEV model ($\text{BIC} = -4619.284, \text{ICL} = -4620.231$) selects an 8 component mixture model, where each component appears to represent a “slice” of the data. The best fitting multivariate-t solution was the 7-component UUUC ($\text{BIC} = -3135.689, \text{ICL} = -3179.907$) mixture model. Interestingly, the multivariate-t UUUC model is able to recover the structure of the bottom 4 ‘slices’ perfectly. However, it uses only 3 components for the top 4 ‘slices.’ The BIC selected a 12-component VEV model as the best fitting Gaussian solution. It appears as though the Gaussian components are trying to separate regions with
Figure 4.1: Clustering results for the best fitting ParSAL (Panel 1), GPCM (Panel 2) and tEIGEN (Panel 3) models on the ‘pizza’ data; the colours highlight the predicted component memberships.

outliers from those where the data are more highly concentrated. This result is not surprising as Gaussian mixtures tend to use extra components to account for skewness.
4.4.3 Hematopoietic Steam Cell Transplant Data

The hematopoietic steam cell transplant (HSCT) data was collected in the Terry Fox Lab at the British Columbia Cancer Agency. Researchers used flow cytometry to collect information about 9780 cells, each stained with four fluorescent dyes. Evaluation of the data using manual gating identified four clusters comprised of a total of 9702 cells with the other 78 cells being deemed “dead.”

Based on the expert clustering results we fit $G = 4$ component ParSAL, GPCM and tEIGEN models to the HSCT Data. The models were initialized using $k$-means clustering and the BIC was used to select the best fitting model. The clustering results for the best fitting ParSAL VEV ($\text{BIC} = -460414.6$) are displayed in Figure 4.2. Note: The ParSAL VEV model was also chosen using the ICL; $\text{ICL} = -460469.3$. The GPCM VVV ($\text{BIC} = -467450.2$) and tEIGEN CUCU ($\text{BIC} = -45057.69, \text{ICL} = -45365.93$) were chosen as the best fitting Gaussian and multivariate-$t$ models.

Table 4.1 summarizes the clustering performance of each model. The results show that the ParSAL VEV model out-preforms both the symmetric alternatives obtaining a near perfect ARI. The ParSAL VEV models success is attributed to being able to better accommodate for the lack of symmetry in these groups. However, the component skewness parameters do not indicate that any component is skewed in one specific direction.
Figure 4.2: The true cell groupings (represented by colour) and clustering results for the ParSAL VEV (represented by symbol).
Table 4.1: Clustering results for the chosen ParSAL, GPCM and tEIGEN models. The fluorescent dyes are cross-tabulated against our predicted classifications (A, B, C, D) in each case.

<table>
<thead>
<tr>
<th></th>
<th>ParSAL VEV</th>
<th>GPCM VVV</th>
<th>tEIGEN CUCU</th>
</tr>
</thead>
<tbody>
<tr>
<td>X..FL1.H</td>
<td>4262</td>
<td>3994</td>
<td>3954</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>277</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>314</td>
<td>3</td>
</tr>
<tr>
<td>X..FL2.H</td>
<td>37</td>
<td>3110</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3424</td>
<td>351</td>
<td>3448</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>X..FL3.H</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>906</td>
<td>143</td>
<td>125</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>763</td>
<td>781</td>
</tr>
<tr>
<td>X..FL4.H</td>
<td>0</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1058</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1062</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

ARI 0.984 0.841 0.891

4.5 Summary and Discussion

A family of parsimonious SAL mixtures, ParSAL, was introduced via an eigen-decomposition of the component scale matrix. Some of the models drastically reduce the number of free covariance parameters, while others maintain the property of being quadratic in $p$. We demonstrate our models using a simulated and real data set from a flow cytometry experiment. The simulated data set displayed the flexibility of our approach and showed the benefit of being able to parameterize skewness. Although the ‘true’ group memberships are not known the chosen ParSAL model had 8 components with each representing one ‘slice’ of the data. On the other hand the best fitting GPCM chosen the maximum number of groups and appears to be using a large number of components to account for skewness; a familiar Gaussian trait.

We also examined a data set collected from a flow cytometry experiment in British Columbia. Based on expert clustering results we gave each family of models the benefit of knowing the true number of groups. This approach could be considered more closely related to model-based classification however, all group labels were removed before the analysis took place. The ParSAL EVE model gave excellent classification performance (ARI = 0.984)
outperforming both the GPCM (ARI = 0.841) and tEIGEN (ARI = 0.891) families. Both of
the symmetric alternatives also preformed well on the data yielding very good classification
results. However, our ParSAL models were better able to capture the asymmetry in the
data. Interestingly, this data set was also analyzed in Lee and McLachlan (2013) using
various multivariate skew-\(t\) and the general mixture of SAL densities. The ParSAL VEV
model gives quite good performance compared to these skewed alternatives; though all the
skew models fit to this data gave excellent classification performance.
Chapter 5

Extending the Mixture of Factor Analyzers

5.1 Introduction

In this chapter we introduce the second family of parsimonious shifted asymmetric Laplace mixture models. Specifically, we extend the mixture of factor analyzers model to the shifted asymmetric Laplace distribution (cf. Section 2.4.2). Imposing constraints on the constitute parts of the resulting decomposed component scale matrices leads to a family of parsimonious models. An explicit two-stage parameter estimation procedure is described and the Bayesian information criterion and integrated completed likelihood measure are compared for model selection. This novel family of models is applied to simulated and real data, where it is compared to its Gaussian analogue within clustering and classification paradigms.
5.2 The Model

Recall that a random variable \( X \sim \text{SAL}(\mu, \Sigma, \alpha) \) can be generated through the relationship

\[
X = \mu + W\alpha + \sqrt{W}N,
\]

where \( \mu \) and \( \alpha \) are the location and skewness parameters, \( W \sim \text{Exp}(1) \) and \( N \sim \mathcal{N}_p(0, \Sigma) \) are independent of one another.

If we let \( N = \Lambda U + \epsilon \) where \( \Lambda, U \) and \( \epsilon \) are defined in Section 2.4.2 we obtain the SAL factor analysis model

\[
X = \mu + W\alpha + \sqrt{W}(\Lambda U + \epsilon).
\]

It follows that the marginal distribution of \( X \) is \( \text{SAL}(\alpha, \Lambda\Lambda' + \Psi, \mu) \), and the mixture of SAL factor analyzers model has density

\[
f(x \mid \vartheta) = \sum_{g=1}^{G} \pi_g \xi_g(x \mid \alpha_g, \Lambda_g\Lambda'_g + \Psi_g, \mu_g).
\]

Setting \( \Psi_g = \omega_g\Delta_g \), we obtain a mixture of modified SAL factor analyzers, whose density is given by

\[
f(x \mid \vartheta) = \sum_{g=1}^{G} \pi_g \xi_g(x \mid \alpha_g, \Lambda_g\Lambda'_g + \omega_g\Delta_g, \mu_g),
\]

and we can proceed in an analogous fashion to McNicholas and Murphy (2010) to obtain a family of twelve parsimonious shifted asymmetric Laplace mixtures (PSALM). The nomenclature for the PSALM family is analogous to that for the PGMM family (Table 2.2) except that the constrains are on component scale matrices rather than component covariance matrices. Note that the most general member of the PSALM family (UUUU), i.e.,
the mixture of (modified) SAL factor analyzers model, has component covariance matrix $\Lambda_g \Lambda_g' + \omega_g \Delta_g + \alpha_g \alpha_g'$.

\section{Parameter Estimation}

Using an EM algorithm to estimate the parameters of our PSALM model still leads to computational issues when updating $\hat{\alpha}_g$ and $\hat{\Sigma}_g$ (cf. Section 3.2.4). To overcome this problem we propose a two-stage parameter estimation procedure for the PSALM models. In stage 1, we use a modified deterministic annealing algorithm, and in stage 2, an AECM algorithm.

\subsection{A Two-Phase Approach}

There are three sources of missing data in each PSALM model, the latent variable $w_{ig}$, the component labels $z_i$, and the latent factors $u_{ig}$.

From Section 5.1, the complete-data log-likelihood can be written

\begin{equation}
    l_c(\pi_g, \mu_g, \Lambda_g, \omega_g, \Delta_g, \alpha_g) = \sum_{i=1}^n \sum_{g=1}^G \left[ \log \pi_g + \log h(w_{ig}) + \log \phi(x_i | \mu_g + w_{ig} \alpha_g, w_{ig} \Lambda_g \Lambda_g' + w_{ig} \omega_g \Delta_g) \right] z_{ig}, \tag{5.1}
\end{equation}

where $\phi(x_i | \mu_g + w_{ig} \alpha_g, w_{ig} \Lambda_g \Lambda_g' + w_{ig} \omega_g \Delta_g)$ is the density of a multivariate Gaussian distribution with mean $\mu_g + w_{ig} \alpha_g$ and covariance matrix $w_{ig} \Lambda_g \Lambda_g' + w_{ig} \omega_g \Delta_g$, and $h(w_{ig})$ is the density of an exponential distribution with rate 1, i.e., $h(z) = e^{-1}$, for real $z \geq 0$.

On the first $E$-step of our modified deterministic annealing algorithm the complete-data
consist of the observed $x_i$, the component membership labels $z_{ig}$, and the latent variable $w_{ig}$, for $i = 1, \ldots, n$ and $g = 1, \ldots, G$. Their expected values are given by:

$$
\mathbb{E}[Z_{ig} | x_{i}] = \frac{[\pi_{g} \xi(x_{i} | \alpha_{g}, \Sigma_{g}, \mu_{g})]^{v}}{\sum_{h=1}^{G} [\pi_{h} \xi(x_{i} | \alpha_{h}, \Sigma_{h}, \mu_{h})]^{v}} := \hat{z}_{ig}^{*},
$$

$$
\mathbb{E}[W_{ig} | x_{i}, z_{ig} = 1] = \sqrt{\frac{b_{ig}}{a_{g}}} R_{\nu} \left( \sqrt{a_{g} b_{ig}} \right) := E_{1ig} \text{ and,}
$$

$$
\mathbb{E}[W_{ig}^{-1} | x_{i}, z_{ig} = 1] = \sqrt{\frac{a_{g}}{\psi + b_{ig}}} R_{\nu} \left( \sqrt{a_{g}(\psi + b_{ig})} \right) - \frac{2\nu}{\psi + b_{ig}} := E_{2ig}^{*},
$$

where $v$ is defined in Section 2.6.4, $a_{g} := 2 + \alpha_{g}' \Sigma_{g}^{-1} \alpha_{g}$ and $b_{ig} := \delta(x_{i}, \mu_{g} | \Sigma_{g})$. We introduce the parameter $\psi$ to prevent $\mu_{g}$ from becoming equal to any observation $x_{i}$. The introduction of $\psi$ is motivated by the limiting cases of the generalized hyperbolic density discussed in Section 2.3. Moreover, it is because of the presence of $\psi$ that we refer to this deterministic algorithm as “modified.”

On the first CM-step the updates for the mixing proportions, skewness and shift parameters are given by $\hat{\pi}_{g} = n_{g}/n$,

$$
\hat{\alpha}_{g} = \frac{\sum_{i=1}^{n} \hat{z}_{ig}^{*} E_{2ig}^{*} (\sum_{j=1}^{n} \hat{z}_{jg}^{*} x_{j}) - n_{g} \sum_{i=1}^{n} \hat{z}_{ig}^{*} E_{2ig}^{*} x_{i}}{\sum_{i=1}^{n} \hat{z}_{ig}^{*} E_{1ig}^{*} (\sum_{j=1}^{n} \hat{z}_{jg}^{*} E_{2jg}^{*}) - n_{g}^{2}},
$$

and

$$
\hat{\mu}_{g} = \frac{\sum_{i=1}^{n} \hat{z}_{ig}^{*} E_{1ig}^{*} (\sum_{j=1}^{n} \hat{z}_{jg}^{*} E_{2jg}^{*} x_{j}) - n_{g} \sum_{i=1}^{n} \hat{z}_{ig}^{*} x_{i}}{\sum_{i=1}^{n} \hat{z}_{ig}^{*} E_{1ig}^{*} (\sum_{j=1}^{n} \hat{z}_{jg}^{*} E_{2jg}^{*}) - n_{g}^{2}},
$$

respectively, where $n_{g} = \sum_{i=1}^{n} \hat{z}_{ig}^{*}$.

On the second $E$-step the complete-data include the same constituents as the first $E$-step plus the latent factors $u_{ig}$, for $i = 1, \ldots, n$ and $g = 1, \ldots, G$. The expected values of the
latent factors are

\[ \mathbb{E}[Z_{ig}W_{ig}^{-1}U_{ig} | x_i, w_{ig}^{-1}] = w_{ig}^{-1}\beta_g(x_i - \mu_g - w_i\alpha_g) \text{ and,} \]

\[ \mathbb{E}[Z_{ig}W_{ig}^{-1}U_{ig} U'_{ig} | x_i, w_{ig}^{-1}] = I_q - \beta_g\Lambda_g + w_{ig}^{-1}\beta_g(x_i - \mu_g - w_i\alpha_g)(x_i - \mu_g - w_i\alpha_g)\beta'_g, \]

where \( \beta_g = \Lambda'_g(\Lambda_g\Lambda'_g + \omega_g\Delta_g)^{-1} \) and \( w_{ig}^{-1} \) will be replaced by \( E_{2ig}^* \). Note that these expected values are similar to those used by McNicholas and Murphy (2008) and others.

On the second CM-step the updates for \( \Lambda_g, \omega_g, \) and \( \Delta_g \) will depend on which PSALM family member is under consideration. Consider, for example, the mixture of SAL factor analyzers model (i.e., UUUU). In this case, the updates are

\[ \hat{\Lambda}_g = S_g\beta'_g(I_q - \beta_g\Lambda_g + \beta_gS_g\beta'_g)^{-1}, \]

\[ \hat{\omega}_g = |\hat{\Psi}_g|^{1/p}, \text{ and} \]

\[ \hat{\Delta}_g = \hat{\Psi}_g/|\hat{\Psi}_g|^{1/p}, \]

where \( \hat{\Psi}_g = \text{diag}\{S_g - \Lambda_g\hat{\beta}_gS_g\} \) and

\[ S_g = \frac{1}{n_g} \sum_{i=1}^{n} z_{ig}^*E_{2ig}^*(x_i - \mu_g)(x_i - \mu_g)' - \alpha_g r_g' - r_g\hat{\alpha}_g + \frac{1}{n_g} \hat{\alpha}_g\hat{\alpha}_g' \sum_{i=1}^{n} \hat{z}_{ig}E_{1ig}, \]

where \( r_g := (1/n_g)\sum_{i=1}^{n} \hat{z}_{ig}(x_i - \hat{\mu}_g) \). Updates for the other models (Table 2.2) are analogous to those given in McNicholas and Murphy (2008, 2010).

After our deterministic annealing algorithm is run, the second phase of our parameter estimation procedure begins. This AECM algorithm is identical to our modified deterministic
annealing algorithm except that, on each $E$-step, the

$$\mathbb{E}[Z_{ig} \mid x_i] = \frac{\pi_g \xi(x_i \mid \alpha_g, \Sigma_g, \mu_g)}{\sum_{h=1}^{G} \pi_h \xi(x_i \mid \alpha_h, \Sigma_h, \mu_h)} := \hat{z}_{ig}, \tag{5.2}$$

and

$$\mathbb{E}[W_{ig}^{-1} \mid x_i, z_{ig} = 1] = \sqrt{\frac{a_g}{b_{ig}}} R_{\nu} \left( \sqrt{a_g b_{ig}} \right) - \frac{2\nu}{b_{ig}} := E_{2ig}.$$

On the first CM step we now update only the mixing proportions, $\hat{\pi}_g$ and skewness parameter, $\hat{\alpha}_g$, for $g = 1, \ldots, G$ where $n_g = \sum_{i=1}^{n} \hat{z}_{ig}$. On the second $CM$-step the updates for $\hat{\Lambda}_g$, $\hat{\omega}_g$, and $\hat{\Delta}_g$ are also the same as in our modified deterministic annealing algorithm but with $\hat{z}_{ig}^*$ and $E_{2ig}$ replaced by $\hat{z}_{ig}$ and $E_{2ig}$, respectively. Accordingly, we now have

$$S_g = \frac{1}{n_g} \sum_{i=1}^{n} \hat{z}_{ig} E_{2ig} (x_i - \mu_g) (x_i - \mu_g)' - \hat{\alpha}_g r_g' - r_g \hat{\alpha}_g' + \frac{1}{n_g} \hat{\alpha}_g \hat{\alpha}_g' \sum_{i=1}^{n} \hat{z}_{ig} E_{1ig}$$

in the updates for $\hat{\Lambda}_g$, $\hat{\omega}_g$, and $\hat{\Delta}_g$. After convergence (cf. Section 2.7), clustering results are reported based on maximum $a posteriori$ (MAP) classification values. That is, $\text{MAP}\{\hat{z}_{ig}\} = 1$ if $\max_h \{\hat{z}_{ih}\}$ occurs in component $h = g$, and $\text{MAP}\{\hat{z}_{ig}\} = 0$ otherwise. In other words, each observation is assigned to the component to which it has the highest $a posteriori$ probability of membership.
5.3.2 Woodbury Identity

To avoid inverting any non-diagonal $p \times p$ matrices we make use of the Woodbury Identity (Woodbury, 1950), which states that

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1},$$

where $A$ is an $m \times m$ matrix, $U$ is an $m \times k$ matrix, $C$ is an $k \times k$ matrix, and $V$ is an $k \times m$ matrix. Setting $U = \Lambda$, $V = \Lambda'$, $A = \omega \Delta$, and $C = I_q$ we can write

$$(\omega \Delta + \Lambda \Lambda)^{-1} =$$

$$(\omega \Delta)^{-1} - (\omega \Delta)^{-1}\Lambda(I_q + \Lambda'(\omega \Delta)^{-1}\Lambda)^{-1}\Lambda'(\omega \Delta)^{-1}.$$

Following from this we can compute the determinant using

$$|\omega \Delta + \Lambda \Lambda'| = |\omega \Delta|/|I_q + \Lambda'(\Lambda \Lambda' + \omega \Delta)^{-1}\Lambda|.$$

These identities provide a major computational advantage as $p$ grows, because $q \ll p$, and have been used by many authors including McLachlan and Peel (2000), McNicholas and Murphy (2008), Andrews and McNicholas (2011a), Andrews and McNicholas (2011b), and Murray et al. (2013).
5.3.3 Model-Based Classification

The PSALM model-based classification likelihood is given by

\[
L(x_1, \ldots, x_n, z_1, \ldots, z_n \mid \theta) = \prod_{i=1}^{k} \prod_{g=1}^{G} \left[ \pi_g \xi \left( x_i \mid \alpha_g, \Lambda_g, \Lambda'_g + \omega_g \Delta_g, \mu_g \right) \right]^{z_{ig}} \times \prod_{j=k+1}^{n} \sum_{h=1}^{H} \pi_h \xi \left( x_j \mid \alpha_h, \Lambda_h, \Lambda'_h + \omega_h \Delta_h, \mu_h \right),
\]

where \( H \geq G \). Parameter estimation is analogous to model-based clustering (cf. Section 5.3).

5.4 Data Analyses

5.4.1 Introduction

We use a simulation study (Section 5.4.2) and three real data sets: Leptograpsus Crabs (Section 5.4.3), Swiss Bank Notes (Section 5.4.4) and the Yeast Data (Section 5.4.5) to demonstrate the classification ability of our PSALM mixture models. Our simulation study is used to demonstrate our models abilities in high-dimensional situations.

5.4.2 Simulation Study

To evaluate the PSALM clustering performance in high dimensions multivariate asymmetric Laplace data was generated from a 2 component mixture model. In total 80 data sets were simulated; 20 for \( p = 10 \), \( p = 20 \), \( p = 50 \) and \( p = 100 \) respectively with \( n = 200 \) observations.
each. The mixture components were simulated with a significant amount of overlap and therefore perfect clustering is not expected. Furthermore, we set $G = 2$ in these simulations and find that, in general, the chosen PSALM models outperform the best fitting PGMM in each dimension, in spite of this benefit.

Table 2 presents the average ARI for both the best fitting PSALM and PGMMs for each dimension. For all 80 data sets the BIC and ICL choose the same model and therefore agreed on ARI in all cases. Figure 5.1 shows an example of a typical PSALM fit compared to the PGMM classification of a $p = 10$ simulated data set.

Table 5.1: Clustering results for the chosen PSALM and PGMM for the simulated data.

<table>
<thead>
<tr>
<th>G = 2</th>
<th>PSALM</th>
<th>PGMM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p = 10</td>
<td>p = 20</td>
</tr>
<tr>
<td>Average ARI</td>
<td>0.823</td>
<td>0.827</td>
</tr>
</tbody>
</table>

5.4.3 Leptograpsus Crabs

Campbell and Mahon (1974) give data on 200 crabs of the species *Leptograpsus variegatus* collected at Fremantle, Western Australia. The data are available in the R package MASS and contain 5 morphological measurements: frontal lobe size (mm), rear width (mm), carapace length (mm), carapace width (mm) and body depth (mm). The crabs are of two genders and two colours (blue and orange). In this section we consider the principal components of the crabs data, and in Section 5.4.6 we consider the full data set.

As one might expect, the variables in the crabs data are highly correlated with one another. The PSALM and PGMM families were fit to the first and third principal components for
Figure 5.1: An example of the predicted group memberships for the chosen PSALM and PGMM models on a simulated data set with $p = 10$ dimensions.

$G = 1, \ldots, 9$ groups and $q = 1$ latent factor. The best PSALM model, as chosen by both the BIC ($-812.9913$) and ICL ($-815.9154$), was the $G = 2$ component UCCC model, which gave perfect classification with respect to gender (ARI = 1.00). The best PGMM also had the UCCC covariance structure (BIC = ICL = $-741.3885$) but used $G = 3$ components to fit the data (ARI = 0.748). Figure 5.2 shows the resulting MAP classifications as well as contours for the fitted PSALM and PGMM models. Note, the classifications for the chosen PGMM
model can be made equal to those for the chosen PSALM model by merging components.

Figure 5.2: The first and third principal components of the crabs data with contours reprint-
ing the fit of the chosen PSALM and PGMM models.
5.4.4 Gaussian Cluster Merging & Swiss Bank Notes

The results for the chosen PGMM model in Section 5.4.3 raise an interesting point, i.e., that merging Gaussian mixture components can sometimes give identical classification performance to a non-Gaussian mixture (cf. Figure 5.2). Baudry et al. (2010) and Hennig (2010) discuss methods for merging components. Now, we will look at another example to reinforce the point that the performance of non-Gaussian approaches can sometimes be matched by Gaussian mixtures followed by merging.

Consider the Swiss bank notes data, which are available in the R package gclus (Hurley, 2004). The bank notes data are composed of 200 Swiss bills, of which 100 are counterfeit and 100 are legitimate. For each bill there are six physical measurements available. The PSALM and PGMM families were fit to these data for $G = 1, \ldots, 9$ components and $q = 1, 2$ and 3 latent factors. The classification results (Table 5.2) show that, like the analysis of the crabs principal components, the classification performance of the chosen PGMM model (CUCU) can be made equal to that of the chosen PSALM model (CCCU) by merging components.

Table 5.2: Clustering results for the chosen PSALM and PGMM for the Swiss bank notes data. The bill types are cross-tabulated against our predicted classifications (A, B) in each case.

<table>
<thead>
<tr>
<th></th>
<th>PSALM</th>
<th></th>
<th>PGMM</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Counterfeit</td>
<td>99</td>
<td>1</td>
<td>21</td>
<td>0</td>
</tr>
<tr>
<td>Legitimate</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>16</td>
<td>0</td>
</tr>
</tbody>
</table>
5.4.5 Yeast Data

The PSALM and PGMM families were fitted to the yeast data (Section 3.3.4) for \( G = 1, \ldots, 9 \) components and \( q = 1 \) latent factor. The chosen PGMM model is a \( G = 6 \) component CCUU model (\( \text{BIC} = \text{ICL} = -4993.137, \text{ARI} = 0.22 \)). For the PSALM family, the ICL \((-5275.705)\) selects a \( G = 2 \) component CUUU model (\( \text{ARI} = 0.86 \)), and the BIC \((-5220.763)\) chooses a \( G = 2 \) component UUCU model (\( \text{ARI} = 0.83 \)). The classification results for the best fitting Gaussian and SAL mixtures, as chosen by BIC, are given in Table 5.3. From this table

Table 5.3: Clustering results for the chosen PSALM and PGMM for the yeast data. The cellular localization sites are cross-tabulated against our predicted classifications (A, B, C, D, E, F) in each case.

<table>
<thead>
<tr>
<th></th>
<th>PSALM UUCU</th>
<th>PGMM CCUU</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>CYT</td>
<td>454</td>
<td>9</td>
</tr>
<tr>
<td>ME3</td>
<td>13</td>
<td>150</td>
</tr>
</tbody>
</table>

and Figure 5.3, we can see that merging components could improve the PGMM solution. However, even under the optimal merging scenario, i.e., combining PGMM components B through F (cf. Table 5.3), the resulting classification performance (\( \text{ARI} = 0.705 \)) is still not as good the chosen PSALM model.

5.4.6 Model-Based Classification

Thus far, our applications have focused on model based clustering. Now, we consider applying the PSALM and PGMM families in model-based classification scenarios to the crabs and yeast data. In each case we take 50 random subsets of 80\% of the labels to be known and we assume that we know the number of classes. Note that we consider the full \( p = 5 \) variable,
Figure 5.3: The yeast data with predicted group memberships highlighted for the chosen PSALM and PGMM models.
$G = 4$ component crabs data set, which presents a much more challenging classification problem than using the principal components. The aggregate classification results are given in Tables 5.4 and 5.5. In each case, the PSALM clearly outperforms the PGMM family.

Table 5.4: Aggregated classification results for the PSALM CCCU (ARI = 0.853) and PGMM UCUU (ARI = 0.737) for the Leptograpsus crabs data. The colour and gender combinations are cross-tabulated against our predicted classifications (A, B, C, D) in each case.

<table>
<thead>
<tr>
<th></th>
<th>PSALM</th>
<th>PGMM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Blue Male</td>
<td>461</td>
<td>47</td>
</tr>
<tr>
<td>Orange Male</td>
<td>32</td>
<td>501</td>
</tr>
<tr>
<td>Blue Female</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Orange Female</td>
<td>0</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 5.5: Clustering results for the chosen PSALM and PGMM for the yeast data. The localization sites are cross-tabulated against our predicted classifications (A, B) in each case.

<table>
<thead>
<tr>
<th></th>
<th>PSALM CCCU</th>
<th>PGMM CCUC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>CYT</td>
<td>4591</td>
<td>48</td>
</tr>
<tr>
<td>ME3</td>
<td>184</td>
<td>1477</td>
</tr>
</tbody>
</table>

5.5 Summary and Discussion

We have extended the mixture of factor analyzers model using SAL mixtures. Based on the resulting mixture of (modified) SAL factor analyzers model, a new family of mixture models, i.e., PSALM, was developed. The PSALM models are well suited for the analysis of high-dimensional data because the covariance structure allows for $p$-dimensional data to be represented by $q$ latent factors where $q \ll p$; crucially, the number of covariance parameters is linear in data dimensionality for each member of the PSALM family.
A two-stage approach was taken to parameter estimation for members of the PSALM family, consisting of deterministic annealing followed by an AECM algorithm. The performance of the PSALM family was compared to the PGMM family in both model-based clustering and classification scenarios. In these applications, the BIC and the ICL were used for model selection. Our PSALM models gave similar or superior classification performance to the PGMM family and although merging could sometimes be used to bring the classification performance of the PGMM models up to that of the PSALM family, this was not always the case.
Chapter 6

Multiple Scaled Shifted Asymmetric Laplace Distributions

6.1 Introduction

An alternative parameterization of the mixture of shifted asymmetric Laplace distributions component scale matrices is introduced. This parameterization allows for the inclusion of a multidimensional weight parameter that creates a generalization of the multivariate SAL density. We refer to this model as a multiple scaled distribution. A parameter estimation procedure is outlined via an EM algorithm and we demonstrate our model using real data.
6.2 Multiple Scaled Distributions

Consider a random variable \( Y \) from a multivariate-\( t \) distribution. It is well-known that the multivariate-\( t \) distribution has many forms and each can be represented as normal variance-mean mixture. One of its most useful representations is given by

\[
t(y \mid \mu, \Sigma, \nu) = \int_0^\infty \phi_p(y \mid \mu, \Sigma/w) g(w \mid v/2, v/2) \, dw
\]

where \( \phi_p(y \mid \mu, \Sigma/w) \) is the density of the multivariate Gaussian distribution with mean \( \mu \) and covariance matrix \( \Sigma/w \), and \( g(v/2, v/2) \) is the density of the Gamma distribution with degrees of freedom parameter \( \nu \) given by

\[
g(x \mid \alpha, \gamma) = x^{\alpha-1} \gamma^{-1} (\alpha)^{-1} \exp(-\gamma x) \gamma^\alpha.
\]  

In general, we can write that the density of a random variable \( Y \in \mathbb{R}^p \) whose distribution is a normal variance-mean mixture as

\[
p(y \mid \mu, \Sigma, \theta) = \int_0^\infty \phi_p(y \mid \mu, \Sigma/w) f_W(w \mid \theta) \, dw
\]

where \( f_W(w \mid \theta) \) is the probability distribution of a univariate random variable \( W \geq 0 \) (cf. Section 2.1).

Forbes and Wraith (2013) present a modelling framework that allows for the incorporation of a multidimensional weight parameter by decomposing the covariance matrix, \( \Sigma \), using the eigen-decomposition. Recall, the eigen-decomposition of a covariance matrix is given by
\[ \Sigma = \mathbf{D} \mathbf{A} \mathbf{D}' \], where \( \mathbf{D} \) is a matrix of eigenvectors and \( \mathbf{A} \) is a diagonal matrix containing the eigenvalues of \( \Sigma \). Given this representation the authors show a \( p \times p \) matrix, \( \Delta_w = \text{diag}(w_1, \ldots, w_p) \) can be incorporated into a normal variance-mean mixtures density simply by replacing the scaled Gaussian density in equation (6.3) with \( \phi_p(y \mid \mu, \mathbf{D} \mathbf{A} \Delta_w \mathbf{D}') \).

Formally, this gives

\[
\begin{align*}
  f(y \mid \mu, \mathbf{D}, \mathbf{A}, \theta) &= \int_0^\infty \cdots \int_0^\infty \phi_p(y \mid \mu, \mathbf{D} \mathbf{A} \Delta_w \mathbf{D}') f_W(w_1, \ldots, w_p \mid \theta) dw_1 \cdots dw_p, \quad (6.4)
\end{align*}
\]

where \( f_W(w_1, \ldots, w_p \mid \theta) = f_W(w_1 \mid \theta_1) \times \cdots \times f_W(w_p \mid \theta_p) \) is now a \( p \)-variate density function such that each weight is independent. Note: If all the weights are equal to one, a standard multivariate Gaussian distribution is recovered.

Now, we can write

\[
\begin{align*}
  \phi_p(y \mid \mu, \mathbf{D} \mathbf{A} \Delta_w \mathbf{D}') &= \prod_{j=1}^p \phi_1\left( \left[ \mathbf{D}'(\mathbf{x}) \right]_j \mid \left[ \mathbf{D}'(\mu) \right]_j, a_j w_j^{-1} \right) \\
  &= \prod_{j=1}^p \phi_1\left( \left[ \mathbf{D}'(\mathbf{x} - \mu) \right]_j \mid 0, a_j w_j^{-1} \right) \\
  &= \prod_{j=1}^p a_j^{-1/2} \phi_1\left( \left[ \mathbf{D}'(\mathbf{x} - \mu) \right]_j \mid 0, w_j^{-1} \right), \quad (6.5)
\end{align*}
\]

where \( \left[ \mathbf{D}'(\mathbf{x} - \mu) \right]_j \) denotes the \( j \)th component vector of \( \mathbf{D}'(\mathbf{x} - \mu) \) and \( a_j \) the \( j \)th diagonal element of the matrix \( \mathbf{A} \) i.e., \( j \)th eigenvalue. Therefore, the density in (6.5) can be written

\[
\begin{align*}
  f(y \mid \mu, \mathbf{D}, \mathbf{A}, \theta) &= \prod_{j=1}^p \int_0^\infty \phi_1\left( \left[ \mathbf{D}'(\mathbf{x} - \mu) \right]_j \mid 0, a_j w_j^{-1} \right) f_{W_j}(w_j \mid \theta_j) dw_j. \quad (6.6)
\end{align*}
\]

Letting \( f_{W_j}(w_j \mid \theta_j) \) be the gamma density \( g(w_j \mid \nu_j/2, \nu_j/2) \) given in (6.2) leads to a gen-
eralization of the multivariate-\(t\) distribution, called a multiple scaled distribution. Formally, the density of the general multiple scaled multivariate-\(t\) distribution is given by

\[
f(y \mid \mu, A, D, \nu) = \prod_{j=1}^{p} \frac{\Gamma((v_j + 1)/2)}{\Gamma(v_j/2)(a_j v_j \pi)^{1/2}} \left(1 + \frac{D'(y - \mu)[j]^2}{a_j v_j}\right)^{-(v_j+1)/2},
\]

where \(\mu\) is a location parameter, \(\nu = v_1, \ldots, v_p\) is a vector containing the degrees of freedom in each dimension and \(A\) and \(D\) are previously defined.

In addition to this generalization Forbes and Wraith (2013) derived three other multiple scaled distributions using the density given in (6.6). Specifically, the authors give multivariate representations of a Pearson type VII distribution (see, e.g. Johnson et al., 1994, vol. 2 chap. 28), the so-called multivariate K model (Eltoft et al., 2006b), and the multivariate normal inverse Gaussian distribution (Karlis and Santourian, 2009). However, only a parameter estimation procedure for only the multivariate-\(t\) generalizations are presented.

### 6.3 Mixtures of Multiple Scaled Shifted Asymmetric Laplace distributions

Recall from Section 2.5.2 that a random variable \(X\) following the shifted asymmetric Laplace distribution can be generated through the relationship

\[
X = \mu + w \alpha + \sqrt{w}N,
\]

(6.8)
where \( w \) follows an exponential distribution with rate 1 and \( N \) is a multivariate Gaussian random variable with mean \( 0 \) and covariance \( \Sigma \). For our purposes, a more fruitful representation of (6.8) is given by

\[
X = \mu + D A \Delta \omega D' \beta + N, \tag{6.9}
\]

where the parameters: \( \mu, D, A \) and \( \Delta \omega \) are previously defined, \( N \sim N_p(0, DA\Delta \omega D') \), and \( \beta \in \mathbb{R}^p \).

It follows from this relationship that the density of \( X \) can be represented by

\[
\xi(x | \beta, D, A, \mu, \theta) = \int_0^\infty \cdots \int_0^\infty \phi_p(x | \mu + DA\Delta \omega D' \beta, DA\Delta \omega D')
\times f_W(w_1, \ldots, w_p | \theta) \, dw_1 \cdots dw_p
\]

\[
= \int_0^\infty \cdots \int_0^\infty \phi_p(D'x | D'\mu + A \Delta \omega D' \beta, A \Delta \omega)
\times f_W(w_1, \ldots, w_p | \theta) \, dw_1 \cdots dw_p
\]

\[
= \prod_{j=1}^p \int_0^\infty \phi_1\left( [D'x]_j | [D'\mu]_j + [A \Delta \omega D' \beta]_j, A_j w_j \right)
\times f_{W_j}(w_j | \theta_j) \, dw_j
\]

\[
= \prod_{j=1}^p \int_0^\infty \phi_1\left( [D'(x - \mu - A \Delta \omega \beta)]_j | 0, a_j w_j \right)
\times f_{W_j}(w_j | \theta_j) \, dw_j, \tag{6.10}
\]

where \( \phi_1\left( [D'(x - \mu - A \Delta \omega \beta)]_j | 0, a_j w_j \right) \) is the multivariate Gaussian density with mean 0 and covariance matrix \( a_j w_j \), and \( f_{W_j}(w_j | \theta_j) = e^{-w_j} \), for \( W_j > 0 \). and our generalization
of the multivariate SAL distribution is given by

\[ h(x | \beta, D, A, \mu) = \prod_{j=1}^{p} \frac{1}{\gamma_j} \exp \left\{ -\frac{|D'[x - \mu]_j|}{a_j} \left[ \gamma_j - [AD'\beta]_j \text{sign}(D'[x - \mu]_j) \right] \right\}, \quad (6.11) \]

where \( \gamma_j = \sqrt{|D'\alpha|_j^2 + 2a_j}, \beta \in \mathbb{R}^p, D \) is the matrix of eigenvectors, \( A \) is a diagonal matrix of eigenvalues, and \( \mu \) is a location parameter (cf. Section 2.5.1). Figure 6.1 gives contour plots of the possible shapes that we can obtain using this distribution. For these plots we let \( \beta \) and eigenvector matrix \( D \) vary. Following Forbes and Wraith (2013) we parameterize \( D \) via an angle \( \zeta \) so that \( D_{11} = D_{22} = \cos \zeta \) and \( D_{21} = -D_{12} = \sin \zeta \).

Herein we say that a random variable \( X \) is distributed multiple scaled shifted asymmetric Laplace (MSSAL) if it has density given by (6.11). Therefore, the density of a mixture of multiple scaled shifted asymmetric Laplace distributions is given by

\[ f(x | \theta) = \sum_{g=1}^{G} \pi_g h(x | \beta_g, D_g, A_g, \mu_g), \quad (6.12) \]

where \( \pi_g \) are the mixing proportions and \( h(x | \mu_g, \beta_g, D_g, A_g) \) is the density of the MSSAL distributions given in (6.11) with component location parameter \( \mu_g \), component eigenvector matrix \( D_g \), component eigenvalue matrix \( A_g \), and \( \beta_g \in \mathbb{R}^p \).

### 6.4 Parameter Estimation

We use an EM algorithm to calculate the maximum likelihood estimates of the mixture of MSSAL distributions. For this model the complete-data is composed of the observed data,
Figure 6.1: Bivariate contour plots of the multiple scaled SAL density with \( \mu = (0, 0)' \), \( A = \text{diag}(1, 1) \) and varying \( \beta \) and \( \zeta \).
\( \mathbf{x}_1, \ldots, \mathbf{x}_n \), the missing weight variables \( \Delta_{wig} = \text{diag}(w_{1g}, \ldots, w_{pg}) \) and the component indicator variables, \( \mathbf{z}_1, \ldots, \mathbf{z}_n \). For each \( i \), we have \( \mathbf{z}_i = (z_{i1}, \ldots, z_{iG}) \), where \( z_{ig} = 1 \) if observation \( i \) is in group \( g \) and \( z_{ig} = 0 \) otherwise.

For our MSSAL mixtures the complete-data likelihood is given by

\[
L_c = \prod_{i=1}^{n} \prod_{g=1}^{G} \pi_g \prod_{j=1}^{p} \int_{0}^{\infty} \phi_1 \left( \left[ D'_g (\mathbf{x} - \mu_g - A_g \Delta_{wig} \beta_g) \right]_j | 0, a_{jg} w_j \right) f_{W_j} (w_j | \theta_j) \right]^{z_{ig}},
\]

where \( \pi_g \) are the mixing proportions, and \( \phi_1 \left( \left[ D'_g (\mathbf{x} - \mu_g - A_g \Delta_{wig} \beta_g) \right]_j | 0, a_{jg} w_j \right) \) and \( f_{W_j} (w_j | \theta_j) \) are defined for (6.10). The expected value of the complete-data log-likelihood can be written

\[
Q = C + \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} \log |A_g| - \frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} (\mathbf{x}_i - \mu_g - \Omega_{ig} \beta_g)' D_g D_g^{-1} D_g (\mathbf{x}_i - \mu_g - \Omega_{ig} \beta_g), \tag{6.13}
\]

where \( \Omega_{ig} = D_g \Delta_{wig} A_g ' D_g, \hat{z}_{ig}, \Delta_{wig}, \) and \( \Delta_{wig}^{-1} \) are, respectively, the expected values of the \( z_{ig}, \Delta_{wig}, \) and \( \Delta_{wig}^{-1} \) (defined in Section 6.4.1), and \( C \) is constant with respect to the model parameters: \( \beta_g, D_g, A_g, \) and \( \mu_g \).

### 6.4.1 E-step

On each iteration we require the following expectations for the E-step, which effectively amounts to replacing the sufficient statistics of the unobserved data by their expected values.
We replace the $z_{ig}$ with their expected value given by

$$
\hat{z}_{ig} := \frac{\pi_g h (x_i \mid \beta_g, D_g, A_g, \mu_g)}{\sum_{h=1}^{G} \pi_h h (x_i \mid \beta_h, D_h, A_h, \mu_h)},
$$

(6.14)

and $\Delta_{wig}$ and $\Delta_{wig}^{-1}$ with the $p \times p$ matrices $\Delta_{wig}$ and $\Delta_{wig}^{-1}$ whose off-diagonal elements are equal to zero and diagonal elements are given by:

$$
\mathbb{E}[W_{ijg} \mid x_i, z_{ig} = 1] = \sqrt{\frac{b_{ijg}}{d_{jg}}} R_\nu \left( \sqrt{d_{jg} b_{ijg}} \right) =: E_{1ijg}, \text{ and}
$$

$$
\mathbb{E}[W_{ijg}^{-1} \mid x_i, z_{ig} = 1] = \sqrt{\frac{d_{jg}}{b_{ijg}}} R_\nu \left( \sqrt{d_{jg} b_{ijg}} \right) - \frac{2\nu}{b_{ijg}} =: E_{2ijg},
$$

(6.15)

where $d_{jg} = 2 + [AD'_g \beta]^2 / a_j$, $b_{ijg} = [D'_g (x_i - \mu_g)]^2 / a_j$, and $R_\nu (c) := K_{\nu + 1} (c) / K_{\nu} (c)$, where $K_\nu (c)$ is the modified Bessel function of the third kind with index $\nu$ for $c \in \mathbb{R}^+$. The expected values given in (6.15) have these convenient forms because $W_{ijg} \mid \mathbf{X} = \mathbf{x} \sim GIG(d_{jg}, b_{ijg})$, where $GIG(d_{jg}, b_{ijg})$ denotes the generalized inverse Gaussian distribution (cf. Section 2.2). Note that the value $[D'_g (x_i - \mu_g)]^2$ can be regarded as the squared Mahalanobis distance between $[D'_g x_i]_j$ and $[D'_g \mu_g]_j$ based on variance $a_j$.

6.4.2 M-step

To maximize $Q$ with respect to the $\mu_g$ and $\beta_g$ we compute maximum likelihood estimates (see Appendix B for details). Specifically, the mixing proportions, location parameter and
\( \beta_g \) are replaced by \( \hat{\pi}_g = n_g/n, \)

\[
\hat{\mu}_g = \left( \sum_{i=1}^{n} \hat{z}_{ig} \Omega_{ig}^{-1} \right)^{-1} \sum_{i=1}^{n} \hat{z}_{ig} \Omega_{ig}^{-1} x_i - n_g \hat{\beta}_g,
\]

and

\[
\hat{\beta}_g = \left( \sum_{i=1}^{n} \hat{z}_{ig} \Omega_{ig} \right)^{-1} \sum_{i=1}^{n} \hat{z}_{ig} x_i - n_g \mu_g
\]

where \( \Omega_{ig}^{-1} = D_g \Delta_{wig}^{-1} A_g^{-1} D_g' \).

To maximize \( Q \) with respect to \( D_g \) we make use of two separate majorization-minimization (MM) procedures (cf. Browne and McNicholas, 2014). Specifically, we are interested in minimizing the objective function

\[
f(D_g) = \sum_{i=1}^{n} \text{tr} \left\{ \hat{z}_{ig} D_g (\tilde{\Delta}_{wig} A_g)^{-1} D_g' (x_i - \mu_g - \Omega_g \beta_g) (x_i - \mu_g - \Omega_g \beta_g)' \right\}
\]

\[
= \sum_{i=1}^{n} \text{tr} \left\{ \hat{z}_{ig} D_g (\tilde{\Delta}_{wig} A_g)^{-1} D_g' (x_i - \mu_g) (x_i - \mu_g)' \right\}
\]

\[
- \sum_{i=1}^{n} \text{tr} \left\{ \hat{z}_{ig} (x_i - \mu_g)' + \hat{z}_{ig} (x_i - \mu_g) \beta_g' - \hat{z}_{ig} D_g \tilde{\Delta}_{wig} A_g D_g \beta_g \beta_g' \right\}
\]

\[
= C + \sum_{i=1}^{n} \text{tr} \left\{ \hat{z}_{ig} D_g (\tilde{\Delta}_{wig} A_g)^{-1} D_g' (x_i - \mu_g) (x_i - \mu_g)' \right\}
\]

\[
- \sum_{i=1}^{n} \text{tr} \left\{ \hat{z}_{ig} D_g \tilde{\Delta}_{wig} A_g D_g \beta_g \beta_g' \right\}
\]

\[
= C + \sum_{i=1}^{n} \text{tr} \left\{ \hat{z}_{ig} D_g (\tilde{\Delta}_{wig} A_g)^{-1} D_g' W_i \right\} - \sum_{i=1}^{n} \text{tr} \left\{ \hat{z}_{ig} D_g \tilde{\Delta}_{wig} A_g D_g B_i \right\}, \quad (6.16)
\]

with respect to \( D_g \), where \( B_i = \hat{z}_{ig} \beta_g \beta_g' \) and \( W_i = \hat{z}_{ig} (x_i - \mu_g) (x_i - \mu_g)' \).
Formally, on iteration \((t)\) of this procedure, we begin by computing

\[
F_1 = \sum_{i=1}^{n} \left[ (\Delta_{wig}A)^{-1} D_y^{(t)} W_i - \omega_{i1} A_g^{-1} D_g^{(t)} \right] - \sum_{i=1}^{n} \left[ (\Delta_{wig}A) D_y^{(t)} B_g - \omega_{i2} A_g D_g^{(t)} \right]
\]

(6.17)
given the current parameter estimates, where \(\omega_{i1}\) and \(\omega_{i2}\) are the largest eigenvalues of the matrices \(W_i\) and \(B_i\), respectively. Following this computation we calculate the elements of the singular value decomposition of \(F_1\), i.e., we set

\[
F_1 = P_1B_1R'_1
\]

and calculate \(P_1\), \(B_1\), and \(R'_1\) where \(P_1\) and \(R'_1\) are orthonormal and \(B_1\) is a diagonal matrix containing the singular values of \(F_1\). It follows that our initial estimate of \(D_g\) for iteration \((t + 1)\) is given by

\[
\hat{D}_g^{(t+1)*} = R_1P'_1.
\]

(6.18)

Given this estimate, denoted \(\hat{D}_g^{(t+1)*}\), we then compute

\[
F_2 = \sum_{i=1}^{n} \left[ W_iD_y^{(t+1)*} (\Delta_{wig}A)^{-1} - \alpha_{i1} W_iD_g^{(t+1)*} \right] - \sum_{i=1}^{n} \left[ B_iD_y^{(t+1)*} (\Delta_{wig}A) - \alpha_{i2} B_iD_g^{(t+1)*} \right],
\]

(6.19)

where \(\alpha_{i1}\) and \(\alpha_{i2}\) are the largest eigenvalues of \((\Delta_{wig}A)^{-1}\) and \((\Delta_{wig}A)\), respectively, and set \(F_2 = P_2B_2R'_2\). Therefore, our final estimate of \(D_g\) on iteration \((t + 1)\) is given by

\[
\hat{D}_g^{(t+1)} = R_2P'_2,
\]

where \(R_2\) and \(P'_2\) are previously defined.

We iterate over both the first and second MM algorithms until the difference in the objective
function, (6.16), at iteration \((t)\) and iteration \((t + 1)\) is found to be less than some small value.

Finally, to maximize \(Q(\mu_g, A_g, D_g, \beta_g)\) with respect to \(A_g\) we use

\[
\hat{A}_g = \text{diag} \left\{ \sqrt{\frac{\sum_{i=1}^{n} E_{2i1g} \hat{z}_{ig} v_{i1g}^2}{n_g/a_1^{(t-1)}} + \sum_{i=1}^{n} E_{1i1g} \hat{z}_{ig} \lambda_1^2}, \ldots, \sqrt{\frac{\sum_{i=1}^{n} E_{2ipg} \hat{z}_{ig} v_{ijg}^2}{n_g/a_p^{(t-1)}} + \sum_{i=1}^{n} E_{1ipg} \hat{z}_{ig} \lambda_j^2} \right\},
\]

(6.20)

where \(v_{ijg} = [D'_g(x - \mu_g)]_j\), \(\lambda_j\) is the \(j\)th element of the matrix \(A_g = D'_g \beta_g\) and all off-diagonal elements of \(\hat{A}_g\) are equal to zero (see Appendix B for details).

After preforming one E-step and one M-step we calculate the value of the log-likelihood. Our EM algorithm is considered to have converged when the difference between the asymptotic log-likelihood \(l^{(t+1)}\) and the likelihood value on iteration \((t)\), \(l^{(t)}\), are less than some small value \(\epsilon\) (McNicholas et al., 2010). At convergence we use the final estimates of the \(\hat{z}_{ig}\) to obtain the maximum a posteriori (MAP) classification values. Specifically, \(\text{MAP}\{\hat{z}_{ig}\} = 1\) if \(\max_h \{\hat{z}_{ih}\}\) occurs in component \(h = g\), and \(\text{MAP}\{\hat{z}_{ig}\} = 0\) otherwise.

### 6.5 Applications

#### 6.5.1 Introduction

We use two real data sets to illustrate the shapes and classification ability of our MSSAL mixtures. In both Section 6.5.2 and Section 6.5.3 we compare our mixtures of MSSAL distributions to a mixture of multivariate Gaussian distributions in model-based clustering scenarios. That is, we removed all group labels and fitted both mixtures for \(G = 1, \ldots, 5\)
groups and chose the best fitting model using the BIC. Furthermore, for both analysis we initialized the models using \( k \)-means starting values.

### 6.5.2 Leptograpsus Crabs

In our first analysis we reconsider the data created by comparing the first and third principal components of the Leptograpsus crabs data set (cf. Section 5.4.3). We present contour plots for a chosen \( G = 2 \) component MSSAL (\( \text{BIC} = -770.0312, ICL = -772.3863 \)) and \( G = 2 \) component GMM (\( \text{BIC} = 721.0288 \)) fitted to the principal components of the Crabs data (cf. Section 5.4.3). The contour plots in 6.2 show the unique shape of our multivariate generalization compared to the common spherical shapes given by a mixture of Gaussian distributions. Both our MSSAL (\( \text{ARI} = 0.98 \)) and the Gaussian mixtures (\( \text{ARI} = 1.00 \)) obtain excellent classification results however, the MSSAL distribution appears to capture the true shape of the data much better than the Gaussian distribution, despite the single misclassification.

### 6.5.3 Italian Wines Data

The Italian wines data was originally reported on by Forina et al. (1986). In total there are 178 wines in this data set and each wine belongs to one of three types: Barolo, Grignolino or Barbera. A complete list of the chemical and physical properties of each wine is given in Table 6.1.

A mixture of MSSAL and multivariate Gaussian distributions were fitted to this data. A 3 component mixture of MSSAL distributions (\( \text{BIC} = -5691.533, ICL = -5691.533 \)) was
Table 6.1: The thirteen chemical and physical properties of the Italian wines.

<table>
<thead>
<tr>
<th>Alcohol</th>
<th>Malic acid</th>
<th>Hue</th>
<th>Flavonoids</th>
</tr>
</thead>
<tbody>
<tr>
<td>Color Intensity</td>
<td>$OD_{280}/OD_{315}$ of diluted wines</td>
<td>Magnesium</td>
<td>Proline</td>
</tr>
<tr>
<td>Ash</td>
<td>Total phenols</td>
<td>Nonflavonoid phenols</td>
<td>Alcalinity of ash</td>
</tr>
<tr>
<td>Proanthocyanins</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 6.2: The first and third principal components of the crabs data with contours reprinting the fit of the chosen MSSAL and GMM models.
selected as the best fitting MSSAL model and a 2 component mixture of multivariate Gaussian distributions \((\text{BIC} = -5595.24)\) was also chosen. The classification results are given in Table 6.2 below and, clearly, the chosen 3 component mixture of MSSAL distributions \((\text{ARI} = 0.829)\) outperforms the chosen Gaussian mixture model \((\text{ARI} = 0.461)\).

Table 6.2: Classification results for the mixture of MSSAL distribution \((\text{ARI} = 0.829)\) and GMM \((\text{ARI} = 0.461)\) for the Italian wines data. The types of wine are cross-tabulated against our predicted classifications (A, B, C) in each case.

<table>
<thead>
<tr>
<th></th>
<th>MSSAL</th>
<th>GMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>57</td>
<td>59</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>66</td>
<td>5</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>48</td>
</tr>
</tbody>
</table>

### 6.6 Summary and Discussion

A multivariate generalization of a shifted asymmetric Laplace distribution, referred to as a multiple scaled shifted asymmetric Laplace distribution, is introduced. Contour plots illustrating the possible shapes are given and a parameter estimation scheme using an EM algorithm is explicitly described. We notice that both diamonds and triangular shapes are possible with this distribution; a far cry from any of the spherical or tear-drop like densities commonly displayed in the literature. Furthermore, our parameter estimation procedure is relatively straight-forward and a comparison with the unrestricted models given in Lee and McLachlan (2013) is the subject of future work.

Our MSSAL mixtures are compared to the general Gaussian mixture model using two real data sets: the Leptograpsus crabs and the 13-variable Italian wines data. The Leptograpsus crabs data demonstrated the flexibility and unique shapes that are possible using these
mixtures and the chosen model gave excellent classification performance compared to the Gaussian mixture model. Furthermore, when fit to the popular Italian wines data the chosen MSSAL mixture (ARI = 0.829) gave a superior classification performance compared to the chosen GMM (ARI = 0.461).

The fact that the levels of our MSSAL density are guaranteed to be convex (e.g., Figures 6.1 and 6.2) makes the mixture of MSSAL distributions ideal for clustering and classification applications. Specifically, the MSSAL distribution is guaranteed to have convex level sets, similar to a Gaussian and other elliptical distributions, because it has the same concentration in each direction from the mode. In contrast, the contours produced for the multiple scaled multivariate $t$-distribution with varying degrees of freedom in the tails, introduced by Forbes and Wraith (2013), will have levels that are not convex; therefore, situations will arise where one component is used to model two clusters, e.g., X-shaped components. This problem cannot arise with our mixture of MSSAL distributions and this a significant advantage of our approach in clustering and classification applications.
Chapter 7

Conclusion

7.1 Summary

This thesis has focused on the development of mixtures of shifted asymmetric Laplace distributions. In addition to the general mixture, two parsimonious families of models are developed and a multivariate generalization was presented that allows for the incorporation of a multidimensional weight function. This work is a significant contribution to the body of literature on using mixtures of skewed distributions for clustering.

7.1.1 Mixtures of shifted asymmetric Laplace distributions

A decade on from the landmark paper of Fraley and Raftery (2002), we have put forth a case for substantial departure from the Gaussian model-based clustering paradigm. Unlike the skew-normal and skew-$t$ approaches, which are perhaps less substantial departures, our ap-
proach is elegant and computationally straightforward. For our mixture of SAL distributions we used an EM algorithm in combination with deterministic annealing for parameter estimation. The ‘best’ fitting SAL mixture was selected using the integrated complete likelihood measure and we evaluated clustering performance using the ARI.

We demonstrated our mixtures in both model-based clustering and model-based classification scenarios using both simulated and real data. 2D and 3D contour plots are given using the Old Faithful Geyser Data and compared to a mixture of Gaussian distributions our SAL mixtures preform favourably on the Yeast Data.

7.1.2 Extensions

We developed two parsimonious families of shifted asymmetric Laplace distributions by decomposing the scale matrix and constraining the elements of each decomposition. We use the well-known eigen-decomposition to create a family of fourteen parsimonious SAL mixtures and extend the mixture of factor analyzers model to include the SAL density. The mixture of SAL factor analyzers has the favourable property that the number of free covariance parameters are all linear in $p$, making them better suited for high dimensional analysis.

We compare both families to their available symmetric analogues and find that they out-preform the opposition on the data sets considered. Notable applications were in clustering data from a flow cytometry experiment, a simulated ‘pizza’ data set and the Leptograpsus crabs data.
7.1.3 Multiple scaled shifted asymmetric Laplace distributions

In this chapter, a multivariate generalization of the SAL density was introduced. This density appears to have a rigid (diamond-like) shape and is characterized by an ability to parameterize skewness in linearly independent directions. An EM algorithm was presented for parameter estimation and encompassed within this procedure was a MM algorithm, used for estimating the common orthogonal matrix, \( D_g \). We used two real data sets to demonstrate the classification ability of our mixtures. For both data sets we found the best fitting model to give a very good classification performance, specifically in the case of the Italian wines data.

7.2 Future Work

7.2.1 Infinite Likelihoods

A better solution to the issue of infinite log-likelihood values in our EM algorithms is a subject of ongoing work. In this thesis we presented three simple ideas to help overcome this problem however, computationally efficient as they are, there is still much work to be done in this area. Potential avenues of exploration include the development of a parameter estimation scheme that restricts the value of the estimated location parameter to be from a specific set of values or introduce a measure that prevents the difference between \( \mathbf{x} \) and \( \mu_g \) from being 0. The second idea arises from studying univariate distributions like the double exponential or Pareto distributions whose densities have a similar form but, with an imposed protection.
7.2.2 Mixtures of Common SAL Factor Analyzers

There are still several extensions of our mixtures of SAL distributions that are possible. One promising area of exploration is to constrain the scale matrices of our SAL mixtures along the lines of work by Baek et al. (2010), who focused on extensions of mixtures of factor analyzers. Such an imposition could allow for the analysis of even higher dimensional data and have applications in fields like bioinformatics where the number of dimensions often exceeds the number of observations.

7.2.3 Model Selection

One issue that arises when using the BIC and ICL for model selection is that they can sometimes pick the model whose classification performance is not the best among the models fit. For example, consider the results of fitting the ParSAL family to the Swiss bank notes Data (Section 5.4.4). The best fitting model (BIC = −1885.69, ICL = −1908.613) is the 5-component VEV model however, six of the possible 2-component models have ARI’s of 0.98. One possible solution to this issue is to derive a new penalty term that penalizes the likelihood properly for skewness. This could be accomplished via an alteration to the prior densities on which the BIC was originally derived.
Bibliography


105


Appendix A

Calculations for the Mixture of SALs

A.1 The distribution of $W \mid X = x$

To find the distribution of the random variable $W \mid X = x$ we use Bayes rule

$$f_x(X \mid Y = y) = \frac{f_{x,y}(x,y)}{f_Y(y)} = \frac{f_Y(y \mid X = x)f_X(x)}{f_Y(y)}.$$  

Therefore, for our mixture of SAL distributions we have

$$f_W(W \mid X = x) = \frac{f_{w,x}(x,w)}{f_X(x)} = \frac{f_X(x \mid W = w)h(w)}{f_X(x)}.$$  

where $f_X(x \mid W = w) \sim MVN(u + w\alpha, w\Sigma)$, $h(w) = e^{-w}$ and $f_X(x) \sim SAL(\alpha, \Sigma, \mu)$. This
which is the GIG density with $a \equiv 2 + \alpha' \Sigma^{-1} \alpha$ and $b \equiv \delta(x, \mu | \Sigma)$. 

\[f_W(W | X = x) = \frac{1}{(2\pi)^{p/2} \alpha^{1/2}} \frac{1}{(2\pi)^{p/2} \alpha^{1/2}} e^{\left(-\frac{1}{2}(x - \mu - w \alpha)^t \Sigma^{-1} (x - \mu - w \alpha)\right) e^{-w}}\]

\[= \frac{1}{w^{p/2}} \left(\frac{\delta(x, \alpha | \Sigma)}{2 + \alpha' \Sigma^{-1} \alpha}\right)^{-\nu/2} e^{\left(-\frac{1}{2}(x - \mu - w \alpha)^t \Sigma^{-1} (x - \mu - w \alpha)\right) e^{-w}}\]

\[= \frac{1}{2w^{p/2}} \left(\frac{\delta(x, \mu | \Sigma)}{2 + \alpha' \Sigma^{-1} \alpha}\right)^{-\nu/2} e^{\left(-\frac{1}{2}(x - \mu - w \alpha)^t \Sigma^{-1} (x - \mu - w \alpha)\right) e^{-w}}\]
A.2 Maximum Likelihood Estimates

For a mixture of shifted Asymmetric Laplace distributions the expected value of the complete-data log-likelihood is given by

$$Q(\mu_g, \Sigma_g, \alpha_g) = \sum_{g=1}^{G} n_g \log \pi_g - \frac{np}{2} \log 2\pi + \sum_{g=1}^{G} \frac{n_g}{2} \log |\Sigma_g| - \frac{p}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} E_{3ig}$$

$$- \frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} E_{2ig} (x_i - \mu_g)' \Sigma_g^{-1} (x_i - \mu_g) + \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} \alpha_g' \Sigma_g^{-1} (x_i - \mu_g)$$

$$- \frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} E_{1ig} \alpha_g' \Sigma_g^{-1} \alpha_g - \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} E_{1ig},$$

where $n_g = \sum_{i=1}^{n} z_{ig}$, $E[ W_{ig} | x_i] := E_{1ig}$, $E[1/W_{ig} | x_i, z_{ig} = 1] := E_{2ig}$ and

$$E[ \log W_{ig} | x_i, z_{ig} = 1 ] := E_{3ig}.$$

Differentiating $Q(\mu_g, \Sigma_g, \alpha_g)$ with respect to find $\mu_g$ and $\alpha_g$ gives the score functions below.

$$S_1(\mu_g, \Sigma_g, \alpha_g) = \frac{\partial Q(\mu_g, \Sigma_g, \alpha_g)}{\partial \mu_g} = \sum_{i=1}^{n} \hat{z}_{ig} (x_i - \mu_g)' \Sigma_g^{-1} - n_g \alpha_g' \Sigma_g^{-1}.$$

$$S_2(\mu_g, \Sigma_g, \alpha_g) = \frac{\partial Q(\mu_g, \Sigma_g, \alpha_g)}{\partial \alpha_g} = \sum_{i=1}^{n} \hat{z}_{ig} (x_i - \mu_g)' \Sigma_g^{-1} - \sum_{i=1}^{n} \hat{z}_{ig} E_{1ig} \alpha_g' \Sigma_g^{-1}.$$

Now, solving $S_1(\hat{\mu}_g^{\text{new}}, \Sigma_g, \alpha_g) = 0$ gives

$$\hat{\mu}_g^{\text{new}} = \frac{\sum_{i=1}^{n} \hat{z}_{ig} E_{2ig} x_i - n_g \alpha_g}{\sum_{i=1}^{n} \hat{z}_{ig} E_{2ig}}, \quad (A.1)$$
and solving $S_2(\hat{\mu}_g^{\text{new}}, \Sigma_g, \hat{\alpha}_g^{\text{new}}) = 0$ gives

$$
\hat{\alpha}_g^{\text{new}} = \frac{\sum_{i=1}^{n} \hat{z}_{ig} (x_i - \hat{\mu}_g^{\text{new}})}{\sum_{i=1}^{n} \hat{z}_{ig} E_{2ig}}. \tag{A.2}
$$

Now, to find the non-conditional MLE for $\mu_g$ we need to take the update in (A.2) and sub it into (A.1). Note: Herein we let $A_{ig} = \hat{z}_{ig} E_{1ig}$ and $B_{ig} = \hat{z}_{ig} E_{2ig}$. Subbing (A.2) into (A.1) gives

$$
\hat{\mu}_g^{\text{new}} = \frac{\sum_{i=1}^{n} B_{ig} x_i - \sum_{i=1}^{n} \hat{z}_{ig} \sum_{i=1}^{n} (x_i - \hat{\mu}_g^{\text{new}})}{\sum_{i=1}^{n} B_{ig}}
$$

$$
\hat{\mu}_g^{\text{new}} \sum_{i=1}^{n} B_{ig} = \sum_{i=1}^{n} B_{ig} x_i - \sum_{i=1}^{n} \hat{z}_{ig} \sum_{i=1}^{n} (x_i - \hat{\mu}_g^{\text{new}})
$$

$$
\hat{\mu}_g^{\text{new}} \sum_{i=1}^{n} A_{ig} \sum_{i=1}^{n} B_{ig} = \sum_{i=1}^{n} A_{ig} \sum_{i=1}^{n} B_{ig} x_i - n_g \sum_{i=1}^{n} \hat{z}_{ig} \sum_{i=1}^{n} (x_i - \hat{\mu}_g^{\text{new}})
$$

$$
\hat{\mu}_g^{\text{new}} \sum_{i=1}^{n} A_{ig} \sum_{i=1}^{n} B_{ig} - n_g^{2} \hat{\mu}_g^{\text{new}} = \sum_{i=1}^{n} A_{ig} \sum_{i=1}^{n} B_{ig} x_i - n_g \sum_{i=1}^{n} \hat{z}_{ig} x_i,
$$

therefore

$$
\hat{\mu}_g^{\text{new}} = \frac{\sum_{i=1}^{n} A_{ig} \sum_{i=1}^{n} B_{ig} x_i - n_g \sum_{i=1}^{n} \hat{z}_{ig} x_i}{\sum_{i=1}^{n} A_{ig} \sum_{i=1}^{n} B_{ig} - n_g^{2}}.
$$

Similarly, to find the non-conditional MLE for $\alpha_g$ we need to take the update in (A.1) and
sub it into (A.2). This gives

\[ \hat{\alpha}_{g}^{\text{new}} = \frac{\sum_{i=1}^{n} \hat{z}_{ig} x_i - \sum_{i=1}^{n} \hat{z}_{ig} \frac{\sum_{i=1}^{n} B_{ig} x_i - n_g \hat{\alpha}_{g}^{\text{new}}}{\sum_{i=1}^{n} B_{ig}}}{\sum_{i=1}^{n} A_{ig}} \]

\[ \hat{\alpha}_{g}^{\text{new}} = \frac{\sum_{i=1}^{n} B_{ig} \sum_{i=1}^{n} \hat{z}_{ig} x_i - n_g \sum_{i=1}^{n} B_{ig} x_i + n_g^2 \hat{\alpha}_{g}^{\text{new}}}{\sum_{i=1}^{n} A_{ig} \sum_{i=1}^{n} B_{ig}} \]

\[ \sum_{i=1}^{n} A_{ig} \sum_{i=1}^{n} B_{ig} \hat{\alpha}_{g}^{\text{new}} - n_g^2 \hat{\alpha}_{g}^{\text{new}} = \sum_{i=1}^{n} B_{ig} \sum_{i=1}^{n} \hat{z}_{ig} x_i - n_g \sum_{i=1}^{n} B_{ig} x_i \]

\[ \hat{\alpha}_{g}^{\text{new}} \left( \sum_{i=1}^{n} A_{ig} \sum_{i=1}^{n} B_{ig} - n_g^2 \right) = \sum_{i=1}^{n} B_{ig} \sum_{i=1}^{n} \hat{z}_{ig} x_i - n_g \sum_{i=1}^{n} B_{ig} x_i, \]

therefore

\[ \hat{\alpha}_{g}^{\text{new}} = \frac{\sum_{i=1}^{n} B_{ig} \sum_{i=1}^{n} \hat{z}_{ig} x_i - n_g \sum_{i=1}^{n} B_{ig} x_i}{\sum_{i=1}^{n} A_{ig} \sum_{i=1}^{n} B_{ig} - n_g^2} \]

Finally, Differentiating \( Q(\mu_g, \Sigma_g, \alpha_g) \) with respect to find \( \Sigma_g^{-1} \) gives the score functions below.

\[ S_1(\mu_g, \Sigma_g, \alpha_g) = \frac{\partial Q(\mu_g, \Sigma_g, \alpha_g)}{\partial \Sigma_g^{-1}} = \frac{n_g}{2} \Sigma_g - \frac{1}{2} \sum_{i=1}^{n} B_{ig} \left( x_i - \mu_g \right) \left( x_i - \mu_g \right)' + \sum_{i=1}^{n} \hat{z}_{ig} \alpha_g \left( x_i - \mu_g \right)' - \frac{1}{2} \sum_{i=1}^{n} A_{ig} \alpha_g \alpha_g' \]

Solving \( S_1(\mu_g^{\text{new}}, \Sigma_g^{\text{new}}, \alpha_g^{\text{new}}) = 0 \) gives

\[ \Sigma_g^{\text{new}} = \frac{1}{n_g} \sum_{i=1}^{n} B_{ig} \left( x_i - \hat{\mu}_g^{\text{new}} \right) \left( x_i - \hat{\mu}_g^{\text{new}} \right)' \]

\[ - \frac{2}{n_g} \hat{\alpha}_g^{\text{new}} \sum_{i=1}^{n} \hat{z}_{ig} \left( x_i - \hat{\mu}_g^{\text{new}} \right)' + \frac{1}{n_g} \sum_{i=1}^{n} A_{ig} \hat{\alpha}_g^{\text{new}} \hat{\alpha}_g^{\text{new}}' \quad (A.3) \]
Appendix B

Calculations for the MSSAL

B.1 Maximum Likelihood Estimates

For a mixture of multiple scaled shifted asymmetric Laplace distributions the complete-data log-likelihood is given by

\[ Q(\mu_g, A_g, D_g, \beta_g) = C + \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} \log |A_g| \]

\[ - \frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} (x_i - \mu_g - \Omega_{ig} \beta_g)^\prime D_g \Delta^{-1} \omega_{ig} A_g^{-1} D_g^\prime (x_i - \mu_g - \Omega_{ig} \beta_g) \]  

(B.1)

where \( C \) is constant of the model parameters: \( \mu_g, \beta_g, A_g \) and \( D_g, \Omega_{ig} = D_g \Delta_{wig} A_g D_g^\prime \) such that
\[ \tilde{\Delta}_{wig} = \begin{pmatrix} \mathbb{E}[W_{1ig} | x, z_{ig} = 1] & 0 & 0 & 0 \\ 0 & \mathbb{E}[W_{2ig} | x, z_{ig} = 1] & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \mathbb{E}[W_{ipg} | x, z_{ig} = 1] \end{pmatrix} \]

and

\[ \tilde{\Delta}_{wig}^{-1} = \begin{pmatrix} \mathbb{E}[W_{1ig}^{-1} | x, z_{ig} = 1] & 0 & 0 & 0 \\ 0 & \mathbb{E}[W_{2ig}^{-1} | x, z_{ig} = 1] & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \mathbb{E}[W_{ipg}^{-1} | x, z_{ig} = 1] \end{pmatrix} \]

We obtain score functions for the relevant parameters by differentiating \( Q(\mu_g, A_g, D_g, \beta_g) \) with respect to \( \mu_g \) and \( \beta_g \). Note: Herein we let \( E_{1ijg} = \mathbb{E}[W_{ijg} | x, z_{ig} = 1] \) and \( E_{2ijg} = \mathbb{E}[W_{ijg}^{-1} | x, z_{ig} = 1] \).

Specifically,

\[
S_1(\mu_g, A_g, D_g, \beta_g) = \frac{\partial Q(\mu_g, A_g, D_g, \beta_g)}{\partial \mu_g} = \sum_{i=1}^{n} \hat{z}_{ig} \Omega_{ig} (x_i - \mu_g)' D_g \Delta_{wig}^{-1} A_g^{-1} D'_g \\
- \sum_{i=1}^{n} \hat{z}_{ig} \beta'_g
\]

\[
S_2(\mu_g, A_g, D_g, \beta_g) = \frac{\partial Q(\mu_g, A_g, D_g, \beta_g)}{\partial \alpha_g} = \sum_{i=1}^{n} \hat{z}_{ig} (x_i - \mu_g)' - \sum_{i=1}^{n} \hat{z}_{ig} \beta'_g \Omega_{ig}
\]
Solving $S_1(\hat{\mu}^\text{new}_g, A_g, D_g, \beta_g) = 0$ gives

$$\hat{\mu}^\text{new}_g = \left( \sum_{i=1}^{n} \hat{z}_{ig} \Omega_{ig}^{-1} \right)^{-1} \sum_{i=1}^{n} \hat{z}_{ig} \Omega_{ig}^{-1} x_i - n_g \beta_g$$

and solving $S_2(\hat{\mu}^\text{new}_g, A_g, D_g, \hat{\beta}^\text{new}_g) = 0$ gives

$$\hat{\beta}^\text{new}_g = \left( \sum_{i=1}^{n} \hat{z}_{ig} \Omega_{ig}^{-1} \right)^{-1} \sum_{i=1}^{n} \hat{z}_{ig} x_i - n_g \hat{\mu}^\text{new}_g$$

To maximize $Q(\mu_g, A_g, D_g, \beta_g)$ with respect to $A_g$ we derive the update as follows. Simplifying $Q(\mu_g, A_g, D_g, \beta_g)$ gives

$$Q(\mu_g, A_g, D_g, \beta_g) = C + \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} \log |A_g|$$

$$- \frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} (x_i - \mu_g - \Omega_{ig} \beta_g)' D_g \Delta_{wig}^{-1} D_g (x_i - \mu_g - \Omega_{ig} \beta_g)$$

$$= C - n \sum_{j=1}^{p} \log A_{jg} - \sum_{i=1}^{n} \text{tr}\left\{ \hat{z}_{ig} D_g (\Delta_{ig} A_g)^{-1} D_g' (x_i - \mu_g) (x_i - \mu_g)' \right\}$$

$$- \sum_{i=1}^{n} \text{tr}\left\{ \hat{z}_{ig} D_g \Delta_{ig} A_g D_g' \beta_g \beta_g' \right\},$$

where $A_{jg}$ is the $j$th diagonal element of the matrix $A_g$ and letting $v_{ig} = D_g' (x_i - \mu_g)$ and $\lambda_g = D_g' \beta_g$ gives

$$Q(\mu_g, A_g, D_g, \beta_g) = C - \sum_{j=1}^{p} \log A_{jg} - \sum_{i=1}^{n} \text{tr}\left\{ \hat{z}_{ig} (\Delta_{ig} A_g)^{-1} v_{ig} v_{ig}' \right\}$$

$$- \sum_{i=1}^{n} \text{tr}\left\{ \hat{z}_{ig} \Delta_{ig} A_g \lambda_g \lambda_g' \right\}.$$
Since $\mathbf{v}_g = (v_{1g}, \ldots, v_{pg})$, $\lambda = (\lambda_1, \ldots, \lambda_p)$ and $\mathbf{A}_g = \text{diag}(A_{1g}, \ldots, A_{pg})$ we can write

$$l(\mathbf{A}_g) = -n \sum_{j=1}^{p} \log A_{jg} - \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{1}{A_{jg}} E_{2ijg} \hat{z}_{ig} v_{ijg}^2 - \sum_{i=1}^{p} \sum_{j=1}^{p} A_{jg} E_{1ijg} \hat{z}_{ig} \lambda_{jg}^2.$$  \hspace{1cm} (B.2)

Recall, if $\log(x)$ is a concave function then for all $x$ and $x_0$

$$\log x \leq \log x_0 + \frac{1}{x_0} (x - x_0),$$

and

$$-\log x \geq -\log x_0 - \frac{1}{x_0} (x - x_0),$$

therefore

$$-l(\mathbf{A}_g) \geq g(A_g, A^{(t)}_g) = -n \sum_{j=1}^{p} \log A_{jg} - n \sum_{j=1}^{p} \frac{A_{jg} - A^{(t)}_{jg}}{A_{jg}} - \sum_{j=1}^{p} \frac{1}{A_{jg}} \left( \sum_{i=1}^{n} E_{2ijg} \hat{z}_{ig} v_{ijg}^2 \right)$$

$$- \sum_{j=1}^{p} A_{jg} \left( \sum_{i=1}^{n} E_{1ijg} \hat{z}_{ig} \lambda_{jg}^2 \right).$$

Taking the derivative of with respect to each diagonal element of $\mathbf{A}_g$ gives the score function

$$S_3(\mu_g, \mathbf{A}_g, \mathbf{D}_g, \beta_g) = \frac{\partial g(A_g, A^{(t)}_g)}{\partial A_g} = -n \sum_{j=1}^{p} \frac{A_{jg} - A^{(t)}_{jg}}{A_{jg}} + \frac{1}{A_{jg}^2} \left( \sum_{i=1}^{n} E_{2ijg} \hat{z}_{ig} v_{ijg}^2 \right) - \left( \sum_{i=1}^{n} E_{1ijg} \hat{z}_{ig} \lambda_{jg}^2 \right).$$

Solving $S_2(\mu_{g_{\text{new}}}, \hat{A}_{g_{\text{new}}}, \hat{D}_{g_{\text{new}}}, \beta_{g_{\text{new}}}) = 0$ gives

$$\hat{A}_g = \text{diag} \left\{ \sqrt{\frac{\sum_{i=1}^{n} E_{2ig} \hat{z}_{ig} v_{1ig}^2}{\sum_{i=1}^{n} E_{1ig} \hat{z}_{ig} \lambda_{1g}^2}}, \ldots, \sqrt{\frac{\sum_{i=1}^{n} E_{2igt} \hat{z}_{igt} v_{igt}^2}{\sum_{i=1}^{n} E_{1igt} \hat{z}_{igt} \lambda_{tg}^2}} \right\}.$$  \hspace{1cm} (B.3)