Variational Approximations and Other Topics
in Mixture Models

by
Sanjeena Dang

A Thesis
Presented to
The University of Guelph

In partial fulfilment of requirements
for the degree of
Doctor of Philosophy
in
Mathematics and Statistics
(Applied Statistics)

Guelph, Ontario, Canada

© Sanjeena Dang, July, 2012
Mixture model-based clustering has become an increasingly popular data analysis technique since its introduction almost fifty years ago. Families of mixture models are said to arise when the component parameters, usually the component covariance matrices, are decomposed and a number of constraints are imposed. Within the family setting, it is necessary to choose the member of the family — i.e., the appropriate covariance structure — in addition to the number of mixture components. To date, the Bayesian information criterion (BIC) has proved most effective for this model selection process, and the expectation-maximization (EM) algorithm has been predominantly used for parameter estimation.

We deviate from the EM-BIC rubric, using variational Bayes approximations for parameter estimation and the deviance information criterion (DIC) for model selection. The variational Bayes approach alleviates some of the computational complexities associated with the EM algorithm. We use this approach on the most famous family of Gaussian mixture models known as Gaussian parsimonious clustering mod-
els (GPCM). These models have an eigen-decomposed covariance structure.

Cluster-weighted modelling (CWM) is another flexible statistical framework for modelling local relationships in heterogeneous populations on the basis of weighted combinations of local models. In particular, we extend cluster-weighted models to include an underlying latent factor structure of the independent variable, resulting in a novel family of models known as parsimonious cluster-weighted factor analyzers. The EM-BIC rubric is utilized for parameter estimation and model selection.

Some work on a mixture of multivariate t-distributions is also presented, with a linear model for the mean and a modified Cholesky-decomposed covariance structure leading to a novel family of mixture models. In addition to model-based clustering, these models are also used for model-based classification, i.e., semi-supervised clustering. Parameters are estimated using the EM algorithm and another approach to model selection other than the BIC is also considered.
To my husband Utkarsh

With love
Acknowledgments

First of all, I would like to express my most sincere gratitude to my advisor Dr. Paul D. McNicholas for his guidance, dedication and support during my study and research at the University of Guelph. I’m also thankful for all the opportunities for self development through conferences and workshops that he always encouraged. His dedication towards his students and research is truly an inspiration.

I’m also grateful for having a very friendly learning environment in the University of Guelph and the endless support and motivations provided by my office mates and friends. I’m also very grateful for the Postgraduate Scholarship from the Natural Science and Engineering Research Council of Canada.

I would also like to express my thanks to Dr. Salvatore Ingrassia and Dr. Antonio Punzo from University of Catania and my advisor Dr. Paul McNicholas for providing me with the opportunity for collaborative research in University of Catania, Italy. I would also like to thank my examining committee Dr. Charles Bouveyron, Dr. Paul D. McNicholas, Dr. Rob Deardon, Dr. Ryan Browne and Dr. Zeny Feng for their time.

Furthermore, I would like to express my gratitude towards my family and friends whose unconditional love and support made it possible for me. Last but not least, I am greatly indebted to my husband, Utkarsh Dang for his unconditional love and encouragement at all times. It is so wonderful to have you beside me, now and forever. Thank you for always believing in me and for being such an amazing husband.
Publications

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


# Table of Contents

List of Tables \hspace{1.5cm} x

List of Figures \hspace{1.5cm} xii

1 Introduction \hspace{1.5cm} 1
  1.1 Overview \hspace{1.5cm} 1
  1.2 Model-based approaches \hspace{1.5cm} 1
  1.3 Finite Mixture Models \hspace{1.5cm} 2
  1.4 Cluster-Weighted Models \hspace{1.5cm} 3
  1.5 Family of Models \hspace{1.5cm} 3
    1.5.1 Gaussian Parsimonious Clustering Models (GPCM) \hspace{1.5cm} 4
    1.5.2 Mixtures of Factor Analysers (MFA)/ Parsimonious Gaussian
        Mixture Models (PGMM) \hspace{1.5cm} 4
    1.5.3 Modified Cholesky-decomposed Gaussian Mixture Models \hspace{1.5cm} 5
  1.6 Overview of the thesis \hspace{1.5cm} 5

2 Parameter Estimation and Model Selection \hspace{1.5cm} 7
  2.1 Parameter Estimation \hspace{1.5cm} 7
    2.1.1 EM-algorithm \hspace{1.5cm} 8
    2.1.2 Variational Bayes Approach \hspace{1.5cm} 8
    2.1.3 Other Approaches \hspace{1.5cm} 10
  2.2 Model selection \hspace{1.5cm} 11
    2.2.1 Akaike Information Criterion (AIC) \hspace{1.5cm} 11
    2.2.2 Bayesian Information Criterion (BIC) \hspace{1.5cm} 12
    2.2.3 Integrated Completed Likelihood (ICL) \hspace{1.5cm} 12
    2.2.4 Deviance Information Criterion (DIC) \hspace{1.5cm} 13

3 Parameter Estimation and Mixture Model Selection for a Family of
  Gaussian Parsimonious Clustering models using Variational Bayes
  Approximations \hspace{1.5cm} 14
  3.1 Methodology \hspace{1.5cm} 14
    3.1.1 Introducing Parsimony \hspace{1.5cm} 14
    3.1.2 Priors and Approximating Densities \hspace{1.5cm} 16
    3.1.3 Convergence \hspace{1.5cm} 19
    3.1.4 Model Selection \hspace{1.5cm} 21
    3.1.5 Performance Assessment \hspace{1.5cm} 21
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1.6</td>
<td>Model-Based Classification</td>
<td>22</td>
</tr>
<tr>
<td>3.2</td>
<td>Results</td>
<td>23</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Simulation Study 1</td>
<td>23</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Simulation Study 2</td>
<td>24</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Clustering of Leptograpsus Crabs Data</td>
<td>25</td>
</tr>
<tr>
<td>3.2.4</td>
<td>Classification of Olive Oil Data</td>
<td>26</td>
</tr>
<tr>
<td>3.3</td>
<td>Discussion</td>
<td>27</td>
</tr>
<tr>
<td>4</td>
<td>A family of Cluster-Weighted Factor Analyzers</td>
<td>37</td>
</tr>
<tr>
<td>4.1</td>
<td>Methodology</td>
<td>37</td>
</tr>
<tr>
<td>4.1.1</td>
<td>Cluster-Weighted Modeling</td>
<td>38</td>
</tr>
<tr>
<td>4.1.2</td>
<td>The Factor Regression Model</td>
<td>39</td>
</tr>
<tr>
<td>4.1.3</td>
<td>The Gaussian Cluster-Weighted Factor Analyzer</td>
<td>41</td>
</tr>
<tr>
<td>4.1.4</td>
<td>The Likelihood Function and the EM algorithm</td>
<td>42</td>
</tr>
<tr>
<td>4.1.5</td>
<td>A Set of Parsimonious Covariance Structures</td>
<td>43</td>
</tr>
<tr>
<td>4.1.6</td>
<td>Semi-supervised Classification via CWFA models</td>
<td>50</td>
</tr>
<tr>
<td>4.1.7</td>
<td>Initialization</td>
<td>51</td>
</tr>
<tr>
<td>4.1.8</td>
<td>Convergence Criterion and Model Selection</td>
<td>53</td>
</tr>
<tr>
<td>4.2</td>
<td>Data Analyses</td>
<td>55</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Simulated data</td>
<td>55</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Real data</td>
<td>60</td>
</tr>
<tr>
<td>4.3</td>
<td>Discussion</td>
<td>66</td>
</tr>
<tr>
<td>5</td>
<td>Clustering Gene Expression Time Course Data Using Mixtures of Multivariate $t$-Distributions</td>
<td>67</td>
</tr>
<tr>
<td>5.1</td>
<td>Background</td>
<td>67</td>
</tr>
<tr>
<td>5.2</td>
<td>Methodology</td>
<td>69</td>
</tr>
<tr>
<td>5.2.1</td>
<td>The multivariate $t$-distribution, the correlation structure, and the mean</td>
<td>69</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Model-based clustering</td>
<td>70</td>
</tr>
<tr>
<td>5.2.3</td>
<td>Parameter estimation</td>
<td>72</td>
</tr>
<tr>
<td>5.2.4</td>
<td>Model selection</td>
<td>75</td>
</tr>
<tr>
<td>5.2.5</td>
<td>Model-based classification</td>
<td>76</td>
</tr>
<tr>
<td>5.3</td>
<td>Data analyses</td>
<td>77</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Simulation Studies</td>
<td>77</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Yeast sporulation time course data</td>
<td>81</td>
</tr>
<tr>
<td>5.3.3</td>
<td>Bronchial HAHRma data</td>
<td>89</td>
</tr>
<tr>
<td>5.4</td>
<td>Discussion</td>
<td>91</td>
</tr>
<tr>
<td>6</td>
<td>Conclusion</td>
<td>94</td>
</tr>
<tr>
<td>6.1</td>
<td>Variational Bayes approach for Gaussian parsimonious clustering models (GPCM)</td>
<td>94</td>
</tr>
<tr>
<td>6.1.1</td>
<td>Summary</td>
<td>94</td>
</tr>
<tr>
<td>6.1.2</td>
<td>Future work</td>
<td>95</td>
</tr>
<tr>
<td>6.2</td>
<td>Cluster-Weighted Factor Analysers</td>
<td>96</td>
</tr>
</tbody>
</table>
List of Tables

3.1 Geometric interpretation of the eigen-decomposition of the covariance matrix where I refers to identity matrix and E refers to equal, V refers to variable and S refers to Spherical across groups. ........................................... 15
3.2 Prior distributions for the parameters of the eigen-decomposed covariance structures. ................................................................. 33
3.3 Summary of the variational Bayes analysis of the two-dimensional simulated data using 10 different random starts. ......................... 34
3.4 Estimated parameters along with true parameters of the two-dimensional simulated data for one run. ........................................... 34
3.5 Summary of the variational Bayes and \texttt{mclust} analysis of the two-dimensional simulated data using \texttt{hclass} starting values. .......... 34
3.6 Summary of variational Bayes analysis of the three-dimensional simulated data using 10 different random starts. ......................... 35
3.7 Estimated and true mean parameters for the EEE model of the three-dimensional simulated data. .................................................. 35
3.8 Summary of variational Bayes analysis of the three-dimensional simulated data using classification from hierarchical clustering as the starting values. .......................................................... 35
3.9 Summary of the variational Bayes analysis of the principal components of the Leptograpsus crab data using 10 different random starts. The ARI was computed using species. .......................................... 36
3.10 Classification of the principal components of the Leptograpsus crabs data using the VVV and EVV models. ........................................ 36
3.11 Summary of the variational Bayes analysis of the olive oil data using 10 different random starts. .................................................. 36
4.1 Covariance structure for the $X$ variable where U refers to unconstrained and C refers to constrained. ............................................. 50
4.2 Estimated parameters along with true parameters .......................... 58
4.3 Estimated parameters along with true parameters .......................... 60
4.4 Classification of \texttt{f.voles} data using different clustering approaches. 62
4.5 Classification of \texttt{f.voles} data assuming 50\% of the observations have known group membership. ............................................... 66
5.1 The nomenclature, covariance structure, and number of covariance parameters for the models used by McNicholas and Murphy (2010a). . 71
5.2 True and estimated values for $m_{\beta_1}, m_{\beta_2}, D_1, D_2, \nu_1$, and $\nu_2$ from the first simulation study. ................................................................. 78
5.3 True and estimated values for $m_{\beta_1}, m_{\beta_2}, D_1, D_2, \nu_1$, and $\nu_2$ from the second simulation study. ................................................................. 79
5.4 True and estimated values for $m_{\beta_1}, m_{\beta_2}, D_1, D_2, \nu_1$, and $\nu_2$ from the third simulation study using mixtures of $t$-distributions. ................. 80
5.5 True classifications (1,2) tabulated against predicted classifications (A,B,C) resulting from the Gaussian mixture model analysis of the data from the third simulation study. ........................................ 80
5.6 True and estimated values of $m_{\beta_1}, m_{\beta_2}, D_1$, and $D_2$ from the Gaussian mixture models applied to data from the third simulation study. 81
5.7 Maximum a posteriori classifications for our best model (1–5) cross-tabulated against those from McNicholas and Murphy (2010a). .............. 86
5.8 Maximum a posteriori classifications for our best model (4–5) cross-tabulated against the model expression profiles from Chu et al. (1998). 88
A.1 Posterior expected value of the parameters of the eigen-decomposed covariance matrix ................................................................. 113
A.2 Posterior expected value of the parameters of the eigen-decomposed covariance matrix ................................................................. 134
List of Figures

3.1 The three component simulated data (left) along with the best classification given by the random start selected by the DIC for the EII model (right). .................................................. 23
3.2 Scatter plot matrix showing the relationships among the variables of Leptograpsus crab data. ............................................................. 28
3.3 Scatter plot matrix showing the relationships among the uncorrelated principal components. ......................................................... 29
4.1 Relationships among the models in the 5-step random-hierarchical initialization procedure. Arrows are oriented from the model used to initialize to the model to be estimated. ................................. 53
4.2 Example ???: scatterplot matrix of data. ........................................ 56
4.3 Example ???: group-factor plot of BIC values ordered in an increasing order. Numbers denote the estimated number of groups and colors denote the number of latent factors (red: $q = 1$, blue: $q = 2$). The green point denotes the ‘true’ model. ................................................................. 57
4.4 Example ???: scatterplot matrix of data. ........................................ 59
4.5 Example ???: group-factor plot of BIC values ordered in increasing order. Numbers denote the estimated number of groups and colors denote the number of latent factor (red: $q = 1$, blue: $q = 2$). The green point denotes the “true” model. ................................................................. 61
4.6 f.voles data: group-factor plot of BIC values ordered in an increasing order. Numbers denote the estimated number of groups and colors denote the number of latent factor (red: $q = 1$, blue: $q = 2$, green: $q = 3$) ................................................................. 63
4.7 Scatterplot matrix of f.voles data showing the classification observed using CWFA using clustering framework where black and red indicates sub-cluster of M. ochrogaster species and green indicates Microtus californicus species. ................................................................. 64
5.1 The 1,000 expression profiles from the third simulation; the green lines represent the members of component C under the chosen Gaussian model. ................................................................. 82
5.2 The forty model expression profiles of Chu et al. (1998), viewed over all seven time points: 38 genes that met the selection criterion in grey and the other two genes in black. ................................................................. 84
5.3 The forty model expression profiles of Chu et al. (1998), adjusted to six time points: 38 genes that met the selection criterion in grey and the other two genes in black.

5.4 The temporal patterns for the five clusters selected by our best model on the yeast data.

5.5 The temporal patterns for the three clusters selected by our best model on the HAHrma data.

A.1 The temporal patterns for the four clusters selected by our best model with constrained degrees of freedom on the yeast data.

A.2 Two apparently distinct temporal patterns for component 3 of our best model with constrained degrees of freedom on the yeast data.
Chapter 1

Introduction

1.1 Overview

This thesis is an exposition of three recent pieces of work on mixture model-based clustering and classification. Cluster analyses are used in virtually every area of scientific endeavour to select a number of similar, or homogenous, sub-populations from a given population. Identification and analysis of such sub-populations in mainstream scientific applications is well established (e.g., Sneath, 1957; Sokal and Michener, 1958). Most early clustering algorithms were based on heuristic approaches and some such methods, including hierarchical agglomerative clustering and $k$-means clustering (MacQueen, 1967; Hartigan and Wong, 1979; McLachlan and Peel, 2000a), are still widely used.

1.2 Model-based approaches

The use of mixture models to account for population heterogeneity has been very well established for over a century (e.g., Pearson, 1893), but only in the 1960s were they applied to clustering (Wolfe, 1963). Due to the lack of suitable computing
equipment, it was even later before the use of mixture models (Banfield and Raftery, 1993; Celeux and Govaert, 1995) and, more generally, the use of probability models (Bock, 1996, 1998a,b) for clustering took off. Since the turn of the century, the use of mixture models for clustering has burgeoned into a popular subfield of cluster analysis (recent examples include McLachlan and Peel, 2000b; Fraley and Raftery, 2002; McLachlan et al., 2003; Raftery and Dean, 2006; McLachlan et al., 2007; Bouveyron et al., 2007; Lin et al., 2007; Gormley and Murphy, 2008; McNicholas and Murphy, 2008, 2010a; Scrucca, 2010; Greselin and Ingrassia, 2010; Andrews and McNicholas, 2011a; Browne et al., 2012; McNicholas and Subedi, 2012).

1.3 Finite Mixture Models

A $d$-dimensional random vector $Y$ is said to arise from a parametric finite mixture distribution if, for all $y \subset Y$, we can write its density as $f(y \mid \vartheta) = \sum_{g=1}^{G} \rho_{g} p_{g}(y \mid \theta_{g})$, where $\rho_{g} > 0$ such that $\sum_{i=1}^{G} \rho_{g} = 1$ are the mixing proportions, $p_{g}(y \mid \theta_{g})$ are component densities, and $\vartheta = (\rho_{1}, \ldots, \rho_{G}, \theta_{1}, \ldots, \theta_{G})$ is the vector of parameters. Typically, each component probability density represents a cluster and is of the same type. For instance, when these densities are Gaussian, the density function is

$$f(y \mid \vartheta) = \sum_{g=1}^{G} \rho_{g} \phi(y \mid \mu_{g}, \Sigma_{g})$$
where $\phi(.)$ is the probability density function of Gaussian distribution with mean $\mu_g$ and covariance $\Sigma_g$. The likelihood is

$$L(\theta \mid y_1, \ldots, y_n) = \prod_{i=1}^{n} \sum_{g=1}^{G} \rho_g \phi(y_i \mid \mu_g, \Sigma_g)$$

where $G$ is the number of components and $n$ is the number of observations. When the component parameters $\theta_1, \ldots, \theta_G$ are decomposed and constraints are imposed on the resulting decompositions, the result is a family of mixture models. In Gaussian families, it is usually the component covariance matrices $\Sigma_1, \ldots, \Sigma_G$ that are decomposed (cf. Section 1.5).

1.4 Cluster-Weighted Models

If the data arise from a random vector $(X', Y)'$, where $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$, and a linear dependence of $Y$ on $x$ is assumed for each mixture-component, under these assumptions, the linear cluster-weighted model (Gershenfeld, 1997) in the mixture framework is an ideal choice. The (linear) cluster-weighted model factorizes the joint density of $(X', Y)'$, in each mixture-component, into the product of the conditional density of $Y|x$ and the marginal density of $X$. When a Gaussian distribution is used for both the conditional density of $Y|x$ and the marginal density of $X$, the resulting model can be termed a (linear) Gaussian cluster-weighted model. This takes into account the potential of finite mixtures of Gaussian regressions (see Frühwirth-Schnatter, 2006, Chapter 8) and finite mixtures of Gaussian distributions.
simultaneously (see Titterington et al., 1985 and McLachlan and Peel, 2000a); the idea of the former approach is adopted to model the conditional density of $Y | x$, while the principle of the latter is used to model both the joint density of $(X', Y)'$ and the marginal density of $X$ (see Ingrassia et al., 2012, which considers the relationships between the linear Gaussian cluster-weighted model and some well-known mixture-based approaches).

1.5 Family of Models

Several different constraints have been proposed on the covariance structures that results in a parsimonious family of models.

1.5.1 Gaussian Parsimonious Clustering Models (GPCM)

Banfield and Raftery (1993), Celeux and Govaert (1995), Fraley and Raftery (1998), and Fraley and Raftery (2002) created a family of parsimonious Gaussian clustering models (GPCM) by the imposition of eigen-decomposition on the covariance matrix such that $\Sigma_g = \lambda_g D_g A_g D_g'$ where $D_g$ is the orthogonal matrix of eigenvectors, $A_g$ is a diagonal matrix proportional to the eigenvalues of $\Sigma_g$ such that $|A_g| = 1$, and $\lambda_g$ is a constant. These constraints have a geometrical interpretation such that the parameter $\lambda_g$ controls the cluster volume, $A_g$ controls the cluster shape, and $D_g$ controls the cluster orientation. Additional details will be given in Chapter 3.
1.5.2 Mixtures of Factor Analysers (MFA)/ Parsimonious Gaussian Mixture Models (PGMM)

Another family of parsimonious Gaussian mixture models (PGMMs; McNicholas and Murphy, 2008) is based on a mixture of factor analyzers (Ghahramani and Hinton, 1987). Factor analysis models assume that the observed $d$-dimensional data $\mathbf{x}$ can be modelled as $\mathbf{x}_i = \mu_g + \Lambda_g \mathbf{u}_i + \epsilon_{ig}$, where $\mathbf{u}_i$ is $q$-dimensional vector of unobserved latent factors such that $\mathbf{u}_i \sim N(0, \mathbf{I}_q)$ and $\epsilon_{ig} \sim N(0, \Psi_g)$. This allows the covariance matrix to have the form $\Sigma_g = \Lambda_g \Lambda_g' + \Psi_g$ where $\Lambda_g$ is a $(d \times q)$ matrix where $d$ is the dimensionality of the data, $q \ll d$ is the number of latent factors, $\Psi_g$ is a diagonal matrix giving the error variance. Additional details will be given in Chapter 4.

1.5.3 Modified Cholesky-decomposed Gaussian Mixture Models

In order to analyze longitudinal data, Pourahmadi (1999, 2000) explored the modified Cholesky decomposition of the covariance matrix such that $\mathbf{T}_g \Sigma_g \mathbf{T}_g' = \mathbf{D}_g$, where $\mathbf{T}_g$ is the unique lower triangular matrix with 1 in the diagonal and $-\phi_{ij}$ in the off-diagonals, $\Sigma_g$ is the covariance matrix, and $\mathbf{D}_g$ is the diagonal matrix of innovation variances. This decomposition can alternatively be written as $\Sigma_g^{-1} = \mathbf{T}_g' \mathbf{D}_g^{-1} \mathbf{T}_g$. McNicholas and Murphy (2010a) extended this structure in a Gaussian mixture modelling framework resulting in a family of models for longitudinal data. Additional details will be given in Chapter 5.
1.6 Overview of the thesis

In Chapter 2, we will give some background information on parameter estimation using the most popular expectation-maximization framework. In addition, we will also discuss several other alternatives for parameter estimation. In Chapter 3, we will illustrate the use of variational Bayes framework for the parameter estimation for the family of GPCMs with the use of deviance information criterion (DIC) for model selection.

In Chapter 4, we will extend cluster weighted models to include an underlying latent factor structure of the independent variable resulting in a family of parsimonious cluster-weighted factor analyzers (CWFA). This provides the model with the flexibility of clustering of high-dimensional data.

In Chapter 5, we will present work on a novel family of mixture of multivariate t-distributions with a modified Cholesky-decomposed covariance structure for both model-based clustering, and model-based classification, i.e., semi-supervised clustering.

In Chapter 6, we will present some concluding remarks.
Chapter 2

Parameter Estimation and Model Selection

2.1 Parameter Estimation

When mixture models are used for clustering, we call this ‘model-based clustering’ and when they are used for classification, we use the term ‘model-based classification’. In model-based clustering, the cluster/group membership of all observations are unknown and are treated as missing data. Generally, estimation of group memberships and parameters are performed in an alternating iterative process. Although the term ‘model-based classification’ is not frequently used (recent examples include McNicholas, 2010; Andrews et al., 2011), work has been conducted in this direction for around 30 years now (cf. McLachlan, 1982). Model-based classification is a semi-supervised alternative to model-based clustering, where a proportion of the observed data have a known group membership. Parameter estimation under model-based classification is obtained by jointly modelling data with known and unknown group memberships. These parameters are then used to estimate the unknown group memberships. Some of the existing parameter estimation techniques are discussed in this section.
2.1.1 EM-algorithm

The most popular approaches to parameter estimation for model-based clustering and classification remain well-routed within the expectation-maximization (EM) paradigm (Dempster et al., 1977). The EM algorithm is an iterative approach for maximizing the likelihood when the data are incomplete or are treated as incomplete. It involves two main steps, the expectation step (E-step) and the maximization step (M-step). In model-based clustering and classification, the unknown group memberships are taken to be missing data. Let $z_i = (z_{i1}, \ldots, z_{iG})$ denote the component, or group, membership where $z_{ig} = 1$ if observation $i$ belongs to group $g$ and $z_{ig} = 0$ otherwise. In the E-step, the expected value of the complete-data log-likelihood, $Q$ say, is computed based on the observed data $y_1, \ldots, y_n$ using current parameter estimates. In order to form $Q$, the unknown memberships are replaced by their conditional expected values. In the M-step, $Q$ is then maximized with respect to the model parameters. The E and M-steps are repeated until convergence.

2.1.2 Variational Bayes Approach

Bayesian approaches to mixture modelling offer the flexibility of sampling from computationally complex models that use various Markov chain Monte Carlo (MCMC) sampling algorithms. While this has led to increased interest in their use (Diebolt and Robert, 1994; Richardson and Green, 1997; Stephens, 1997, 2000; Casella et al., 2002), difficulties have been encountered with, inter alia, computational overhead and convergence (Celeux et al., 2000; Jasra et al., 2005). Variational Bayes
approximations present an alternative to MCMC algorithms for mixture modelling parameter estimation and are gaining popularity due to their fast and deterministic nature (cf. Jordan et al., 1999; Corduneanu and Bishop, 2001; Ueda and Ghahramani, 2002; McGrory and Titterington, 2007, 2009; McGrory et al., 2009). With the use of a computationally convenient approximating density in place of a more complex but ‘true’ posterior density, the variational algorithm overcomes the hurdles of MCMC sampling while still retaining the benefits of the Bayesian approach.

For observed data $y$, the joint conditional distribution of parameters $\theta$ and missing data $z$ are approximated by using another computationally convenient distribution $q(\theta, z)$. This distribution $q(\theta, z)$ is obtained by minimizing the Kullback-Leibler (KL) divergence between the true and the approximating densities, where

$$
\text{KL}(q(\theta, z) \mid p(\theta, z \mid y)) = \int_{\Theta} \sum_{z} q(\theta, z) \log \left\{ \frac{q(\theta, z)}{p(\theta, z \mid y)} \right\} d\theta.
$$

The approximating density is restricted to have a factorized form for computational convenience, so that $q(\theta, z) = q_{\theta}(\theta)q_{z}(z)$. Upon choosing a conjugate prior, the appropriate hyper-parameters approximating density $q_{\theta}(\theta)$ for data can be obtained by solving a set of coupled non-linear equations.

The variational Bayes algorithm is initialized with more components than expected. As the algorithm iterates, if two components have similar parameters then one component dominates the other causing the dominated component’s weighting to be zero. If a component’s weight becomes sufficiently small, in our analyses less than
or equal to two observations, the component is removed from consideration. Therefore, the variational Bayes approach allows for simultaneous parameter estimation and selection of the optimal number of components. However, while selecting among a family of models, variational Bayes approach needs to be used in conjunction with model selection criterion.

2.1.3 Other Approaches

**MM Algorithm**

MM can mean either ‘minorization-maximization’ or ‘majorization-minimization’ depending on the application. MM algorithms date back to as early as 1970 in the context of a line search method Ortega (1970) but it has gained attention in the finite mixture model application. The much more well known EM algorithm Dempster et al. (1977) is a special case. The conditional expectation of the complete data log-likelihood (calculated in the E-step of the EM algorithm) is a minorizing function which is then maximized with respect to the parameters of the model (M-step). Thus, the MM algorithm is a generalization of an EM algorithm. The MM-algorithm creates a surrogate function which will maximize the log-likelihood or some other objective function (Young, 2008).

**Method of Moments**

The most influential paper on the use of method of moments for inference on finite mixture models was by Pearson (1894). This was the most widely used
approach till the late 20th century until the advent of computers (Titterington et al., 1985). The method of moments estimated the model parameters ($\theta$) by selecting a set of moments of the data $y$ for the parameters ($\mathbb{E}[H_j(y)|\theta]$) such that the theoretical moments and the empirical moment ($\hat{H}_j$) are the same. The empirical moments are obtained by the sample average of $H_j(.)$ for the observed values of $y$. Details on the algorithm are available in Frühwirth-Schnatter (2006).

2.2 Model selection

When families of mixture models are used, all the above discussed algorithms must be employed in conjunction with a model selection criterion to select the ‘best’ model among the family of models and, in many cases, the number of components. The log-likelihood value on its own favours the model with more parameters and therefore, these model selection criteria penalize the log-likelihood for model complexity. There are many model selection criteria, such as the Akaike information criterion (AIC; Akaike, 1973), the Bayesian information criterion (BIC; Schwarz, 1978), the integrated completed likelihood (ICL; Biernacki et al., 2000), and the deviance information criterion (DIC Spiegelhalter et al., 2002).

2.2.1 Akaike Information Criterion (AIC)

The AIC is a well known model selection criterion based on penalizing the log-likelihood by the addition of a penalty term based on the number of parameters.
Mathematically,
\[ AIC = -2l(y, \hat{\vartheta}) + 2k, \]

where \( k \) is the number of parameters, \( l(y, \hat{\vartheta}) \) is the maximized log-likelihood, and \( \hat{\vartheta} \) is the maximum likelihood estimate of \( \vartheta \). The model with the smallest AIC is selected as the best model. However, some work by Shibata (1976) and Katz (1981) has shown that AIC tends to favour more complex models.

### 2.2.2 Bayesian Information Criterion (BIC)

The BIC is often used for model selection in model-based clustering applications involving a family of mixture models. To date, the Bayesian information criterion (BIC) has proved most effective for this model selection process, and the EM-BIC rubric has monopolized the literature on families of mixture models. For a model with parameters \( \vartheta \), the BIC is given by

\[ BIC = 2l(y, \hat{\vartheta}) - k \log n, \]

where \( l(y, \hat{\vartheta}) \) is the maximized log-likelihood, \( \hat{\vartheta} \) is the maximum likelihood estimate of \( \vartheta \), \( k \) is the number of free parameters in the model, and \( n \) is the number of observations. The use of BIC in a mixture model context is further discussed in Section 5.2.4.
2.2.3 Integrated Completed Likelihood (ICL)

The ICL essentially penalizes the BIC for estimated mean entropy, thereby punishing mixture components that are more spread out. In practice, an approximate ICL is used and this is given by

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

where MAP\{\hat{z}_{ig}\} is the maximum \textit{a posteriori} classification given \(\hat{z}_{ig}\), that is MAP\{\hat{z}_{ig}\} = 1 if \(\max_g\{z_i1,\ldots,z_iG\}\) occurs in component \(g\) and MAP\{\hat{z}_{ig}\} = 0 otherwise. The use of ICL in mixture model context is further discussed in Section 5.2.4

2.2.4 Deviance Information Criterion (DIC)

The deviance information criterion (DIC; Spiegelhalter \textit{et al.}, 2002) is a model selection criterion that measures complexity of the model and fit of the data in a Bayesian framework. Mathematically,

\[
DIC = -2 \log p(y \mid \hat{\theta}) + 2p_D
\]

where \(p_D\) which is based on deviance gives a measure of the complexity and \(\log p(y \mid \hat{\theta})\) is the posterior log-likelihood of the data. It measures the difference between the posterior mean of the deviance and deviance calculated at the posterior mean. More mathematical details are provided in Section 3.1.4.
Chapter 3

Parameter Estimation and Mixture Model

Selection for a Family of Gaussian Parsimonious Clustering models using Variational Bayes Approximations

3.1 Methodology

3.1.1 Introducing Parsimony

If $d$-dimensional data $y_1, \ldots, y_n$ arise from a finite mixture of Gaussian distributions, then the log-likelihood is

$$
\log p(y \mid \theta) = \sum_{i=1}^{n} \log \left[ \sum_{g=1}^{G} \rho_g \frac{\left| \Sigma_g^{-1} \right|}{(2\pi)^{\frac{d}{2}}} \exp \left\{ -\frac{1}{2} (y_i - \mu_g)' \Sigma_g^{-1} (y_i - \mu_g) \right\} \right].
$$

The number of parameters in the component covariance matrices of this mixture model is $Gd(d + 1)/2$, which is quadratic in $d$. When dealing with real data, the
parameters to be estimated can easily exceed the sample size. Hence, the introduction of parsimony through the imposition of additional structure on the covariance matrices is desirable.

Banfield and Raftery (1993); Celeux and Govaert (1995); Fraley and Raftery (1998, 2002) exploited geometrical constraints on the covariance matrices of Gaussian distribution using the eigen-decomposition of the covariance matrices, such that $\Sigma_g = \lambda_g D_g A_g D_g'$, where $D_g$ is the orthogonal matrix of eigenvectors and $A_g$ is a diagonal matrix proportional to the eigenvalues of $\Sigma_g$, such that $|A_g| = 1$ and $\lambda_g$ is a constant. The parameter $\lambda_g$ controls the cluster volume, $A_g$ controls the cluster shape, and $D_g$ controls the cluster orientation. This allows for imposition of several constraints on the covariance matrix that have geometrical interpretation giving rise to a family of 14 models (Table 3.1) known as Gaussian Parsimonious clustering models (GPCM).

Table 3.1: Geometric interpretation of the eigen-decomposition of the covariance matrix where I refers to identity matrix and E refers to equal, V refers to variable and S refers to Spherical across groups.

<table>
<thead>
<tr>
<th>Model (Covariance)</th>
<th>Volume</th>
<th>Shape</th>
<th>Orientation</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>EII ($\lambda I$)</td>
<td>E</td>
<td>S</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>VII ($\lambda_g I$)</td>
<td>V</td>
<td>S</td>
<td>-</td>
<td>$G$</td>
</tr>
<tr>
<td>EEI ($\lambda A$)</td>
<td>E</td>
<td>E</td>
<td>Ax-Alg</td>
<td>$d$</td>
</tr>
<tr>
<td>VEI ($\lambda_g A$)</td>
<td>V</td>
<td>E</td>
<td>Ax-Alg</td>
<td>$d + G - 1$</td>
</tr>
<tr>
<td>EVI ($\lambda A_g$)</td>
<td>E</td>
<td>V</td>
<td>Ax-Alg</td>
<td>$dG - G + 1$</td>
</tr>
<tr>
<td>VVI ($\lambda_g A_g$)</td>
<td>V</td>
<td>V</td>
<td>Ax-Alg</td>
<td>$dG$</td>
</tr>
<tr>
<td>EEE ($\lambda D A D'$)</td>
<td>E</td>
<td>E</td>
<td>E</td>
<td>$d(d + 1)/2$</td>
</tr>
<tr>
<td>VEE* ($\lambda_g D A D'$)</td>
<td>V</td>
<td>E</td>
<td>E</td>
<td>$d(d + 1)/2 + G$</td>
</tr>
<tr>
<td>EVE* ($\lambda D A_g D'$)</td>
<td>E</td>
<td>V</td>
<td>E</td>
<td>$dG - G + 1 + d(d+1)/2$</td>
</tr>
<tr>
<td>VVE* ($\lambda_g D A_g D'$)</td>
<td>V</td>
<td>V</td>
<td>E</td>
<td>$dG + (d-1)d/2$</td>
</tr>
<tr>
<td>EEV ($\lambda D_g A D_g'$)</td>
<td>E</td>
<td>E</td>
<td>V</td>
<td>$Gd(d + 1)/2 - (G - 1)d$</td>
</tr>
<tr>
<td>VEV ($\lambda_g D_g A D_g'$)</td>
<td>V</td>
<td>E</td>
<td>V</td>
<td>$Gd(d + 1)/2 - (G - 1)(d - 1)$</td>
</tr>
<tr>
<td>EVV* ($\lambda D_g A_g D_g'$)</td>
<td>E</td>
<td>V</td>
<td>V</td>
<td>$Gd(d + 1)/2 - (G - 1)$</td>
</tr>
<tr>
<td>VVV ($\lambda_g D_g A_g D_g'$)</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>$Gd(d + 1)/2$</td>
</tr>
</tbody>
</table>
All models except those marked * or ** (Table 3.1) are implemented within the EM framework in the R package mclust (Fraley and Raftery, 1998). Bayesian regularization of some of these models has been considered by Fraley and Raftery (2007). After assigning a highly dispersed conjugate prior, they replaced the maximum likelihood estimator of the group membership obtained using the EM algorithm by a maximum a posteriori probability (MAP) estimator such that \( \text{MAP}(\hat{z}_{ig}) = 1 \) if \( \max_g(\hat{z}_{ig}) \) occurs in component \( g \) and \( \text{MAP}(\hat{z}_{ig}) = 0 \) otherwise. A modified BIC using the maximum a posteriori probability was then used for model selection. However, here we implement those models and the models denoted by ** using variational Bayes approximations; conjugate priors for the models denoted by * were not available.

### 3.1.2 Priors and Approximating Densities

As suggested by McGrory and Titterington (2007), the Dirichlet distribution was used as the conjugate prior for the mixing proportion, such that

\[
p(\rho) = \text{Dir}(\rho; \alpha_1^{(0)}, \ldots, \alpha_G^{(0)})
\]

and, conditional on the precision matrix \( T_g \), independent normal distributions were used as the conjugate priors for the means, such that

\[
p(\mu \mid T_1, \ldots, T_G) = \prod_{g=1}^{G} \phi(\mu_g; m_g^{(0)}, (\beta_g^{(0)} T_g)^{-1}),
\]
where $\alpha_g^{(0)}$, $m_g^{(0)}$, and $\beta_g^{(0)}$ are the hyperparameters.

Fraley and Raftery (2007) assigned priors on the parameters for the covariance matrix and its components in a Bayesian regularization application. However, we assign priors on the precision matrix with the hyperparameters shown in Table 3.2. Note that it was not possible to put a prior on the matrix $A_g$ for the models EVI and VVI and $A$ for models VEV and VEI that met the constraint their determinant is 1. We therefore put a prior on $c_gA_g^{-1}$ or $cA^{-1}$, where $c_g$ or $c$ is a fixed constant. Using the expected value of $c_gA_g^{-1}$ or $cA^{-1}$, the expected value of $A_g^{-1}$ or $A^{-1}$ was determined to satisfy the constraint that their determinant is 1. Because $D_g$ is the orthogonal matrix of eigenvectors, the matrix von Mises-Fisher (or Langevin) distribution (Downs, 1972; Khatri and Mardia, 1977) is used as the prior for $D_g$. The von Mises-Fisher distribution is a probability distribution on a set of orthonormal matrices that is widely used in orientation statistics and has recently been used in multivariate analysis and matrix decomposition methods (Hoff, 2009). The density of the von Mises-Fisher distribution as defined by Downs (1972) is

$$p(D) = a(C) \exp \text{tr}(CD^t),$$

for $D \in O(n,p)$, where $O(n,p)$ is the Stiefel manifold of $n \times p$ matrices, parameter $C$ is a fixed $n \times p$ matrix and $a(C)$ is a normalizing constant. The resulting posterior distribution in this case is a matrix Bingham-von Mises-Fisher or matrix Langevin-Bingham distribution (Khatri and Mardia, 1977). The density of a matrix Bingham-
von Mises-Fisher distribution is given by

\[ p(D|A,B,C) \propto \exp(\text{tr}(C'D + BD'AD)), \]

where \( A \) and \( B \) are symmetric and diagonal matrices, respectively.

Samples from the matrix Bingham-von Mises-Fisher distribution can be obtained using the Gibbs sampling algorithm implemented in the R package \texttt{rstiefel} (Hoff, 2012).

The approximating densities that minimize the KL divergence are as follows. For the mixing proportions, \( q_\pi(\pi) = \text{Dir}(\pi, \alpha_1, \ldots, \alpha_G) \), where \( \alpha_g = \alpha_g^{(0)} + \sum_{i=1}^n \hat{z}_{ig} \).

For the mean, \( q_\mu(\mu | T_1, \ldots, T_G) = \prod_{g=1}^G N_p(\mu_g; m_g, (\beta_g T_g)^{-1}) \), where \( \beta_g = \beta_g^{(0)} + \sum_{i=1}^n \hat{z}_{ig} \) and

\[ m_g = \frac{\beta_g^{(0)} m_g^{(0)} + \sum_{i=1}^n \hat{z}_{ig} y_i}{\beta_g}. \]

The probability that the \( i \)th observation belongs to a group \( g \) is then given by \( \hat{z}_{ig} = \frac{\hat{\varphi}_{ig}}{\sum_{j=1}^G \hat{\varphi}_{ij}} \), where

\[ \hat{\varphi}_{ig} = \exp \left( \mathbb{E}[\log \rho_g] + \frac{1}{2} \mathbb{E}[\log |T_g|] - \frac{1}{2} \text{tr} \left\{ \mathbb{E}[T_g](y_i - \mathbb{E}[\mu_g])(y_i - \mathbb{E}[\mu_g])' + \frac{1}{\beta_g} I_p \right\} \right). \]
\[ E[\mu_g] = m_g, \ E[\log(\rho_g)] = \Psi(\hat{\alpha}_g) - \Psi \left(\sum_{g=1}^{G} \hat{\alpha}_g\right), \text{ and } \Psi(\cdot) \text{ is the digamma function.} \]

The values of \( E[T_g] \) and \( E[\log |T_g|] \) vary depending on the model (cf. Table A.2). The posterior distribution of the parameters \( \lambda_g^{-1} \) and the diagonal elements of \( A_g \) are well known gamma distributions and, therefore, the expected value of \( E[\lambda_g^{-1}], E[\log |\lambda_g^{-1}|], E[A_g], \) and \( E[\log |A_g|] \) have a closed form. The posterior distribution for \( D_gA_gD'_g \) is Wishart which has a closed form solution for \( E[D_gA_gD'_g] \) and \( E[\log |D_gA_gD'_g|] \). The posterior distribution of the parameter \( D_g \) was a matrix Bingham-von Mises-Fisher distribution (see Appendix A.1.1 for the mathematical details). The Bingham-von Mises-Fisher distribution does not have a closed form solution to find the expected values of \( E[T_g] \) and \( E[\log |T_g|] \) and, hence, Monte Carlo integration was used.

### 3.1.3 Convergence

The posterior log-likelihood of the observed data is

\[
\log p(y_1, \ldots, y_n | \tilde{\theta}) = \sum_{i=1}^{n} \log \left[ \sum_{g=1}^{G} \bar{\rho}_g \frac{|T_g|}{2\pi^{d/2}} \right] \times \exp \left\{ \frac{1}{2}(y_i - \tilde{\mu}_g)'\tilde{T}_g(y_i - \tilde{\mu}_g) \right\},
\]

where \( \bar{\rho}_g = \alpha_j / \sum_{g=1}^{G} \alpha_g \) and \( \tilde{\mu}_g = m_g. \) The expected precision matrix \( \tilde{T}_g \) varies according to the model. Convergence of the algorithm for these models is determined using a modified Aitken acceleration criterion. The Aitken acceleration (Aitken, 1926)
is given by

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

where \( l^{(m-1)}, l^m, \) and \( l^{(m+1)} \) are values of the posterior log-likelihoods at iterations \( m - 1, m, \) and \( m + 1, \) respectively. Convergence is achieved when \( l^{(m+1)} - l^{(m+1)} < \epsilon, \)

where \( \epsilon \) is a small constant and \( l^{(m+1)} \) is an asymptotic estimate of the log-likelihood (Böhning et al., 1994) given by

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

The VEV and EEV models utilize Gibbs sampling and Monte Carlo integration to find both the expected value of the parameter \( T_g \) and the expectations of functions of \( T_g. \) As the Gibbs sampling chain approaches the stationary posterior distribution, the posterior likelihood oscillates around the maximum likelihood rather than increasing at every new iteration. This would lead our modified Aitken’s acceleration criterion to fail to determine convergence. Hence, our variational Bayes algorithm was modified to ensure that the log-likelihood increases at every iteration. This modification is simple: if the parameter estimates obtained using Gibbs sampling fail to increase the posterior log-likelihood, those estimates are discarded and resampled using different random starts. Hence, the check for convergence can be achieved using a modified Aitken’s acceleration criterion. However, due to the use of the Monte Carlo approximation, the posterior likelihood at every iteration is not monotonic. The maximum posterior likelihood (or a value very close to it) can be
reached before the difference between successive likelihoods is small enough to be detected by the modified Aitken's acceleration criterion. No further values will then be accepted if the maximum posterior likelihood is reached or very few values will be accepted if the likelihood is close to the maximum. In such scenarios, the algorithm will take an extremely long time to converge sometimes even fail to converge. Hence, to reduce the computational burden, if no values were accepted with 50 different random starts for Gibbs sampling at an iteration, convergence to maximum posterior likelihood was assumed.

3.1.4 Model Selection

Despite the benefits of simultaneously obtaining parameter estimates along with the number of components, a model selection criterion is needed to determine the covariance structure. For the selection of the model with the best fit, the deviance information criterion (DIC; Spiegelhalter et al., 2002) is used as suggested by McGrory and Titterington (2007). The DIC is given by

$$
\text{DIC} = -2 \log p(y | \tilde{\theta}) + 2p_D
$$

where

$$
2p_D \approx -2 \int q_\theta(\theta) \log \left( \frac{q_\theta(\theta)}{p(\theta)} \right) d\theta + 2 \log \left( \frac{q_\theta(\tilde{\theta})}{p(\tilde{\theta})} \right),
$$

and \( \log p(y | \tilde{\theta}) \) is the posterior log-likelihood of the data.
3.1.5 Performance Assessment

The adjusted Rand index (ARI; Hubert and Arabie, 1985) is used to assess the performance of the classification techniques applied in Section 3.2. The Rand index (Rand, 1971) is based on the pairwise agreement between the predicted and true classifications. The ARI corrects the Rand index to account for agreement by chance: a value of ‘1’ indicates perfect agreement, ‘0’ indicates random classification, and a negative values indicate a classification that is worse than would be expected by guessing.

3.1.6 Model-Based Classification

Model-based classification (cf. McLachlan, 1982), a semi-supervised alternative to model-based clustering, has been garnering much attention in recent literature (Dean et al., 2006; McNicholas, 2010; Andrews et al., 2011; Andrews and McNicholas, 2011b; Browne et al., 2012). Dean et al. (2006) and McNicholas (2010) demonstrated that model-based classification outperformed model based clustering. Model-based classification is best explained through likelihoods.

In the model-based clustering framework, where the group membership of all the observations are taken to be unknown, the likelihood is given by

$$\mathcal{L}(\theta \mid y_1, \ldots, y_n) = \prod_{i=1}^{n} \sum_{g=1}^{G} \rho_g \phi(y_i \mid \mu_g, \Sigma_g).$$

In the model-based classification framework, suppose that there are $k$ observations
with known group memberships. Without loss of generality, order the data so that the first \( k \) observations have known group memberships. Then, the likelihood is

\[
\mathcal{L}(\vartheta \mid y_1, \ldots, y_n) = \prod_{i=1}^{k} \prod_{g=1}^{G} [\rho_g \phi(y_i \mid \mu_g, \Sigma_g)]^{z_{ig}} \times \prod_{j=k+1}^{n} \sum_{h=1}^{G} \rho_h \phi(y_j \mid \mu_h, \Sigma_h).
\]

Parameter estimation under model-based classification is obtained by jointly modelling data with known and unknown group memberships. These parameters are then used to estimate the unknown group memberships.

### 3.2 Results

#### 3.2.1 Simulation Study 1

Our variational Bayes algorithm was run on a simulated two-dimensional Gaussian data set with three components and known mean and covariance structures \( \Sigma_g = \lambda_g I \) (VII, see Table 3.4). We ran multiple simulations, with different random starts, and we set the maximum number of components to ten each time.

The model with the minimum DIC was the VII model, which gave perfect classifications for all 10 random starts. Note that the models with closer covariance structures, such as VEI (\( \lambda_g A \)) and VVI (\( \lambda_g A_g \)), also give perfect classifications at every random start. However, their DIC is slightly higher than the true model VII (cf. Table 3.3). Also, it should be noted that the VEV and EEV models had a larger
Figure 3.1: The three component simulated data (left) along with the best classification given by the random start selected by the DIC for the EII model (right).

range of DIC, indicating more sensitivity to starting values. Estimation for the VEV and EEV models utilizes Gibbs sampling. The estimated parameters of model VII, \( \mu_g \) and \( \lambda_g \) were very close to the true parameters (See Table 3.4).

As seen in Figure 3.1, the data are clearly spherical with unequal covariances. Hence, forcing an incorrect covariance structure might result in misrepresentation of cluster membership. For example, let us consider results for our simulated data. If forced to have covariance structure EII, which imposes clearly inappropriate spherical clusters, the result is over-estimation of the number of components. Figure 3.1 depicts results for the EII model with the minimum DIC from the 10 runs. Hence, it can be argued that despite the true model being VII with three components, in so far as the EII model is concerned the best model actually has five components. Also, it should be noted that the best classification, as chosen by the DIC, for all models with
varying volume, i.e., $\lambda_g$, always gave a perfect classification (Table 3.5).

The algorithm was also compared to the widely used EM-framework within the \texttt{mclust} package in R. To facilitate a comparison with our approach, we ran our variational Bayes-DIC rubric using starting values from the \texttt{hclass} function from \texttt{mclust}.

Note that the VEE and EVV models are not implemented within the EM-framework in \texttt{mclust}. For the models with perfect classification, the DIC using the random and the hierarchical clustering starts were the same; however, the DIC of the models with the classification from hierarchical clustering were better for the models without perfect classifications (see Tables 3.3 and 3.5). Even though the DIC of the models chosen by \texttt{mclust} were smaller, the number of components of these models was always greater. This brings us back to the aforementioned argument that the model with $G = 3$ components might give the correct classification if the covariance structure is misspecified. We also analyzed the data within the EM-framework using the \texttt{mclust} package. The EM approach in conjunction with the BIC also chose the VII model with perfect classification, but all other models also gave a perfect classification. For comparison, the BIC for each model was also calculated using the posterior log-likelihood of the model with the \texttt{mclust} package in R. The best model selected by the BIC was again VII, which also had the highest BIC. Hence, model selection using the DIC and BIC seem to be in agreement with one another for these data.
3.2.2 Simulation Study 2

We ran another simulation study with three component three dimensional Gaussian distributions with known mean and covariance structure $\Sigma_g = \Sigma = \lambda \mathbf{DAD}'$. Again, 10 different runs with different random starts were used and the maximum number of components was set to 10. The best model selected by the DIC was the true model (EEE), which consistently gave perfect classification. Again, the range of the DIC for the EEV and VEV was also comparatively large (Table 3.6).

The estimated parameters for the EEE model were very close to the true parameters; the values of $\mu_g$ and $\hat{\mu}_g$ using one random start are given in Table 3.7 and the values for $\Sigma$ and $\hat{\Sigma}$ are

$$\Sigma = \begin{bmatrix} 0.50 & 0.35 & 0.25 \\ 0.35 & 1.00 & 0.45 \\ 0.25 & 0.45 & 1.20 \end{bmatrix}, \quad \hat{\Sigma} = \begin{bmatrix} 0.48 & 0.28 & 0.25 \\ 0.28 & 0.97 & 0.44 \\ 0.25 & 0.44 & 1.14 \end{bmatrix}.$$

A comparison with mclust was carried out in exactly the same way as before (Section 3.2.1). The results (Table 3.8) again suggest a greater tendency for selection of more components using this approach. In terms of the best model overall, the BIC and DIC were in agreement with one another.
3.2.3 Clustering of Leptograpsus Crabs Data

The Leptograpsus crab data set, publicly available in the package MASS in R, consists of biological measurements on 100 crabs from two different species (orange and blue) with 50 males and 50 females of each species. The biological measurements (in millimeters) include frontal lobe size, rear width, carapace length, carapace width, and body depth. Although this data set has been analyzed quite often in the literature using several different clustering approaches, the correlation among the variables makes it difficult to cluster (Figure 3.2). Due to this known issue with the data set, we introduced an initial step of processing using principal component analysis. Principal component analysis used orthogonal transformation to convert these correlated variables into linearly uncorrelated principal components (Figure 3.3). Finally, the variational Bayes algorithm was run on these uncorrelated principal components with a maximum of $G = 6$ components.

The VVV model was selected by the DIC criterion ($\text{DIC} = 2594.893$) and an adjusted Rand index of 0.44 relative to the partition given by species (Table 3.11). However, the EVV model had an ARI of 1 but a much higher DIC. Note that the classification obtained by the VVV model (Table 3.10) leads to the blue crabs having membership in clusters 1, 2, and 4, with clusters 3 and 5 containing orange crabs only. Cluster 6, however, contains only 6 observations — 3 blue and 3 orange — which could potentially be a group of outliers. On the other hand, the EVV model had a perfect classification but a higher DIC. The VVV model seems to create sub-clusters, consequently resulting in a spike in the log-likelihood and thereby lowering the DIC.
Figure 3.2: Scatter plot matrix showing the relationships among the variables of Leptograpsus crab data.
Figure 3.3: Scatter plot matrix showing the relationships among the uncorrelated principal components.
However, such sub-clusters might be equally informative as they could explain some other unknown variations within a given cluster.

It should be noted that it is not clear that choosing the model with the minimum DIC leads to the model with the best classification. The crabs could be classified based on species (blue and orange) only, or sex only, or by both sex and species. The authors suspect that a model selection criterion such as the DIC might be more appropriate for choosing the best model among models with the same covariance structure rather than between covariance structures.

### 3.2.4 Classification of Olive Oil Data

The olive oil data set, originally reported by Forina and Tiscornia (1982); Forina et al. (1983), consists of the percentage composition of eight fatty acids obtained through lipid fractionation of 572 olive oils from nine different regions in Italy: North Apulia, Calabria, South Apulia, Sicily, Inland Sardinia, East Linguria, West Linguria, and Umbria. These data are publicly available in the R package `pgmm` (McNicholas et al., 2011) and have been previously used for classification and clustering examples; they are known to be a very challenging data set for clustering (Cook and Swayne, 2007). We take a model-based classification approach, assuming that 50% of the data have known classifications and the remaining 50% are unknown. Our algorithm was run with 10 randomly selected 50/50 partitions of the data. The best model chosen by the DIC was VVV, with a DIC of 3755.207–3934.210 and ARI values in the range 0.91–0.95 (Table 3.11). The model EVV also has a very close DIC at every run, ranging from 3757.207–3936.210 with an ARI of 0.91–0.95. Precisely how
to handle close DIC values, or indeed close BIC values, remains problematic. For these data, the best classification results were given by the VEV model with an ARI of 0.96 but a much higher DIC.

The classification performance using the variational Bayes algorithm was compared with \texttt{mclust} discriminant analysis available in the \texttt{mclust} package in \texttt{R}. MCLUST discriminant analysis (\texttt{mclustDA}) is a classification technique that performs parameter estimation using a training set of observations with known group memberships and predicts the group membership of the test set using the posterior MAP classifications. Applying \texttt{mclustDA} to these data resulted in an ARI ranging from 0.19–0.68 over all 10 runs.

This data set has also been analyzed using latent Gaussian mixture models in a classification framework (McNicholas, 2010). These models mostly outperformed \texttt{mclustDA} on these data. The misclassification rate for all 10 runs using our variational Bayes algorithm ranged from 3.15–7.69, which is comparable to the performance reported by McNicholas (2010). Comparing our approach with a MCLUST model-based classification approach would be interesting, but there is no model-based classification facility built into \texttt{mclust}.

### 3.3 Discussion

The performance of the variational Bayes approach seems comparable to the EM approach for model-based clustering. The parameters estimated using variational Bayes approximations were very close to the true parameters and perfect classification
was obtained using the true model. Variational Bayes also allows us to use two more models than \texttt{mclust}. This paper revisits the question of whether the original cluster structure is still the true cluster structure if the covariance structure is misspecified. We would argue that the true cluster structure changes according to the covariance structure and, hence, correct classification using one covariance structure might not be the correct classification for other covariance structures.

In summary, we have explored an alternative Bayesian approach to the most widely used family of Gaussian mixture models: MCLUST. The use of variational Bayes in conjunction with the DIC for a family of mixture models is a novel idea and lends itself nicely to further research. It also provides the flexibility to model complex structures, for example the EVV and VEE models that are not implemented in \texttt{mclust}. As opposed to the traditional EM approach, the variational Bayes approach seems less sensitive to starting values overall, the models that utilize the Gibbs sampling technique being the exception. This will be the subject of future work. Moreover, the DIC, which is very well suited to Bayesian models, provides and alternative model selection criterion to the traditional BIC. This paper also explores several issues with the cluster structure and provides promising avenues for further research.
Table 3.2: Prior distributions for the parameters of the eigen-decomposed covariance structures.

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Covariance</th>
<th>Parameter</th>
<th>Prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>EII</td>
<td>$\lambda I$</td>
<td>$\lambda^{-1}$</td>
<td>Gamma ($a^{(0)}, b^{(0)}$)</td>
</tr>
<tr>
<td>VII</td>
<td>$\lambda_g I$</td>
<td>$\lambda_g^{-1}$</td>
<td>Gamma ($a_g^{(0)}, b_g^{(0)}$)</td>
</tr>
<tr>
<td>EEI</td>
<td>$\lambda A$</td>
<td>$k$th diagonal element of $(\lambda A)^{-1}$</td>
<td>Gamma ($a_k^{(0)}, b_k^{(0)}$)</td>
</tr>
<tr>
<td>VEI</td>
<td>$\lambda_g A$</td>
<td>$\lambda_g^{-1}$</td>
<td>Gamma ($a_g^{(0)}, b_g^{(0)}$)</td>
</tr>
<tr>
<td>EVI</td>
<td>$\lambda A_g$</td>
<td>$k$th diagonal elements of $c A^{-1}$</td>
<td>Gamma ($a^{(0)}, b^{(0)}$)</td>
</tr>
<tr>
<td>VVI</td>
<td>$\lambda_g A_g$</td>
<td>$\lambda_g^{-1}$</td>
<td>Gamma ($a_g^{(0)}, b_g^{(0)}$)</td>
</tr>
<tr>
<td>EEE</td>
<td>$\lambda DAD'$</td>
<td>$T = (\lambda DAD')^{-1}$</td>
<td>Wishart ($\nu^{(0)}, \Sigma^{(0)-1}$)</td>
</tr>
<tr>
<td>VEE</td>
<td>$\lambda_g DAD'$</td>
<td>$\lambda_g^{-1}$</td>
<td>Gamma ($a_g^{(0)}, b_g^{(0)}$)</td>
</tr>
<tr>
<td>EEV</td>
<td>$\lambda D_g A D_g'$</td>
<td>$k$th diagonal elements of $(\lambda A)^{-1}$</td>
<td>Gamma ($a_k^{(0)}, b_k^{(0)}$)</td>
</tr>
<tr>
<td>VEV</td>
<td>$\lambda_g D_g A D_g'$</td>
<td>$\lambda_g^{-1}$</td>
<td>Gamma ($a_g^{(0)}, b_g^{(0)}$)</td>
</tr>
<tr>
<td>EVV</td>
<td>$\lambda D_g A_g D_g'$</td>
<td>$\lambda^{-1}$</td>
<td>Gamma ($a^{(0)}, b^{(0)}$)</td>
</tr>
<tr>
<td>VVV</td>
<td>$\lambda_g D_g A_g D_g'$</td>
<td>$T_g = (\lambda_g D_g A_g D_g')^{-1}$</td>
<td>Wishart ($\nu_g^{(0)}, \Sigma_g^{(0)}$)</td>
</tr>
</tbody>
</table>
Table 3.3: Summary of the variational Bayes analysis of the two-dimensional simulated data using 10 different random starts.

<table>
<thead>
<tr>
<th>Model</th>
<th>Range of DIC</th>
<th>G(ARI)</th>
<th>max. ARI</th>
</tr>
</thead>
<tbody>
<tr>
<td>EII</td>
<td>2667.063–2694.071</td>
<td>5(0.69)</td>
<td>1</td>
</tr>
<tr>
<td>VII</td>
<td>2608.909–2608.909</td>
<td>3(1)</td>
<td>1</td>
</tr>
<tr>
<td>EEI</td>
<td>2671.156–3431.661</td>
<td>5(0.69)</td>
<td>1</td>
</tr>
<tr>
<td>V EI</td>
<td>2610.728–2610.728</td>
<td>3(1)</td>
<td>1</td>
</tr>
<tr>
<td>EVI</td>
<td>2692.363–3405.379</td>
<td>4(0.99)</td>
<td>1</td>
</tr>
<tr>
<td>VVI</td>
<td>2615.471–2615.471</td>
<td>3(1)</td>
<td>1</td>
</tr>
<tr>
<td>EEE</td>
<td>2690.559–2691.912</td>
<td>3(1)</td>
<td>1</td>
</tr>
<tr>
<td>VEE</td>
<td>2612.663–2612.663</td>
<td>3(1)</td>
<td>1</td>
</tr>
<tr>
<td>EVV</td>
<td>2685.597–2798.992</td>
<td>4(0.87)</td>
<td>1</td>
</tr>
<tr>
<td>VVV</td>
<td>2600.353–3494.347</td>
<td>3(1)</td>
<td>1</td>
</tr>
<tr>
<td>EVV</td>
<td>2614.778–2617.280</td>
<td>4(0.92)</td>
<td>1</td>
</tr>
<tr>
<td>VVV</td>
<td>2613.617–2613.617</td>
<td>3(1)</td>
<td>1</td>
</tr>
</tbody>
</table>

*True model.

Table 3.4: Estimated parameters along with true parameters of the two-dimensional simulated data for one run.

<table>
<thead>
<tr>
<th>n</th>
<th>(\mu_g)</th>
<th>(\hat{\mu}_g)</th>
<th>(\lambda_g)</th>
<th>(\hat{\lambda}_g)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>(-2.2)</td>
<td>(-2.11, 2.03)</td>
<td>0.5</td>
<td>0.52</td>
</tr>
<tr>
<td>150</td>
<td>(8,0)</td>
<td>(8.11,-0.02)</td>
<td>1.2</td>
<td>1.31</td>
</tr>
<tr>
<td>75</td>
<td>(-7,-7)</td>
<td>(-7.17,-7.11)</td>
<td>2.5</td>
<td>2.10</td>
</tr>
</tbody>
</table>

Table 3.5: Summary of the variational Bayes and mclust analysis of the two-dimensional simulated data using hclass starting values.

<table>
<thead>
<tr>
<th>Model</th>
<th>Variational Bayes</th>
<th>mclust</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>G(ARI)</td>
<td>DIC</td>
</tr>
<tr>
<td>EII</td>
<td>7(0.66)</td>
<td>2651.3</td>
</tr>
<tr>
<td>VII</td>
<td>3(1)</td>
<td>\textbf{2608.909}</td>
</tr>
<tr>
<td>EEI</td>
<td>7(0.66)</td>
<td>2652.002</td>
</tr>
<tr>
<td>V EI</td>
<td>3(1)</td>
<td>2610.728</td>
</tr>
<tr>
<td>EVI</td>
<td>3(1)</td>
<td>2692.443</td>
</tr>
<tr>
<td>VVI</td>
<td>3(1)</td>
<td>2613.952</td>
</tr>
<tr>
<td>EEE</td>
<td>7(0.67)</td>
<td>2652.462</td>
</tr>
<tr>
<td>VEE</td>
<td>3(1)</td>
<td>2612.663</td>
</tr>
<tr>
<td>EVV</td>
<td>7(0.68)</td>
<td>2650.854</td>
</tr>
<tr>
<td>VEV</td>
<td>7(0.83)</td>
<td>2749.065</td>
</tr>
<tr>
<td>EVV</td>
<td>3(1)</td>
<td>2615.618</td>
</tr>
<tr>
<td>VVV</td>
<td>3(1)</td>
<td>2613.617</td>
</tr>
</tbody>
</table>
Table 3.6: Summary of variational Bayes analysis of the three-dimensional simulated data using 10 different random starts.

<table>
<thead>
<tr>
<th>Model</th>
<th>Range of DIC</th>
<th>G(ARI)</th>
<th>max. ARI</th>
</tr>
</thead>
<tbody>
<tr>
<td>EII</td>
<td>3212.374–3315.574</td>
<td>8(0.60)</td>
<td>0.95</td>
</tr>
<tr>
<td>VII</td>
<td>3218.001–3249.194</td>
<td>8(0.54)</td>
<td>0.73</td>
</tr>
<tr>
<td>EEI</td>
<td>3189.586–3251.990</td>
<td>7(0.64)</td>
<td>0.87</td>
</tr>
<tr>
<td>VEI</td>
<td>3211.733–3225.936</td>
<td>6(0.63)</td>
<td>0.71</td>
</tr>
<tr>
<td>EVI</td>
<td>3228.900–3283.098</td>
<td>6(0.65)</td>
<td>0.96</td>
</tr>
<tr>
<td>VVI</td>
<td>3264.784–3566.454</td>
<td>5(0.70)</td>
<td>0.98</td>
</tr>
<tr>
<td>EEE*</td>
<td><strong>3146.962–3146.962</strong></td>
<td><strong>3(1)</strong></td>
<td><strong>1</strong></td>
</tr>
<tr>
<td>VEE</td>
<td>3148.806–3154.764</td>
<td>4(0.94)</td>
<td>1</td>
</tr>
<tr>
<td>EEV</td>
<td>3203.836–3855.437</td>
<td>4(0.62)</td>
<td>1</td>
</tr>
<tr>
<td>VEV</td>
<td>3189.757–3225.281</td>
<td>5(0.82)</td>
<td>1</td>
</tr>
<tr>
<td>EEV</td>
<td>3159.890–3165.685</td>
<td>4(0.93)</td>
<td>1</td>
</tr>
<tr>
<td>VVV</td>
<td>3157.890–3167.064</td>
<td>4(0.93)</td>
<td>1</td>
</tr>
</tbody>
</table>

*True model.

Table 3.7: Estimated and true mean parameters for the EEE model of the three-dimensional simulated data.

<table>
<thead>
<tr>
<th>n</th>
<th>( \mu_g )</th>
<th>( \hat{\mu}_g )</th>
</tr>
</thead>
<tbody>
<tr>
<td>150</td>
<td>(-2, -2, -2)</td>
<td>(-2.08, -1.96, -1.84)</td>
</tr>
<tr>
<td>100</td>
<td>(4, 0, 0)</td>
<td>(3.95, -0.08, -0.06)</td>
</tr>
<tr>
<td>75</td>
<td>(-5, 0, 2)</td>
<td>(-5.04, -0.03, 1.81)</td>
</tr>
</tbody>
</table>

Table 3.8: Summary of variational Bayes analysis of the three-dimensional simulated data using classification from hierarchical clustering as the starting values.

<table>
<thead>
<tr>
<th>Model</th>
<th>Variational Bayes</th>
<th>mclust</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>G(ARI)</td>
<td>DIC</td>
</tr>
<tr>
<td>EII</td>
<td>9(0.57)</td>
<td>3212.726</td>
</tr>
<tr>
<td>VII</td>
<td>9 (0.54)</td>
<td>3213.432</td>
</tr>
<tr>
<td>EEI</td>
<td>8(0.65)</td>
<td>3191.768</td>
</tr>
<tr>
<td>VEI</td>
<td>8(0.55)</td>
<td>3194.519</td>
</tr>
<tr>
<td>EVI</td>
<td>9 (0.55)</td>
<td>3216.638</td>
</tr>
<tr>
<td>VVI</td>
<td>6(0.88)</td>
<td>3245.793</td>
</tr>
<tr>
<td>EEE</td>
<td><strong>3 (1)</strong></td>
<td><strong>3146.962</strong></td>
</tr>
<tr>
<td>VEE</td>
<td>4(0.94)</td>
<td>3149.325</td>
</tr>
<tr>
<td>EEV</td>
<td>10(0.54)</td>
<td>3220.821</td>
</tr>
<tr>
<td>VEV</td>
<td>4 (0.96)</td>
<td>3216.937</td>
</tr>
<tr>
<td>EVV</td>
<td>4 (0.93)</td>
<td>3166.813</td>
</tr>
<tr>
<td>VVV</td>
<td>6(0.97)</td>
<td>3163.685</td>
</tr>
</tbody>
</table>
Table 3.9: Summary of the variational Bayes analysis of the principal components of the Leptograpsus crab data using 10 different random starts. The ARI was computed using species.

<table>
<thead>
<tr>
<th>Model</th>
<th>range DIC</th>
<th>ARI Values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EII</td>
<td>4196.889–4220.479</td>
<td>0.016</td>
</tr>
<tr>
<td>VII</td>
<td>4110.647–4903.835</td>
<td>0.19</td>
</tr>
<tr>
<td>EEI</td>
<td>2724.222–2885.731</td>
<td>0.33</td>
</tr>
<tr>
<td>VEI</td>
<td>2624.506–2744.128</td>
<td>0.33</td>
</tr>
<tr>
<td>EVI</td>
<td>2606.964–2624.034</td>
<td>0.22</td>
</tr>
<tr>
<td>VVI</td>
<td>2967.315–2968.406</td>
<td>0.24</td>
</tr>
<tr>
<td>EEE</td>
<td>2703.988–2960.721</td>
<td>0.34</td>
</tr>
<tr>
<td>VEE</td>
<td>2716.454–2849.483</td>
<td>0.38</td>
</tr>
<tr>
<td>EEV</td>
<td>2698.490–2826.421</td>
<td>0.23</td>
</tr>
<tr>
<td>VEV</td>
<td>2690.440–2792.741</td>
<td>0.32</td>
</tr>
<tr>
<td>EVV</td>
<td>2791.910–2813.595</td>
<td>0.23</td>
</tr>
<tr>
<td>VVV</td>
<td>2594.893–2760.433</td>
<td>0.44</td>
</tr>
</tbody>
</table>

Table 3.10: Classification of the principal components of the Leptograpsus crabs data using the VVV and EVV models.

<table>
<thead>
<tr>
<th></th>
<th>VVV</th>
<th>EVV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue</td>
<td>38</td>
<td>0</td>
</tr>
<tr>
<td>Orange</td>
<td>0</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 3.11: Summary of the variational Bayes analysis of the olive oil data using 10 different random starts.

<table>
<thead>
<tr>
<th>Model</th>
<th>range DIC</th>
<th>ARI Ranges</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>mclustDA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EII</td>
<td>8865.603–9030.999</td>
<td>0.84 (0.84–0.90)</td>
</tr>
<tr>
<td>VII</td>
<td>8437.173–8484.987</td>
<td>0.87(0.87–0.91)</td>
</tr>
<tr>
<td>EEI</td>
<td>8382.770–8503.314</td>
<td>0.87(0.85–0.91)</td>
</tr>
<tr>
<td>VEI</td>
<td>7965.48–8024.94</td>
<td>0.78(0.84–0.93)</td>
</tr>
<tr>
<td>EVI</td>
<td>7502.372–7672.412</td>
<td>0.86(0.84–0.91)</td>
</tr>
<tr>
<td>VVI</td>
<td>6521.262–6663.063</td>
<td>0.87(0.87–0.93)</td>
</tr>
<tr>
<td>EEE</td>
<td>5846.456–5922.331</td>
<td>0.89(0.89–0.92)</td>
</tr>
<tr>
<td>VEE</td>
<td>5508.114–5623.301</td>
<td>0.87(0.86–0.91)</td>
</tr>
<tr>
<td>EEV</td>
<td>10132.15–10270.94</td>
<td>0.85(0.85–0.98)</td>
</tr>
<tr>
<td>VEV</td>
<td>9052.345–9422.970</td>
<td>0.96(0.88–0.98)</td>
</tr>
<tr>
<td>EVV</td>
<td>3755.207–3936.210</td>
<td>0.95(0.91–0.95)</td>
</tr>
<tr>
<td>VVV</td>
<td>3755.207–3934.210</td>
<td>0.95(0.91–0.95)</td>
</tr>
</tbody>
</table>
Chapter 4

A family of Cluster-Weighted Factor Analyzers

4.1 Methodology

4.1.1 Cluster-Weighted Modeling

Let \((X, Y)\) be a pair of random vector \(X\) and random variable \(Y\) defined on a given population \(\Omega\) with joint probability distribution \(p(x, y)\), where \(X\) is the \(d\)-dimensional input vector taking values in some space \(X \subseteq \mathbb{R}^d\) and \(Y\) is a response variable with values in \(Y \subseteq \mathbb{R}\). Thus, \((x, y) \in X \times Y \subseteq \mathbb{R}^{d+1}\). Suppose that \(\Omega\) can be partitioned into \(G\) disjoint groups, say \(\Omega_1, \ldots, \Omega_G\), such that \(\Omega = \Omega_1 \cup \cdots \cup \Omega_G\). Cluster-weighted modelling (CWM; cf. Gershenfeld, 1997; Ingrassia et al., 2012) decomposes the joint probability as

\[
p(x, y; \vartheta) = \sum_{g=1}^{G} p(y|x, \Omega_g) p(x|\Omega_g) \pi_g,
\]

(4.1)

where \(p(y|x, \Omega_g)\) is the conditional density of the response variable \(Y\) given the predictor vector \(x\) and \(\Omega_g\), \(p(x|\Omega_g)\) is the probability density of \(x\) given \(\Omega_g\), \(\pi_g = p(\Omega_g)\) is the mixing proportion for sub-population \(g \in 1, \ldots, G\), so that \(\pi_g > 0\) and \(\sum_{g=1}^{G} \pi_g = 1\), and \(\vartheta\) denotes the set of all model parameters.
Hence, the joint density of \((X, Y)\) can be viewed as a mixture of local models \(p(y|x, \Omega_g)\) weighted (in a broader sense) on both the local densities \(p(x|\Omega_g)\) and mixing proportions \(\pi_g\). The posterior probability \(p(\Omega_g|x, y)\) that \((x, y)\) comes from the \(g\)th component, is given by

\[
p(\Omega_g|x, y) = \frac{p(x, y, \Omega_g)}{p(x, y)} = \frac{p(y|x, \Omega_g)p(x|\Omega_g)\pi_g}{\sum_{j=1}^{G} p(y|x, \Omega_j)p(x|\Omega_j)\pi_j},
\]

for \(g = \{1, \ldots, G\}\). In particular, the classification of each unit depends on both the marginal and conditional densities.

Traditionally, the local densities \(p(x|\Omega_g)\) are assumed to be multivariate Gaussian, so that \(X|\Omega_g \sim \mathcal{N}_d(\mu_g, \Sigma_g)\), for \(g = 1, \ldots, G\). Moreover, the conditional densities \(p(y|x, \Omega_g)\) are also modelled by Gaussian distributions with variance \(\sigma_{\varepsilon,g}^2\) around some deterministic function of \(x\), say \(\mu(x; \beta_g)\), so that the relationship between \(Y\) and \(X\) in the \(g\)th group can be written \(Y = \mu(x, \beta_g) + \varepsilon_g\), where \(\varepsilon_g \sim \mathcal{N}(0, \sigma_{\varepsilon,g}^2)\). Such a model will be referred to as the Gaussian cluster-weighted model:

\[
p(x, y; \theta) = \sum_{g=1}^{G} \phi(y|x; \mu(x; \beta_g), \sigma_g^2) \phi(x; \mu_g, \Sigma_g) \pi_g,
\]

where \(\phi(\cdot)\) denotes the multivariate Gaussian density. Consider the case where \(\mu(x; \beta_g) = b_g'x + b_{g0}\), with \(\beta = (b_g', b_{g0})'\), \(b_g \in \mathbb{R}^d\), and \(b_{g0} \in \mathbb{R}\), so that

\[
p(x, y; \theta) = \sum_{g=1}^{G} \phi(y; b_g'x + b_{g0}, \sigma_{\varepsilon,g}^2) \phi_d(x; \mu_g, \Sigma_g) \pi_g.
\]
This model will be referred to as the linear Gaussian cluster-weighted model.

4.1.2  The Factor Regression Model

In this section, we introduce the factor regression model that has gained much attention in the recent literature (West, 2003; Wang et al., 2007; Carvalho et al., 2008). A common assumption is that the relationship between $X$ and $Y$ can be written as $Y = \beta_0 + \beta'_1 X + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$ and $\beta = (\beta_0, \beta'_1)'$, with $\beta \in \mathbb{R}^{d+1}$. Now, assume a factor analysis model (Spearman, 1904; Bartlett, 1953) for $X$ so that $X = \mu + \Lambda U + e$, where $U \sim N_q(0, I_q)$ is a $q$-dimensional ($q < d$) vector of latent factors, $\Lambda$ is a $d \times q$ matrix of factor loadings, and $e \sim N_d(0, \Psi)$, with $\Psi = \text{diag}(\psi_1^2, \ldots, \psi_d^2)$, independent of $U$. Then $X \sim N_d(\mu, \Lambda \Lambda' + \Psi)$ and, conditional on $u$, results in $X|u \sim N(\mu + \Lambda u, \Psi)$.

In addition to the usual factor analysis assumptions (that is, $U$ and $e$ are independent), we also assume that $\varepsilon$ is independent of $U$ and $e$. Hence, the model $Y = \beta_0 + \beta'_1 X + \varepsilon$ with $X = \mu + \Lambda U + e$ gives

$$Y = \beta_0 + \beta'_1 (\mu + \Lambda U + e) + \varepsilon = (\beta_0 + \beta'_1 \mu) + \beta'_1 \Lambda U + (\beta'_1 e + \varepsilon), \quad (4.5)$$

which is known as the factor regression model. The mean and the variance of $Y$,
based on the independence assumptions, are given by

\[ E(Y) = \beta_0 + \beta'_1 \mu \]

\[ \text{Var}(Y) = \text{Var}(\beta'_1 \Lambda U) + \text{Var}(\beta'_1 e) + \text{Var}(\varepsilon) \]

\[ = \beta'_1 \Lambda \Lambda' \beta_1 + \beta' \Psi \beta_1 + \sigma^2_\varepsilon = \beta'_1 (\Lambda \Lambda' + \Psi) \beta_1 + \sigma^2_\varepsilon, \]

and so \( Y \sim N(\beta_0 + \beta'_1 \mu, \beta'_1 (\Lambda \Lambda' + \Psi) \beta_1 + \sigma^2_\varepsilon). \)

Consider the triplet \((X, U, Y)\). The mean is given by

\[
\mathbb{E} \begin{pmatrix} Y \\ X \\ U \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta'_1 \mu \\ \mu \\ 0 \end{pmatrix},
\]

and because \( \text{Cov}(X, Y) = (\Lambda \Lambda' + \Psi) \beta_1 \) and \( \text{Cov}(U, Y) = \Lambda' \beta_1 \),

\[
\text{Cov}(Y, X, U) = \begin{pmatrix} \beta'_1 \Sigma \beta_1 + \sigma^2_\varepsilon & \beta'_1 \Sigma & \beta'_1 \Lambda \\ \Sigma \beta_1 & \Sigma & \Lambda \\ \Lambda' \beta_1 & \Lambda' & I_q \end{pmatrix}, \tag{4.6}
\]

where \( \Sigma = \Lambda \Lambda' + \Psi \).

Now, we can write the joint density of \((X, U, Y)\) as

\[ p(x, u, y) = p(y|x, u)p(x|u)p(u). \]
We have shown that $X|u \sim N(\mu + \Lambda u, \Psi)$. Details on computation of the distribution of $Y|X, u$ are given in Appendix A.2.1. Of particular importance is that $E(Y|x, u) = E(Y|x)$ and $\text{Var}(Y|x, u) = \text{Var}(Y|x)$, and so $Y|x, u \sim N(\beta_0 + \beta_1' x, \sigma^2)$. This implies that $p(y|x, u) = p(y|x)$ and, therefore, $Y$ is conditionally independent of $U$ given $X = x$, so that the joint density of $(X, U, Y)$ can be written

$$p(x, u, y) = p(y|x, u)p(x|u)p(u) = p(y|x)p(x|u)p(u).$$

Similarly, $U|y, x \sim N(\gamma(x - \mu), I_q - \gamma \Lambda)$, where $\gamma = \Lambda'(\Lambda \Lambda' + \Psi)^{-1}$, and thus $U$ is conditionally independent on $Y$ given $X = x$. Therefore,

$$E[U|x; \mu, \Lambda, \Psi] = \gamma(x - \mu) \quad \text{and} \quad E[UU'|x; \mu, \Lambda, \Psi] = I_q - \gamma \Lambda + \gamma(x - \mu)(x - \mu)' \gamma'.$$

(4.7)

4.1.3 The Gaussian Cluster-Weighted Factor Analyzer

In this section we extend model (4.1) to include factor analyzers. Let $(X, Y)$ be a pair of a random $d$-dimensional vector $X$ and a random variable $Y$ defined on some space $\Omega = \Omega_1 \cup \cdots \cup \Omega_G$. Assume that for each group $\Omega_g$ ($g = 1, \ldots, G$), the pair $(X, Y)$ satisfies a factor regression model

$$Y = \beta_{g0} + \beta_{1g}' X + \varepsilon_g \quad \text{with} \quad X = \mu_g + \Lambda_g U_g + e_g,$$

(4.8)
where $\Lambda_g$ is a $d \times q$ matrix of factor loadings, $U_g \sim N(0, I_q)$ is the vector of factors, $e_g \sim N(0, \Psi_g)$ are the errors, where $\Psi_g = \text{diag}(\psi_{g1}, \ldots, \psi_{gd})$, and $\varepsilon_g \sim N(0, \sigma^2_g)$. Then the CW model introduced in (4.1) can be extended in order to include the underlying factor structure (4.8) for the $X$ variable. In particular, under the Gaussian assumptions given in the previous section (implying that $Y$ is conditionally independent of $U$ given $X = x$), we get

$$p(x, y; \theta) = \sum_{g=1}^G \phi(y|x; \beta_g, \sigma^2_g) \phi_d(x; \mu_g, \Sigma_g) \pi_g = \sum_{g=1}^G \phi(y|x; \beta_g, \sigma^2_g) \phi_d(x; \mu_g, \Lambda_g, \Psi_g) \pi_g,$$

(4.9)

where $\Sigma_g = \Lambda_g \Lambda'_g + \psi_g$ and $\theta$ denote the overall parameters. Model (4.9) will be referred to as the linear Gaussian Cluster-Weighted Factor Analyzers (CWFA) Model. The posterior probability (4.2) specializes as

$$p(\Omega_g|x, y) = \frac{\phi(y|x; \beta_{g0}, \sigma^2_{g0}) \phi_d(x|\mu_g, \Lambda_g, \Psi_g) \pi_g}{\sum_{j=1}^G \phi(y|x; \beta_j, \sigma^2_j) \phi_d(x|\mu_j, \Lambda_j, \Psi_j) \pi_j} \quad g = 1, \ldots, G.$$

(4.10)

**4.1.4 The Likelihood Function and the EM algorithm**

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a sample of size $n$ from density (4.9) and set $X = (x_1', \ldots, x_n')$, $Y = (y_1, \ldots, y_n)$. We assume that the pair $(x_i, y_i)$ ($i = 1, \ldots, n$) is the realization of

$$Y_i = \beta_{g0} + \beta_{1g}x_i + \varepsilon_{ig} \quad \text{and} \quad X_i = \mu_g + \Lambda_g U_{ig} + e_{ig} \quad \text{with probability } \pi_g, \quad g = 1, \ldots, G,$$
where \( \mathbf{U}_{ig} \sim N(\mathbf{0}, \mathbf{I}_q) \) and \( \mathbf{e}_{ig} \sim N(\mathbf{0}, \mathbf{\Psi}_g) \) so that, unconditionally, the density of \((\mathbf{X}_i, \mathbf{Y}_i)\) is

\[
p(\mathbf{x}_i, \mathbf{y}_i; \theta) = \sum_{g=1}^{G} \phi(y_i|\mathbf{x}_i; \beta_g, \sigma^2_g) \phi_d(\mathbf{x}_i; \mu_g, \Lambda_g, \mathbf{\Psi}_g) \pi_g. \tag{4.11}
\]

Moreover, we set \( n \) realizations of \( \mathbf{X} \) as \( \mathbf{X}^\prime \), the \((n \times d)\) data matrix, and \( n \) realizations of \( \mathbf{Y} \) as \( n \)-dimensional vector \( \mathbf{y} \). For given data \((\mathbf{X}, \mathbf{Y})\), parameters in (4.11) can be estimated according to the likelihood approach via the EM algorithm, where the likelihood function is given by:

\[
L(\theta; \mathbf{X}, \mathbf{y}) = \prod_{i=1}^{n} \left\{ \sum_{g=1}^{G} \phi(y_i|\mathbf{x}_i; \beta_g, \sigma^2_g) \phi_d(\mathbf{x}_i; \mu_g, \Sigma_g) \pi_g \right\},
\]

where we set \( \Sigma_g = \Lambda_g \Lambda_g' + \Psi_g \) (\( g = 1, \ldots, G \)). Consider the augmented data \( \{(\mathbf{x}_i, \mathbf{y}_i, \mathbf{u}_{ig}, \mathbf{z}_i), i = 1, \ldots, n\} \), where \( \mathbf{z}_i = (z_{i1}, \ldots, z_{ig})' \), with \( z_{ig} = 1 \) if \((\mathbf{x}_i, \mathbf{y}_i)\) comes from the \( g \)-th population and \( z_{ig} = 0 \) otherwise. Then, the complete-data likelihood function can be written in the form:

\[
L_c(\theta; \mathbf{X}, \mathbf{y}) = \prod_{i=1}^{n} \prod_{g=1}^{G} \left[ \phi(y_i|\mathbf{x}_i; \beta_g, \sigma^2_g) \phi_d(\mathbf{x}_i|\mathbf{u}_i; \mu_g, \Lambda_g, \mathbf{\Psi}_g) \phi_d(\mathbf{u}_i; \mu_{ig}, \Lambda_g, \mathbf{\Psi}_g) \pi_g \right]^{z_{ig}}. \tag{4.12}
\]

In particular, here we use the alternating expectation-conditional maximization (AECM) algorithm due to the factor structure of the model (see Meng and van Dyk (1997)). This extension of the EM algorithm uses different specifications of missing data at
each stage. The idea is to partition $\theta = (\theta_1', \theta_2')'$ in such a way that $L(\theta; X, y)$ is easy to maximize for $\theta_1$ given $\theta_2$ and vice versa. The application of the AECM algorithm consists of two cycles, with an E-step and a CM-step for each cycle. The two CM-steps correspond to the partition of $\theta$ into $\theta_1$ and $\theta_2$. Then, we can iterate between these two conditional maximizations until convergence. In the first stage, $\theta_1 = \{(\pi_g, \beta_g, \mu_g, \sigma^2_g), g = 1, \ldots, G\}$ and the missing data are the unobserved group labels $Z = (z_1', \ldots, z_n')$. In the second stage, $\theta_2 = \{(\Lambda_g, \Psi_g), g = 1, \ldots, G\}$ and the missing data are the group labels $Z$ and the unobserved latent factors $\bar{U} = (U_{11}, \ldots, U_{nG})$. Details are given in Appendix A.2.1.

Hence, the maximum likelihood estimates for $\Lambda$ and $\Psi$ are obtained by iteratively computing

$$
\Lambda^+_g = S^{(k+1)}_g \gamma_g'(\Theta^{-1}) \\
\Psi^+_g = \text{diag} \left\{ S^{(k+1)}_g - \Lambda^+_g \gamma_g' S^{(k+1)}_g \right\}
$$

(4.13)

(4.14)

where the superscript $^+$ denotes the update estimate and $S^{(k+1)}_g = (1/n_g^{(k+1)}) \sum_{i=1}^n z_i^{(k+1)} (x_i - \mu^{(k+1)}_g)(x_i - \mu^{(k+1)}_g)'$. Now, using (A.9) and (A.10), we get

$$
\gamma^+_g = \Lambda^+_g \Lambda^+_g + \Psi^+_g)^{-1}
$$

(4.15)

$$
\Theta^+_g = I_q - \gamma^+_g \Lambda^+_g + \gamma^+_g S^{(k+1)}_g \gamma^+_g
$$

(4.16)

**Algorithm 1.** In summary, the procedure can be described as follows. For a given
initial guess \( \theta^{(0)} \), on the \( k + 1 \)-th iteration, the algorithm carries out the following steps for \( g = 1, \ldots, G \):

1. Compute \( \pi_g^{(k+1)}, \mu_g^{(k+1)}, \beta_g^{(k+1)}, \sigma_g^{2(k+1)} \);

2. Set \( \Lambda_g \leftarrow \Lambda_g^{(k)} \) and \( \Psi \leftarrow \Psi_g^{(k)} \), compute \( \gamma_g \) and \( \Theta_g \);

3. Repeat the following steps until convergence on \( \Lambda_g \) and \( \Psi_g \):
   
   (a) Set \( \Lambda_g^+ \leftarrow S_g^{(k+1)} \gamma_g (\Theta_g)^{-1} \) and \( \Psi_g^+ \leftarrow \text{diag}\left\{ S_g^{(k+1)} - \Lambda_g^+ \gamma_g S_g^{(k+1)} \right\} \);
   
   (b) Set \( \gamma_g^+ \leftarrow \Lambda_g^+ (\Lambda_g^+ \Lambda_g^+ + \Psi_g^+)^{-1} \) and \( \Theta_g^+ \leftarrow I_q - \gamma_g^+ \Lambda_g^+ + \gamma_g^+ S_g^{(k+1)} \gamma_g^+ \);
   
   (c) Set \( \Lambda_g \leftarrow \Lambda_g^+ \), \( \Psi_g \leftarrow \Psi_g^+ \), \( \gamma_g \leftarrow \gamma_g^+ \) and \( \Theta_g \leftarrow \Theta_g^+ \).

### 4.1.5 A Set of Parsimonious Covariance Structures

In this section we propose a set of sixteen parsimonious CWFA models that are a combination of constraints on the covariance structure of the input variable \( X \) (in the framework of McNicholas and Murphy (2008)) and a constraint on the variance term of the response variable \( Y \).

**Constraints on the input variable \( X \)**

Extending results from McNicholas and Murphy (2008) we can consider the following constraints on the covariance structure, for \( g = 1, \ldots, G \): i) equal loading matrices \( \Lambda_g = \Lambda \), ii) equal error variance \( \Psi_g = \Psi \), and iii) isotropic assumption: \( \Psi_g = \psi_g I_d \). In such cases, the \( g \)-th term of the expected complete-data log-likelihood \( Q_2(\theta; \theta^{(k+1/2)}) \) and the estimates (A.16) and (A.17) in Algorithm 1 are computed as follows.
Isotropic assumption: \( \Psi_g = \psi_g I_d \). In this case, Equation (A.11) can be written as

\[
Q_2(\Lambda_g, \psi_g; \theta^{(k+1/2)}) = C(\theta_1^{(k+1)}) + \frac{1}{2} n_g^{(k+1)} \ln |\psi_g^{-1} I_p| - \frac{1}{2} n_g^{(k+1)} \psi_g^{-1} \text{tr} \left\{ S_g^{(k+1)} \right\} + n_g^{(k+1)} \psi_g^{-1} \text{tr} \left\{ \gamma_g^{(k)} S_g^{(k+1)} \Lambda_g \right\} - \frac{1}{2} n_g^{(k+1)} \psi_g^{-1} \text{tr} \left\{ \Lambda_g \Theta_g^{(k)} \Lambda'_g \right\},
\]

yielding

\[
\frac{\partial Q_2}{\partial \psi_g^{-1}} = \frac{1}{2} n_g^{(k+1)} \left[ p \psi_g - \text{tr} \left\{ S_g^{(k+1)} \right\} + 2 \text{tr} \left\{ \gamma_g^{(k)} S_g^{(k+1)} \Lambda_g \right\} - \text{tr} \left\{ \Lambda_g \Theta_g^{(k)} \Lambda'_g \right\} \right].
\]

(4.17)

Then the estimated \( \hat{\psi}_g \) is obtained by satisfying

\[
\frac{\partial Q_2}{\partial \psi_g^{-1}} = 0 \quad \Rightarrow \quad p \psi_g - \text{tr} \left\{ S_g^{(k+1)} \right\} + 2 \text{tr} \left\{ \gamma_g^{(k)} S_g^{(k+1)} \Lambda_g \right\} - \text{tr} \left\{ \Lambda_g \Theta_g^{(k)} \Lambda'_g \right\} = 0.
\]

Thus, according to (A.16), for \( \Lambda_g = \hat{\Lambda}_g = S_g^{(k+1)} \gamma_g^{(k)} \Theta_g^{-1} \) we get \( \text{tr} \left\{ \Lambda_g \Theta_g^{(k)} \Lambda'_g \right\} = \text{tr} \left\{ \gamma_g^{(k)} S_g^{(k+1)} \Lambda_g \right\} \) and finally

\[
\hat{\psi}_g = \frac{1}{p} \text{tr} \left\{ S_g^{(k+1)} - \hat{\Lambda}_g \gamma_g^{(k)} S_g^{(k+1)} \right\}.
\]

(4.18)
Thus,

\[ \psi_g^+ = \frac{1}{p} \text{tr}\left\{ S_g^{(k+1)} - \Lambda_g \gamma_g^+ S_g^{(k+1)} \right\} \tag{4.19} \]

\[ \gamma_g^+ = \Lambda_g' (\Lambda_g \Lambda_g' + \psi_g^+ I_d)^{-1}. \tag{4.20} \]

Note that \( \Theta_g^+ \) is computed according to (4.16).

**Equal error variance: \( \Psi_g = \Psi \).** In this case, Equation (A.11) can be written as

\[
Q_2(\Lambda_g, \Psi; \theta^{(k+1/2)}) = C(\theta_1^{(k+1)}) - \frac{1}{2} n_g^{(k+1)} \ln |\Psi| - \frac{1}{2} n_g^{(k+1)} \text{tr} \left\{ S_g^{(k+1)} \Psi^{-1} \right\} \\
+ n_g^{(k+1)} \text{tr} \left\{ \Lambda_g \gamma_g^{(k)} S_g^{(k+1)} \Psi^{-1} \right\} - \frac{1}{2} n_g^{(k+1)} \text{tr} \left\{ \Lambda_g' \Psi^{-1} \Lambda_g \Theta_g^{(k)} \right\},
\]

yielding

\[
\frac{\partial Q_2(\Lambda_g, \Psi; \theta^{(k+1/2)})}{\partial \Psi^{-1}} = \frac{1}{2} n_g^{(k+1)} \Psi - \frac{1}{2} n_g^{(k+1)} S_g^{(k+1)} + n_g^{(k+1)} S_g^{(k+1)} \gamma_g^{(k)} \Lambda_g' \\
- \frac{1}{2} n_g^{(k+1)} \Lambda_g \Theta_g^{(k)} \Lambda_g'.
\]

Then the estimated \( \hat{\Psi} \) is obtained by satisfying

\[
\sum_{g=1}^{G} \frac{\partial Q_2(\Lambda_g, \Psi; \theta^{(k+1/2)})}{\partial \Psi^{-1}} = 0,
\]
that is

\[
\frac{n}{2} \Psi - \frac{1}{2} \sum_{g=1}^{G} n_{g}^{(k+1)} S_{g}^{(k+1)} + \sum_{g=1}^{G} n_{g}^{(k+1)} S_{g}^{(k+1)'} \gamma_{g}^{(k)'} \Lambda_{g}' - \frac{1}{2} \sum_{g=1}^{G} n_{g}^{(k+1)} \Lambda_{g} \Theta_{g}^{(k)} \Lambda_{g}' = \\
\frac{n}{2} \Psi - \frac{1}{2} \sum_{g=1}^{G} n_{g}^{(k+1)} \left[ S_{g}^{(k+1)} + 2 S_{g}^{(k+1)'} \gamma_{g}^{(k)'} \Lambda_{g}' - \Lambda_{g} \Theta_{g}^{(k)} \Lambda_{g}' \right] = 0,
\]

with \( \sum_{g=1}^{G} n_{g}^{(k+1)} = n \). Again, according to (A.16), for \( \Lambda_{g} = \hat{\Lambda}_{g} = S_{g}^{(k+1)'} \gamma_{g}^{(k)'} \Theta_{g}^{-1} \), we get \( \hat{\Lambda}_{g} \Theta_{g}^{(k)} \Lambda_{g}' = \hat{\Lambda}_{g} \gamma_{g}^{(k)} S_{g}^{(k+1)} \) and, afterwards,

\[
\hat{\psi} = \sum_{g=1}^{G} \frac{n_{g}}{n} \text{diag} \{ S_{g}^{(k+1)} - \hat{\Lambda}_{g} \gamma_{g}^{(k)'} S_{g}^{(k+1)} \} = \sum_{g=1}^{G} \pi_{g}^{(k+1)} \text{diag} \{ S_{g}^{(k+1)} - \hat{\Lambda}_{g} \gamma_{g}^{(k)'} S_{g}^{(k+1)} \}.
\]

(4.21)

Thus,

\[
\Psi^+ = \sum_{g=1}^{G} \pi_{g}^{(k+1)} \text{diag} \{ S_{g}^{(k+1)} - \Lambda_{g}^{+} \gamma_{g} S_{g}^{(k+1)} \} 
\]

(4.22)

\[
\gamma_{g}^{+} = \Lambda_{g}' \left( \Lambda_{g}^{+} \Lambda_{g}' + \Psi^+ \right)^{-1}.
\]

(4.23)

where \( \Theta_{g}^{+} \) is computed according to (4.16).
Equal loading matrices: $\Lambda_g = \Lambda$. In this case, Equation (A.11) can be written as

$$Q_2(\Lambda, \Psi_g, \theta^{(k+1/2)}) = C(\theta_1^{(k+1)}) + \frac{1}{2} n_g^{(k+1)} \ln |\Psi_g^{-1}| - \frac{1}{2} n_g^{(k+1)} \text{tr} \left\{ S_g^{(k+1)} \Psi_g^{-1} \right\}$$

$$+ n_g^{(k+1)} \text{tr} \left\{ \Lambda \gamma_g^{(k)} S_g^{(k+1)} \Psi_g^{-1} \right\} - \frac{1}{2} n_g^{(k+1)} \text{tr} \left\{ \Lambda' \Psi_g^{-1} \Lambda \Theta_g^{(k)} \right\},$$

(4.24)

yielding

$$\frac{\partial Q_2(\Lambda, \Psi_g, \theta^{(k+1/2)})}{\partial \Lambda} = n_g^{(k+1)} \Psi_g^{-1} S_g^{(k+1)} \gamma_g^{(k)} - n_g^{(k+1)} \Psi_g^{-1} \Lambda \Theta_g^{(k)} = 0.$$

Then the estimated $\hat{\Lambda}$ is obtained by satisfying

$$\sum_{g=1}^{G} \frac{\partial Q_2(\Lambda, \Psi_g; \theta^{(k+1/2)})}{\partial \Lambda} = \sum_{g=1}^{G} n_g^{(k+1)} \Psi_g^{-1} \left[ S_g^{(k+1)} \gamma_g^{(k)} - \Lambda \Theta_g^{(k)} \right] = 0.$$  

(4.25)

with $\gamma_g^{(k)} = \Lambda' \gamma_g^{(k)} = \Lambda \Lambda' + \Psi_g^{(k)}$. In this case, the loading matrix cannot be solved directly and must be solved in a row-by-row manner as in McNicholas and Murphy (2008). Therefore,

$$\lambda_i^+ = r_i \left( \sum_{g=1}^{G} \frac{n_g}{\psi_{g(i)}} \Theta_g \right)^{-1}$$

(4.26)

$$\gamma_g^+ = \Lambda' (\Lambda^+ \Lambda'^+ + \Psi_g^+)^{-1}$$

(4.27)

$$\Theta_g^+ = I_q - \gamma_g^+ \Lambda^+ + \gamma_g^+ S_g^{(k+1)} \gamma_g'^+,$$

(4.28)
where $\lambda_i^+$ is the $i$th row of the matrix $\Lambda^+$, $\psi_{g(i)}$ is the $i$th diagonal element of $\Psi_g$, and $r_i$ represents the $i$th row of the matrix $\sum_{g=1}^{G} \sigma_{g}^{(k+1)} S_g^{(k+1)}$.

By combining these constraints we can obtain a set of eight models, as described in Table 4.1, where “constrained” means $\Lambda_g = \Lambda$ in the loading matrix term and $\Psi_g = \Psi$ in the error variance term, $\Psi_g = \psi_g I_d$ means isotropic assumption.

Table 4.1: Covariance structure for the $X$ variable where U refers to unconstrained and C refers to constrained.

<table>
<thead>
<tr>
<th>Covariance Structure</th>
<th>Loading Matrix</th>
<th>Error Variance</th>
<th>Isotropic</th>
<th>$\Lambda_g^+$</th>
<th>$\Psi_g^+$</th>
<th>$\gamma_g^+$</th>
<th>$\Theta_g^+$</th>
</tr>
</thead>
<tbody>
<tr>
<td>UUU</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>(4.13)</td>
<td>(4.14)</td>
<td>(4.15)</td>
<td>(4.16)</td>
</tr>
<tr>
<td>UUC</td>
<td>U</td>
<td>U</td>
<td>C</td>
<td>(4.13)</td>
<td>(4.19)</td>
<td>(4.20)</td>
<td>(4.16)</td>
</tr>
<tr>
<td>UCU</td>
<td>U</td>
<td>C</td>
<td>U</td>
<td>(4.13)</td>
<td>(4.22)</td>
<td>(4.23)</td>
<td>(4.16)</td>
</tr>
<tr>
<td>UCC</td>
<td>U</td>
<td>C</td>
<td>C</td>
<td>(4.13)</td>
<td>(A.19)</td>
<td>(A.20)</td>
<td>(4.16)</td>
</tr>
<tr>
<td>CUU</td>
<td>C</td>
<td>U</td>
<td>U</td>
<td>(4.26)</td>
<td>(A.21)</td>
<td>(4.27)</td>
<td>(4.28)</td>
</tr>
<tr>
<td>CCU</td>
<td>C</td>
<td>C</td>
<td>U</td>
<td>(A.27)</td>
<td>(A.28)</td>
<td>(A.29)</td>
<td>(4.28)</td>
</tr>
<tr>
<td>CCC</td>
<td>C</td>
<td>U</td>
<td>U</td>
<td>(A.27)</td>
<td>(A.30)</td>
<td>(A.31)</td>
<td>(4.28)</td>
</tr>
</tbody>
</table>

**Constraint on the Response Variable $Y$**

For the response variable $Y$, according to (4.8), the error variance terms $\sigma_g^2$ can be constrained to be equal across groups, i.e., $\sigma_g^2 = \sigma^2$ for $g = 1, \ldots, G$, and then (A.5) yields

$$
\sigma^2(k+1) = \frac{1}{n} \sum_{i=1}^{n} \sum_{g=1}^{G} s_{ig}^{(k+1)} \left\{ y_i - (\beta_{0g}^{(k+1)} + \beta_{1g}^{(k+1)} x_i) \right\}^2.
$$

(4.29)
Hence, we have a family of 16 models as a combination of the constraints on $Y$ and $X$ to (4.29) and to the eight models listed in Table 4.1. For example, the model CUCU means equal $\sigma^2_g$ for $Y$ and covariance structure UCU for $X$.

### 4.1.6 Semi-supervised Classification via CWFA models

Consider the classification scenario where there are $n$ observations, $k$ of which have known group memberships. Within the model-based classification framework, we use all $n$ observations to estimate the group memberships for the $n - k$ observations with unknown group memberships. We let $Z_{ig}$ denote the component membership of observation $i$, where $Z_{ig} = 1$ if observation $i$ belongs to component $g$ and $Z_{ig} = 0$ otherwise. Then, arranging the data so that it is the first $k$ observations that have known group memberships, the likelihood (4.12) is

$$ L_c(\theta; X, y) = \prod_{i=1}^{k} \prod_{g=1}^{G} \left[ \phi \left(y_i | x_i, \beta_g, \sigma^2_{\epsilon,g} \right) \phi_d \left( x_i | u_i; \mu_g, \Lambda_g, \Psi_g \right) \phi_q(u_{ig}) \pi_g \right]^{z_{ig}} \cdot \prod_{j=k+1}^{n} \prod_{g=1}^{G} \left[ \phi \left(y_i | x_i, \beta_g, \sigma^2_{\epsilon,g} \right) \phi_d \left( x_i | u_i; \mu_g, \Lambda_g, \Psi_g \right) \phi_q(u_{ig}) \pi_g \right]. \quad (4.30) $$

Parameter estimation under semi-supervised classification is obtained by jointly modeling data with known and unknown group memberships and maximizing the complete data log-likelihood. These parameters are then used to estimate the unknown group memberships.
4.1.7 Initialization

The choice of starting values is a well known and important issue with respect to EM-based algorithms. The standard approach consists of selecting a value for $\theta^{(0)}$. An alternative method consists of choosing a value for $z^{(0)}_i, i = 1, \ldots, n$ (see McLachlan and Peel, 2000a, p. 54). Within this approach, and due to the hierarchical structure of the CWFA family of models, we propose a 5-step hierarchical (5RH) initialization procedure that improves upon the technique adopted by McNicholas and Murphy (2008, p. 288) for their parsimonious family of mixtures of factor analyzers.

For a fixed number of groups $G$, let $z^{(0)}_i, i = 1, \ldots, n$ denote the initial classification for the EM algorithm, so that $z^{(0)}_{ig} \in \{0, 1\}$ and $\sum_g z^{(0)}_{ig} = 1$. The set $\{z^{(0)}_i, i = 1, \ldots, n\}$ can be obtained through some clustering procedure (here we considered the $k$-means method but the same procedure applies for random initialization, e.g., sampling from a multinomial distribution).

Then, at the first step of the 5RH procedure, the most constrained CCCC model is run from these starting values. At the second step, the resulting (AECM-estimated) $\hat{z}_{ig}$ are then taken as the starting group membership labels for the four models (UCCC, CUCC, CCUC, CCCU) obtained by relaxing one of the four constraints. At the third step, each of the six models with two constraints (CCUU, CUCU, UCCU, UCUC, UCUC, UUCC) is then initialized using the $\hat{z}_{ig}$ from the previous step of the model with the highest likelihood. To initialize CCUU, for example, we used the $\hat{z}_{ig}$ from the model having the highest likelihood between CCCU and CCUC. In this fashion, the 5RH procedure continues according to the scheme
displayed in Figure 4.1, until the less constrained model UUUU is estimated at the fifth step.

Figure 4.1: Relationships among the models in the 5-step random-hierarchical initialization procedure. Arrows are oriented from the model used to initialize to the model to be estimated.

For all the models in the CWFA family, in analogy with McNicholas and Murphy (2008), the initial values for the elements of $\Lambda_g$ and $\Psi_g$ are generated from the eigen-decomposition of $S_g$ as follows. The $S_g$ are computed based on the values of $z_{ig}^{(0)}$. The eigen-decomposition of each $S_g$ is obtained using the function `eigen` in R software. Then the initial values of the elements of $\Lambda_g$ are set as $\lambda_{ij} = \sqrt{d_j \rho_{ij}}$, where $d_j$ is the $j$th largest eigenvalue of $S_g$ and $\rho_{ij}$ is the $i$th element of the eigenvector corresponding to the $j$th largest eigenvalue of $S_g$, where $i \in \{1, 2, \ldots, d\}$ and $j \in \{1, 2, \ldots, q\}$. The $\Psi_g$ are then initialized as $\Psi_g = \text{diag}(S_g - \Lambda_g \Lambda'_g)$. 
4.1.8 Convergence Criterion and Model Selection

The Aitken acceleration procedure (Aitken, 1926) is used to estimate the asymptotic maximum of the log-likelihood at each iteration of the AECM algorithm. Based on this estimate, a decision can be made whether or not the algorithm has reached convergence; that is, if the log-likelihood is sufficiently close to its estimated asymptotic value or not. The Aitken acceleration at iteration $k$ is given by

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

where $l^{(k+1)}$, $l^{(k)}$, and $l^{(k-1)}$ are the log-likelihood values from iterations $k + 1$, $k$, and $k - 1$, respectively. Then the asymptotic estimate of the log-likelihood at iteration $k + 1$ is given by

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

(Böhning et al., 1994). In the analyses in Section 4.2, we follow McNicholas (2010) and stop our algorithms when $l_{\infty}^{(k+1)} - l^{(k)} < \epsilon$ (we use $\epsilon = 0.05$).

Several model selection criteria are used in the literature, such as the Bayesian information criterion (BIC; Schwarz, 1978), the integrated completed likelihood (ICL; Biernacki et al., 2000), and the Akaike information criterion (AIC; Sakamoto et al., 1986). Among these, the BIC is the most predominant in the literature and is given by

$$\text{BIC} = 2 \ln L - k \ln(n),$$
where $L$ is the likelihood of the data, $n$ is the sample size, and $k$ is the number of free parameters.

The performance of the algorithm is assessed using the adjusted Rand index (ARI; Hubert and Arabie, 1985). The Rand index (Rand, 1971) is based on the pairwise agreement between the predicted and true classifications. The ARI corrects the Rand index to account for agreement by chance: a value of ‘1’ indicates perfect agreement, ‘0’ indicates random classification, and negative values indicate a classification that is worse than would be expected by guessing.

### 4.2 Data Analyses

In this section we present numerical results based on both simulated and real data.

#### 4.2.1 Simulated data

Numerical analyses based on simulated data are presented. These assist in evaluating classification properties of the proposed models.

**Example 1.** The first data set consisted of a sample of size $n = 175$ drawn from model (4.9) with $G = 2$ groups ($n_1 = 75, n_2 = 100$) and $q = 2$ latent factors according to a covariance structure UUCU (see Figure 4.2 for details). The parameters used for simulation of the data are given in Table 4.2.

All sixteen CWFA models were fitted to the data for $G = 2, 3$ and $q = 1, 2$, resulting in a total of 64 models. Initialization of the $z$ matrix for the most constrained
model (CCCC) for each combination of $G$ and $q$ was done using $k$-means. The remaining 15 models for each $G$ and $q$ were initialized using the 5-step hierarchical initialization procedure. The BIC values for all models were computed and the model with the largest BIC value was selected as the best model. In this example, the model corresponding to the largest BIC value (-5845.997) was a two component UUCU model with two latent factors, same as the model used to generate the data. The model selected by the BIC, UUCU ($G = 2, q = 2$), gave a perfect classification and the
estimated parameters were very close to the parameters used for data simulation (Table 4.2, Appendix A.2.2).

Table 4.2: Estimated parameters along with true parameters

(a) True and estimated $\mu_g$

<table>
<thead>
<tr>
<th>$G$</th>
<th>$(14, 18, 25, 14, 22)'$</th>
<th>$(15.88, 19.94, 27.48, 15.81, 23.93)'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(14, 18, 25, 14, 22)'</td>
<td>(15.88, 19.94, 27.48, 15.81, 23.93)'</td>
</tr>
<tr>
<td>2</td>
<td>$(-12, -10, -22, -20, -22)'$</td>
<td>$(-11.95, -10.36, -22.00, -19.67, -22.03)'$</td>
</tr>
</tbody>
</table>

(b) True and estimated $\beta_{1,g}$

<table>
<thead>
<tr>
<th>$G$</th>
<th>$(0.47, 0.02, 0.42, 0.03, 0.87)'$</th>
<th>$(0.50, 0.03, 0.46, 0.02, 0.81)'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$(0.47, 0.02, 0.42, 0.03, 0.87)'$</td>
<td>$(0.50, 0.03, 0.46, 0.02, 0.81)'$</td>
</tr>
<tr>
<td>2</td>
<td>$(-0.02, -0.63, -0.05, -0.85, -0.03)'$</td>
<td>$(-0.04, -0.57, -0.01, -0.85, -0.18)'$</td>
</tr>
</tbody>
</table>

(c) True and estimated $\sigma_g$

<table>
<thead>
<tr>
<th>$G$</th>
<th>$\sigma_g$</th>
<th>$\hat{\sigma}_g$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>1.24</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>3.79</td>
</tr>
</tbody>
</table>

(d) True and estimated $\beta_{0,g}$

<table>
<thead>
<tr>
<th>$G$</th>
<th>$\beta_{0,g}$</th>
<th>$\hat{\beta}_{0,g}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.5</td>
<td>4.34</td>
</tr>
<tr>
<td>2</td>
<td>-4.2</td>
<td>-6.35</td>
</tr>
</tbody>
</table>

Figure 4.3 shows the BIC values of all 64 models sorted in an increasing order, where number denotes the corresponding estimated number of groups $G$ and color denotes the corresponding number of latent variables $q$. The horizontal line separates the models with a BIC value within 1% of the maximum value (the 1% line); that is, the set of models having a BIC value $\xi$ such that

$$\frac{\text{BIC}_{\text{max}} - \xi}{\text{BIC}_{\text{max}}} \leq 0.01,$$

where $\text{BIC}_{\text{max}}$ denotes the largest obtained BIC value. This graphical representation will be referred to as the group-factor plot of BIC values.

As mentioned earlier, the model with the largest BIC value is UUCU ($G = 2, q = 2$). The subsequent two models that were within 1% were UUUU ($G = 2, q = 2$)
Figure 4.3: Example 1: group-factor plot of BIC values ordered in an increasing order. Numbers denote the estimated number of groups and colors denote the number of latent factors (red: $q = 1$, blue: $q = 2$). The green point denotes the ‘true’ model.
with a BIC of -5867.006 and CUCU \((G = 2, q = 2)\) with a BIC of -5869.839. These two models are structurally very close to the UUCU model and also yielded perfect classification. It should also be noted that most of the models with high BIC values have two components and two latent factors.

**Example 2.** The second data set concerned a sample of size \(n = 235\) drawn from model (4.9) with \(G = 3\) groups \((n_1 = 75, n_2 = 100, \text{ and } n_3 = 60)\) and \(q = 2\) latent factors according to covariance structure CUUC (see Figure 4.4 for details).

![Figure 4.4: Example 2: scatterplot matrix of data.](image-url)
All sixteen CWFA models were fitted to the data for \( G = 2, 3, 4 \) and \( q = 1, 2 \), resulting in 96 different models. The algorithm was initialized the same way as for Example 1. The model with the highest BIC was CUUC \((G = 3, q = 2)\) with a BIC of -6579.116 and a perfect classification. The estimated parameters of the CUUC \((G = 3, q = 2)\) were very close to the true parameters (Table 4.3 and Appendix A.2.2).

Table 4.3: Estimated parameters along with true parameters

<table>
<thead>
<tr>
<th>( G )</th>
<th>( \mu_{g} )</th>
<th>( \hat{\mu}_{g} )</th>
<th>( \beta_{1,g} )</th>
<th>( \hat{\beta}_{1,g} )</th>
<th>( \sigma_{g} )</th>
<th>( \hat{\sigma}_{g} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(0, 0, -5, 0, -4)'</td>
<td>(0.82, 0.48, -5.09, -0.21, -3.75)'</td>
<td>(-0.41, -0.87, -0.22, -0.62, -0.06)'</td>
<td>(-0.34, -0.82, -0.32, -0.66, -0.09)'</td>
<td>2.30</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>(14, 18, 25, 14, 22)'</td>
<td>(13.64, 17.44, 25.44, 14.25, 21.44 )'</td>
<td>(0.47, 0.02, 0.42, 0.03, 0.87)'</td>
<td>(0.51, 0.00, 0.38, 0.05, 0.84)'</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>(-12, -10, -22, -20, -22)'</td>
<td>(-12.33, -10.22, -22.25, -20.24, -22.21)'</td>
<td>(-0.02, -0.63, -0.05, -0.85, -0.03)'</td>
<td>(-0.04 -0.68, -0.36, -0.44, -0.18)'</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

The group-factor plot of BIC values (Figure 4.5) shows the BIC values corresponding to all 96 models, sorted in increasing order; the horizontal line is the 1% line. The other three models with BIC values within 1% of the largest BIC value are UUUC \((G = 3, q = 2)\), CUUU \((G = 3, q = 2)\), and UUUU \((G = 3, q = 2)\) with a BIC values of -6583.692, -6637.222, and -6641.798 respectively. Again in this example, the models with BIC values within 1% of the maximum BIC had three components..
Figure 4.5: Example 1: group-factor plot of BIC values ordered in increasing order. Numbers denote the estimated number of groups and colors denote the number of latent factor (red: \( q = 1 \), blue: \( q = 2 \)). The green point denotes the “true” model.

and two latent factors. They had a covariance structure more similar to the “true” covariance structure and also yielded perfect classification.

### 4.2.2 Real data

In addition to the simulated data analyses of Section 4.2.1, the CWFA family of models was also applied to a real data set for both clustering and classification purposes.

The \texttt{f.voles} data set, detailed in Flury (1997, Table 5.3.7) and available in
the Flury package for R, consists of measurements of female voles from two species, *M. californicus* and *M. ochrogaster*. The data consist of 86 observations for which we have a binary variable Species denoting the species (45 *M. ochrogaster* and 41 *M. californicus*), a variable Age measured in days, and six remaining variables related to skull measurements. The names of the variables are the same as in the original analysis of this data set by Airoldi and Hoffmann (1984a): L$_2$ = condylo-incisive length, L$_9$ = length of incisive foramen, L$_7$ = alveolar length of upper molar tooth row, B$_3$ = zygomatic width, B$_4$ = interorbital width, and H$_1$ = skull height. All of the variables related to the skull are measured in units of 0.1 mm.

The original purpose of Airoldi and Hoffmann (1984a) was to study age variation in *M. californicus* and *M. ochrogaster* and to predict age on the basis of the skull measurements. For our purpose, we assume the data are unlabelled with respect to Species and are interested in evaluating clustering and classification using the CWFA family of models as well as comparing the algorithm with some well-established mixture-based techniques. Therefore, age can be considered the natural Y-variable and the skull measurements can be considered the X-variable for the CWFA framework.

**Clustering**

All sixteen models were fitted assuming no known group membership for $G = 2, \ldots, 5$ components and $q = 1, 2, 3$ factors, resulting in total of 192 different models. The model with the largest BIC value was CCCU ($G = 3, q = 1$), with a BIC of -3837.698 and an ARI of 0.72. The classifications resulting from the CCCU
(G=3, q=1) model are given in Table 4.4(a).

Table 4.4: Classification of f. voles data using different clustering approaches.

<table>
<thead>
<tr>
<th></th>
<th>(a) CWFA</th>
<th>(b) PGMM</th>
<th>(c) Mclust</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRUE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Est.</td>
<td></td>
<td>Est.</td>
</tr>
<tr>
<td></td>
<td>1 2 3</td>
<td></td>
<td>1 2</td>
</tr>
<tr>
<td>ochrogaster</td>
<td>24 21 -</td>
<td></td>
<td>ochrogaster</td>
</tr>
<tr>
<td>californicus</td>
<td>- - 41</td>
<td></td>
<td>californicus</td>
</tr>
</tbody>
</table>

As seen from Table 4.4(a), 4.4(b), and 4.4(c), *M. californicus* was classified correctly using all three approaches. However, *M. ochrogaster* was classified into two sub-clusters using CWFA and parsimonious Gaussian mixture models (PGMM) while Mclust classified it into one cluster. Also, the CWFA approach had no misclassifications between the two species but both PGMM and Mclust misclassified two *M. ochrogaster* as *M. californicus*. We evaluated the group-factor plot using the BIC values of all 192 models.

Among the 192 models, 10 models had a BIC within 1% of the largest BIC, of which five were three component models and five were two component models. However, from Figure 4.6, the top four models all had three components, which shows that a three component model was not randomly chosen. Airoldi and Hoffmann (1984b) mention that some unexplained geographic variation may exist among the voles. However, no covariate was available with such information. Hence, we opted for the matrix scatter plot (see Figure 4.7) to evaluate the presence of sub-clusters.
From Figure 4.7, the scatter plot of the variables $B_3$ vs. $B_4$ shows the presence of distinct sub-clusters for *M. ochrogaster*, which supports our results attained using CWFA modelling.

**Classification**

In a classification framework, a portion of the data has a known group membership. In our case, we assumed that 50% of the data had a known group
Figure 4.7: Scatterplot matrix of \texttt{f_voles} data showing the classification observed using CWFA using clustering framework where black and red indicates sub-cluster of \textit{M. ochrogaster} species and green indicates \textit{Microtus californicus} species.

membership. A subset of observations consisting of 50% of the data was randomly selected and these observations were assumed to have a known group membership. To allow for the unobserved sub-cluster, we ran the algorithm for $G = 2, 3$ and $q = 1, 2, 3$. The best model (CCUU; $G = 2, q = 1$) selected by the BIC gave a perfect classification (see Table 4.5(a)) and had a BIC value of -3843.482.

We also ran the classification assuming that the data actually comprise three known groups. Therefore, using the classification observed by clustering, we ran the classification algorithm with 50% known and 50% unknowns. To further allow for the
unobserved sub-cluster, we ran the algorithm for \( G = 3, 4 \) and \( q = 1, 2, 3 \). The model selected using the BIC was CCCU \((G = 3, q = 1)\) with a BIC value of -3837.383. Even though the BIC value observed using the classification approach (with three known groups membership) was very close to the BIC value using clustering, the sub-clusters do not have precisely the same classification using classification and clustering. This could be a consequence of the classification of borderline observations among the sub-clusters using maximum \textit{a posteriori} probability. However, the BIC value for classification using three known groups was higher than the BIC value using two known groups, which again suggests the presence of sub-clusters.

Table 4.5: Classification of \texttt{f.voles} data assuming 50\% of the observations have known group membership.

<table>
<thead>
<tr>
<th>TRUE</th>
<th>Est. 1</th>
<th>Est. 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ochrogaster</td>
<td>45</td>
<td>-</td>
</tr>
<tr>
<td>californicus</td>
<td>-</td>
<td>41</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TRUE</th>
<th>Est. 1</th>
<th>Est. 2</th>
<th>Est. 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ochrogaster</td>
<td>28</td>
<td>17</td>
<td>-</td>
</tr>
<tr>
<td>californicus</td>
<td>-</td>
<td>-</td>
<td>41</td>
</tr>
</tbody>
</table>

4.3 Discussion

In this paper, we introduced a new family of parsimonious CWFA utilizing the covariance structures of the parsimonious Gaussian mixture models by McNicholas and Murphy (2008). Due to the introduction of a latent factor structure, the parameters are linear in dimensionality as opposed to the completely traditional CWM where the parameters grow quadratically; therefore, the CWFA approach is more suitable for modelling complex high-dimensional data than CWM. Using arti-
ficial and real data, we demonstrated that these models have very good clustering performance and the algorithm was able to recover the parameters very well.

Traditionally, the use of the EM approach for parameter estimation makes the algorithm very sensitive to the starting values due to presence of multiple local maxima in high dimensional space. We proposed a 5-step hierarchical initialization procedure that utilizes the nested structures within the family. Because these models have a hierarchical/ nested structure, it guarantees a natural ranking on the likelihoods of the models in our family. It also avoids a model A, which is nested in another model B, to have a greater likelihood than B.
Chapter 5

Clustering Gene Expression Time Course Data

Using Mixtures of Multivariate t-Distributions

5.1 Background

In gene expression time course studies, the expressions of a large number of genes are recorded over time. The role of clustering is paramount in the analysis of such data because the objective is to find groups of genes with similar expression patterns, or expression profiles, over the course of an experiment. Genes with similar expression profiles are said to co-express. In a typical time course experiment, the specific biological functions of certain genes will be known; if a gene with unknown function is found to co-express with genes with known functions, this provides a hint as to the function of the gene with unknown function.

Model-based clustering is a term used to describe the use of finite mixture models for clustering. A $p$-dimensional random vector $\mathbf{X}$ is said to arise from a parametric finite mixture distribution if, for all $\mathbf{x} \subset \mathbf{X}$, we can write $\xi(\mathbf{x} | \vartheta) = \sum_{g=1}^{G} \pi_g \psi_g(\mathbf{x} | \theta_g)$, where $\pi_g$, such that $\sum_{g=1}^{G} \pi_g = 1$ and $\pi_g > 0$, are the mixing proportions, $\vartheta = (\pi_1, \ldots, \pi_G, \theta_1, \ldots, \theta_G)$ is the vector of parameters, and $\psi_1(\mathbf{x} | \theta_1, \ldots, \theta_G)$
\( \theta_g), \ldots, \psi_G(x \mid \theta_g) \) are the component, or group, densities. Gaussian mixture models have received the bulk of the attention in the mixture modelling literature due to their mathematical tractability (recent examples include Bouveyron et al., 2007; McNicholas, 2010). Herein, we will work with the more robust mixtures of multivariate \( t \)-distributions, which were motivated by McLachlan and Peel (1998) as a heavier-tailed alternative to the Gaussian mixture model for model-based clustering. The model density for a mixture of multivariate \( t \)-distributions model with \( G \) components has the form

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

where \( f(x \mid \mu_g, \Sigma_g, \nu_g) \) is the density of a multivariate \( t \)-distribution with mean \( \mu_g \), covariance matrix \( \Sigma_g \), and \( \nu_g \) degrees of freedom.

Typically, a ‘family’ of mixture models is developed based on the imposition of constraints upon decomposed component covariance matrices. Then, the best member of the family of models is chosen based on some likelihood- and parsimony-based criterion. The most well-known family of mixture models for model-based clustering is the MCLUST family (cf. Celeux and Govaert, 1995; Fraley and Raftery, 2002), which utilizes an eigen-decomposed covariance structure. Another family of mixture models, known as parsimonious Gaussian mixture models, is based on constrained versions of the mixture of factor analyzers model (cf. Ghahramani and Hinton, 1987; Tipping and Bishop, 1999; McLachlan and Peel, 2000b; McNicholas and Murphy, 2008, 2010b). In general, the focus on decomposing the covariance structure is sensible because the vast majority of the parameters in a finite mixture model are in the component covariance matrices \( \Sigma_g \) (for all but very low dimensional cases). However, because longitudinal data are under consideration herein, the models developed in
Section 5.2 include a linear model for the mean of the component densities. This innovation is biologically relevant because the time course expression patterns and not the trend of the patterns per se, are of interest. Another consideration that arises due to the longitudinal nature of the data is the correlation structure; we discuss this further in Section 5.2.1.

5.2 Methodology

5.2.1 The multivariate $t$-distribution, the correlation structure, and the mean

McLachlan and Peel (1998) motivate the multivariate $t$-distribution as a heavy-tailed alternative to the multivariate Gaussian distribution, by first considering the Gaussian scale mixture model $(1 - \epsilon)\phi(\mathbf{x} \mid \mu, \Sigma) + \epsilon\phi(\mathbf{x} \mid \mu, c\Sigma)$, where $c$ is large, $\epsilon$ is small, and $\phi(\mathbf{x} \mid \mu, \Sigma)$ is the density of a multivariate Gaussian distribution with mean $\mu$ and covariance matrix $\Sigma$. They then note that this mixture model can be written

$$\int \phi(\mathbf{x} \mid \mu, \Sigma/u)dH(u),$$

for an appropriately defined probability distribution $H$. The multivariate $t$-distribution is then obtained upon replacement of $H$ by the probability density of a random variable $U \sim \text{gamma}(\nu/2, \nu/2)$, where $\nu$ is the degrees of freedom. Extensive details on multivariate $t$-distributions and applications thereof are given by Kotz and Nadarajah (2004).
To capture the longitudinal nature of the time course data under consideration herein, the covariance matrix is parameterized to account for the correlation structure of the data. The covariance matrix $\Sigma$ of a random variable can be decomposed using the relation $T \Sigma T' = D$, where $T$ is a unique lower unitriangular matrix and $D$ is a unique diagonal matrix with strictly positive diagonal entries (Pourahmadi, 1999). This relation is known as the modified Cholesky decomposition and has previously been used for discriminant analysis (Krzanowski et al., 1995) and clustering (McNicholas and Murphy, 2010a). The modified Cholesky decomposition may also be expressed in the form $\Sigma^{-1} = T'D^{-1}T$, and values of $T$ and $D$ have interpretations as generalized autoregressive parameters and innovation variances, respectively (cf. Pourahmadi, 1999). In model-based clustering applications, the mean is not routinely modelled. However, to effectively capture the trend of the time course data under consideration herein, we use a linear model for the mean. The efficacy of this approach is apparent in the results of the analyses in Section 5.3.

5.2.2 Model-based clustering

Consider the mixture of multivariate $t$-distributions model, where the component precision matrix is modified Cholesky-decomposed so that $\Sigma_g^{-1} = T_g'D_g^{-1}T_g$, with $T_g$ and $D_g$ as defined in Section 5.2.1, and the mean is modelled using a linear model. Now, suppose we have data $\mathbf{x} = (x_1, \ldots, x_n)'$, where each $x_i$ is measured at $p$ time points $t_1, \ldots, t_p$. To facilitate modelling of the mean of component $g$ via a line
with intercept $a_g$ and slope $b_g$, define

$$Q = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ t_1 & t_2 & \cdots & t_p \end{pmatrix}, \quad \beta_g = \begin{pmatrix} a_g \\ b_g \end{pmatrix}. $$

The likelihood for our mixture model can then be written

$$L(\theta | x) = \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g f(x_i | Q_\beta_g, (T'_g D_g^{-1} T_g)^{-1}, \nu_g), \quad (5.1)$$

where $f(x_i | Q_\beta_g, (T'_g D_g^{-1} T_g)^{-1}, \nu_g)$ is the density of a $p$-dimensional $t$-distributed random vector with mean $Q_\beta_g$, precision matrix $T'_g D_g^{-1} T_g$, and $\nu_g$ degrees of freedom.

Table 5.1: The nomenclature, covariance structure, and number of covariance parameters for the models used by McNicholas and Murphy (2010a).

<table>
<thead>
<tr>
<th>Model</th>
<th>$T_g$</th>
<th>$D_g$</th>
<th>$D_g$</th>
<th>Number of Covariance Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>EEA</td>
<td>Equal</td>
<td>Equal</td>
<td>Anisotropic</td>
<td>$p(p-1)/2 + p$</td>
</tr>
<tr>
<td>VVA</td>
<td>Variable</td>
<td>Variable</td>
<td>Anisotropic</td>
<td>$G[p(p-1)/2] + Gp$</td>
</tr>
<tr>
<td>VEA</td>
<td>Variable</td>
<td>Equal</td>
<td>Anisotropic</td>
<td>$G[p(p-1)/2] + p$</td>
</tr>
<tr>
<td>EVA</td>
<td>Equal</td>
<td>Variable</td>
<td>Anisotropic</td>
<td>$p(p-1)/2 + Gp$</td>
</tr>
<tr>
<td>VVI</td>
<td>Variable</td>
<td>Variable</td>
<td>Isotropic</td>
<td>$G[p(p-1)/2] + G$</td>
</tr>
<tr>
<td>VEI</td>
<td>Variable</td>
<td>Equal</td>
<td>Isotropic</td>
<td>$G[p(p-1)/2] + 1$</td>
</tr>
<tr>
<td>EVI</td>
<td>Equal</td>
<td>Variable</td>
<td>Isotropic</td>
<td>$p(p-1)/2 + G$</td>
</tr>
<tr>
<td>EEI</td>
<td>Equal</td>
<td>Equal</td>
<td>Isotropic</td>
<td>$p(p-1)/2 + 1$</td>
</tr>
</tbody>
</table>

Following McNicholas and Murphy (2010a), constraints can be imposed upon the component covariance matrices to give a family of eight mixture models with between $p(p-1)/2 + 1$ and $G[p(p-1)/2] + Gp$ covariance parameters (Table 5.1). Specifically, there is the option to constrain both the $T_g$ and the $D_g$ to be equal across components, as well as to impose the isotropic constraint $D_g = \delta_g I_p$. Practically, the isotropic constraint suggests that the variance at each time point is
the same, constraining $D_g = D$ for each $g$ indicates that the innovation variances are equal across components, and setting $T_g = T$ implies that the autoregressive structure is the same within each component. Parameter estimation for each model can be carried out using an expectation-maximization algorithm (cf. Section 5.2.3) and model selection can be carried out using a number of different criteria, two of which are outlined in Section 5.2.4.

5.2.3 Parameter estimation

The expectation-maximization (EM) algorithm (Dempster et al., 1977) is an iterative technique for finding maximum likelihood estimates when there are, or are assumed to be, missing data. In the E-step, the expected value of the complete-data log-likelihood, $C$ say, is computed. Then, in the M-step, $C$ is maximized with respect to the model parameters. To facilitate a clustering application, let $z_{ig}$ denote the component membership of observation $i$, where $z_{ig} = 1$ if observation $i$ belongs to component $g$ and $z_{ig} = 0$ otherwise. Continuing as outlined in Section 5.2.1, we introduce a latent variable $U_{ig}$ so that

$$X_i \mid u_{ig}, z_{ig} = 1 \sim \mathcal{N}(Q_g\beta_g, (T_g'D_g^{-1}T_g)^{-1}/u_{ig}),$$

independently, for $i = 1, \ldots, n$, and $U_{ig} \mid z_{ig} = 1$ follows a gamma distribution with parameters $(\nu_g/2, \nu_g/2)$, independently.

The observed plus missing data, known as the complete-data, then consist of the observed $x_i$, the unobserved $z_{ig}$, and the latent $u_{ig}$. The complete-data log-
likelihood can then be written in the form \( l_c(\vartheta) = l_{1c}(\pi) + l_{2c}(\nu) + l_{3c}(\varsigma) \), where \( \pi = (\pi_1, \ldots, \pi_G) \), \( \nu = (\nu_1, \ldots, \nu_G) \), \( \varsigma = (\beta_1, \ldots, \beta_G; T_1, \ldots, T_G, D_1, \ldots, D_G) \), and

\[
l_{1c}(\pi) = \sum_{g=1}^{G} n_g \log \pi_g,
\]

\[
l_{2c}(\nu) = \sum_{g=1}^{G} \sum_{i=1}^{n} z_{ig} \left[ -\log \Gamma \left( \frac{\nu_g}{2} \right) + \frac{\nu_g}{2} \log \left( \frac{\nu_g}{2} \right) + \frac{\nu_g}{2} (\log u_{ig} - u_{ig}) - \log u_{ig} \right],
\]

\[
l_{3c}(\varsigma) = \frac{np}{2} \log(2\pi) - \sum_{g=1}^{G} \frac{n_g}{2} \log |D_g| - \sum_{g=1}^{G} \frac{n_g}{2} \text{tr} \left\{ T_g S_g T_g' D_g^{-1} \right\},
\]

where \( n_g = \sum_{i=1}^{n} z_{ig} \) and \( S_g = (1/n_g) \sum_{i=1}^{n} z_{ig} u_{ig} (x_i - Q \beta_g)(x_i - Q \beta_g)' \), for \( g = 1, \ldots, G \).

At each E-step, the missing data, \( z_{ig} \) and \( u_{ig} \), are replaced by their conditional expected values

\[
\mathbb{E}[Z_{ig} \mid x_i] = \frac{\hat{\pi}_g f(x_i \mid Q \hat{\beta}_g; (\hat{T}_g' \hat{D}_g^{-1} \hat{T}_g)^{-1}, \hat{\nu}_g)}{\sum_{h=1}^{G} \hat{\pi}_h f(x_i \mid Q \beta_h; (T_h' D_h^{-1} T_h)^{-1}, \nu_h)} =: \hat{z}_{ig}, \tag{5.2}
\]

and

\[
\mathbb{E}[U_{ig} \mid x_i, z_{ig} = 1] = \frac{\hat{\nu}_g + p}{\hat{\nu}_g + \delta(x_i, Q \hat{\beta}_g; (T_g' D_g^{-1} T_g)^{-1})} =: \hat{u}_{ig}, \tag{5.3}
\]

respectively. In the M-step, and from \( l_{1c}(\pi) \), the mixing proportions are updated by \( \hat{\pi}_g = n_g/n \) and, from \( l_{2c}(\nu) \), the estimates for the \( \nu_g \) can be found by solving the
\[ 1 - \varphi\left(\frac{\hat{\nu}_g^{\text{new}}}{2}\right) + \frac{1}{n_g} \sum_{i=1}^{n} \hat{z}_{ig} (\log \hat{u}_{ig} - \hat{u}_{ig}) + \log \left(\frac{\hat{\nu}_g^{\text{new}} + p}{2}\right) + \varphi\left(\frac{\hat{\nu}_g^{\text{old}} + p}{2}\right) - \log \left(\frac{\hat{\nu}_g^{\text{old}} + p}{2}\right) = 0, \] 

(5.4)

for \( \hat{\nu}_g^{\text{new}} \), where \( \hat{\nu}_g^{\text{old}} \) is the estimate of \( \nu_g \) from the previous iteration and \( \varphi(\cdot) \) is the digamma function. Now, from \( l_{3e}(\varsigma) \), the update for \( \beta_g \) is given by

\[ \hat{\beta}_g = (Q' \hat{T}_g \hat{D}_g^{-1} \hat{T}_g Q)^{-1} Q' \hat{T}_g \hat{D}_g^{-1} \hat{T}_g \frac{\sum_{i=1}^{n} \hat{z}_{ig} \hat{u}_{ig} x_i}{\sum_{j=1}^{n} \hat{z}_{jg} \hat{u}_{jg}}. \]

Note that if we do not impose the linear constraint on the mean, then we would have

\[ \hat{\mu}_g = \frac{\sum_{i=1}^{n} \hat{z}_{ig} \hat{u}_{ig} x_i}{\sum_{j=1}^{n} \hat{z}_{jg} \hat{u}_{jg}}. \]

The estimates for the covariance structure depend upon the constraints that are imposed. In any case, as only the lower triangular elements of \( T_g \) need to be estimated, we adopt the usual notation by letting \( \phi_{ij}^{(g)} \) represent those elements of \( T_g \) that are to be estimated, so that

\[
T_g = \begin{pmatrix}
1 & 0 & 0 & 0 & \cdots & 0 \\
\phi_{21}^{(g)} & 1 & 0 & 0 & \cdots & 0 \\
\phi_{31}^{(g)} & \phi_{32}^{(g)} & 1 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
\phi_{p-1,1}^{(g)} & \phi_{p-1,2}^{(g)} & \cdots & \phi_{p-1,p-2}^{(g)} & 1 & 0 \\
\phi_{p1}^{(g)} & \phi_{p2}^{(g)} & \cdots & \phi_{p,p-2}^{(g)} & \phi_{p,p-1}^{(g)} & 1
\end{pmatrix}.
\]

Using this notation and considering the VVI case (cf. Table 5.1), where only the
isotropic constraint is imposed upon the covariance structure, differentiating $l_{3c}(\varsigma)$
with respect to $T_g$ and $\delta_g^{-1}$, respectively, and setting the resulting equations equal to
zero gives

$$
\begin{pmatrix}
\hat{\varphi}_{r1}^{(g)} \\
\hat{\varphi}_{r2}^{(g)} \\
\vdots \\
\hat{\varphi}_{r,r-1}^{(g)}
\end{pmatrix}
= -
\begin{pmatrix}
\hat{T}_g \\
\hat{T}_g' \\
\vdots \\
\hat{T}_g'
\end{pmatrix}
\begin{pmatrix}
S_{11}^{(g)} & S_{21}^{(g)} & \cdots & S_{r-1,1}^{(g)} \\
S_{12}^{(g)} & S_{22}^{(g)} & \cdots & S_{r-1,2}^{(g)} \\
\vdots & \vdots & \ddots & \vdots \\
S_{1,r-1}^{(g)} & S_{2,r-2}^{(g)} & \cdots & S_{r-1,r-1}^{(g)}
\end{pmatrix}^{-1}
\begin{pmatrix}
S_{r1}^{(g)} \\
S_{r2}^{(g)} \\
\vdots \\
S_{r,r-1}^{(g)}
\end{pmatrix},
$$

for $r = 2, \ldots, p$, where $s_{ij}^{(g)}$ is the element in the $i$th row and $j$th column of $S_g$, and
$\hat{\delta}_g = (1/p)\text{tr}\{T_g S_g T_g'\}$. These estimates for $T_1, \ldots, T_G$ and $\delta_1, \ldots, \delta_G$, along with
the covariance parameter estimates in the other seven cases, are analogous to those
given by McNicholas and Murphy (2010a), with but $S_1, \ldots, S_G$ as defined herein.

5.2.4 Model selection

The Bayesian information criterion (BIC; Schwarz, 1978) is often used for model
selection in model-based clustering applications involving a family of mixture models.
For a model with parameters $\vartheta$, the BIC is given by

$$
\text{BIC} = 2l(x, \hat{\vartheta}) - k \log N,
$$

where $l(x, \hat{\vartheta})$ is the maximized log-likelihood, $\hat{\vartheta}$ is the maximum likelihood estimate
of $\vartheta$, $k$ is the number of free parameters in the model, and $N$ is the number of
observations. Leroux (1992) and Keribin (2000) present theoretical results that, under certain regularity conditions, support the use of the BIC for the estimation of the number of components in a mixture model. One problem with the BIC in mixture model selection is the tendency towards assigning multiple mixture components to what is really just one cluster (cf. Biernacki et al., 2000). Biernacki et al. (2000) proposed the integrated completed likelihood (ICL) as an alternative to the BIC. The ICL essentially penalizes the BIC for estimated mean entropy, thereby punishing mixture components that are more spread out. In practice, an approximate ICL is used and this is given by

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

where MAP\{\hat{z}_{ig}\} is the maximum a posteriori classification given \( \hat{z}_{ig} \), that is MAP\{\hat{z}_{ig}\} = 1 if \( \max_g \{z_{ig}\} \) occurs in component \( g \) and MAP\{\hat{z}_{ig}\} = 0 otherwise. In our analyses (Section 5.3), both the BIC and the ICL are used for model selection.

5.2.5 Model-based classification

As mentioned in Section 5.1, the functions of some genes are typically known, or thought to be known, in advance. One way to make use of this information is to treat these genes as having known component, or group, memberships. The resulting semi-supervised clustering framework has been referred to as model-based classification (e.g., McNicholas, 2010; Andrews et al., 2011) but is also known as ‘partial classification’ (McLachlan and Basford, 1988, Section 1.11). Drawing on
Hosmer (1973), Titterington et al. (1985, Section 4.3.3) pointed out that knowing just a small proportion of the $z_i$ \textit{a priori} can lead to improved clustering performance.

Suppose there are $n$ observations and $k$ of these are known to belong to one of $G_1$ groups. Let the total number of components be $G_2$, so that each of the $n$ observations belongs to one of $G_2$ components, with $G_2 \geq G_1$. In model-based classification, we use the $k$ observations with known group memberships to estimate the group memberships for the remaining $n - k$ cases, and we do this within a joint likelihood framework. Using the multivariate $t$-distribution with modified Cholesky-decomposed covariance structure and a linearly-modelled mean, and ordering the data so that the first $k$ observations are known, the likelihood for the mixture model can be written

$$
\mathcal{L}(\vartheta \mid x) = \prod_{i=1}^{k} \prod_{g=1}^{G_1} \left[ \pi_g f(x_i \mid Q\beta_g, (T_g' D_g^{-1} T_g)^{-1}, \nu_g) \right]^{z_{ig}} \\
\times \prod_{j=k+1}^{n} \sum_{h=1}^{G_2} \pi_h f(x_j \mid Q\beta_h, (T_h' D_h^{-1} T_h)^{-1}, \nu_h),
$$

(5.5)

where the $z_{ig}$ are known for $i = 1, \ldots, k$. Note that the estimates of all parameters will be analogous to the model-based clustering case described in detail in Section 5.2.3.
5.3 Data analyses

5.3.1 Simulation Studies

First Simulation Study

Two hundred and seventy-five values were simulated from a two-component four-dimensional mixture of $t$-distributions, with linear models for the component means and modified Cholesky-decomposed covariance structures (VVA, Table 5.1). The component means and covariance matrices were distinct and $(n_1, n_2) = (130, 145)$. We ran model-based clustering on these data, treating all of the group memberships as unknown; we ran all eight covariance structures (cf. Table 5.1), for $G = 1, 2, 3$. We used the BIC and the ICL for model selection. Both criteria selected the same model: a two-component mixture model with only the isotropic constraint on the covariance matrix (VVI). This model gave perfect classification results and recovered parameter estimates close to the true values. The true and estimated values of each parameter, except $\mathbf{T}_1$ and $\mathbf{T}_2$, are given in Table 5.2.

Table 5.2: True and estimated values for $\mathbf{\beta}_1$, $\mathbf{\beta}_2$, $\mathbf{D}_1$, $\mathbf{D}_2$, $\nu_1$, and $\nu_2$ from the first simulation study.

<table>
<thead>
<tr>
<th>$g$</th>
<th>$\beta_y$</th>
<th>$\hat{\beta}_y$</th>
<th>$\nu_y$</th>
<th>$\hat{\nu}_y$</th>
<th>$\text{diag}{\mathbf{D}_y}$</th>
<th>$\text{diag}{\hat{\mathbf{D}}_y}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(1, 2)'</td>
<td>(0.97, 2.02)'</td>
<td>5</td>
<td>6.97</td>
<td>(0.30, 0.35, 0.30, 0.50)</td>
<td>(0.45, 0.45, 0.45, 0.45)</td>
</tr>
<tr>
<td>2</td>
<td>(0, 0)'</td>
<td>(0.22, −0.06)'</td>
<td>5</td>
<td>5.51</td>
<td>(1.74, 1.70, 1.70, 1.75)</td>
<td>(1.80, 1.80, 1.80, 1.80)</td>
</tr>
</tbody>
</table>
The true and estimated values for $T_1$ and $T_2$ are given below.

$$T_1 = \begin{bmatrix} 1.00 \\ 0.75 & 1.00 \\ 0.25 & 0.75 & 1.00 \\ 0.05 & 0.25 & 0.75 & 1.00 \end{bmatrix}, \quad T_2 = \begin{bmatrix} 1.00 \\ 0.20 & 1.00 \\ 0.05 & 0.3 & 1.00 \\ 0.01 & 0.02 & 0.35 & 1.00 \end{bmatrix}$$

$$\hat{T}_1 = \begin{bmatrix} 1.00 \\ 0.70 & 1.00 \\ 0.06 & 0.42 & 1.00 \\ -0.06 & 0.22 & 0.62 & 1.00 \end{bmatrix}, \quad \hat{T}_2 = \begin{bmatrix} 1.00 \\ 0.19 & 1.00 \\ 0.06 & 0.42 & 1.00 \\ 0.06 & 0.22 & 0.62 & 1.00 \end{bmatrix}$$

**Second Simulation Study**

In theory, our approach should be able to recover underlying multivariate Gaussian distributions by estimating larger values of the component degrees of freedom parameters. A second simulation study was conducted to investigate this ability. We simulated values ($n_1 = 130, n_2 = 145$) from a two-component four-dimensional mixture model, where the first component was a multivariate $t$-distribution and the second was a multivariate Gaussian distribution. The same mean and covariance structures were used but with different parameter values. We applied our model-based clustering approach to these data, again treating all group memberships as unknown and running all eight covariance structures (cf. Table 5.1), for $G = 1, 2, 3$. Using the BIC and the ICL for model selection, both criteria again selected the VVI
model. This model gave perfect classification and recovered parameter estimates close to the true values (Table 5.3 and Appendix A.3.1). For completeness, the true and estimated values of $T_1$ and $T_2$ are given in Appendix A.3.1. As expected, the estimated degrees of freedom $\hat{\nu} = (3.43, 84.60)$ reflect the fact that the second component was Gaussian.

Table 5.3: True and estimated values for $\beta_1$, $\beta_2$, $D_1$, $D_2$, $\nu_1$, and $\nu_2$ from the second simulation study.

<table>
<thead>
<tr>
<th>$g$</th>
<th>$\beta_g$</th>
<th>$\hat{\beta}_g$</th>
<th>$\nu_g$</th>
<th>$\hat{\nu}_g$</th>
<th>$\text{diag}{D_g}$</th>
<th>$\text{diag}{\hat{D}_g}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(0, 0)'</td>
<td>(0.09, -0.01)'</td>
<td>3.43</td>
<td>5</td>
<td>(1.74, 1.70, 1.70, 1.75)'</td>
<td>(1.74, 1.74, 1.74, 1.74)'</td>
</tr>
<tr>
<td>2</td>
<td>(1, 1)'</td>
<td>(1.10, 0.98)'</td>
<td>84.60</td>
<td>(1.30, 0.85, 0.83, 1.50)'</td>
<td>(1.13, 1.13, 1.13, 1.13)'</td>
<td></td>
</tr>
</tbody>
</table>

Third Simulation Study

In the case of heavy tailed data, model-based clustering using Gaussian mixture distributions tends to overestimate the number of clusters by creating additional groups to accommodate observations in the tails. In such cases, model-based clustering using a mixture of $t$-distributions can outperform the mixture of Gaussian distributions; this phenomenon has been demonstrated for other model-based clustering techniques (e.g., Andrews and McNicholas, 2011a) and we perform a third simulation herein to illustrate it for our family of models. We simulated 1,000 values from a two-component four-dimensional mixture model and treated the group membership labels as entirely unknown. These data where then analyzed using our mixtures of $t$-distributions model-based clustering paradigm, with all eight covariance structures (cf. Table 5.1), and its Gaussian analogue. This is the first time that Gaussian mixture models have been used for clustering with modified Cholesky-decomposed covariance
structure and a liner model for the mean.

Table 5.4: True and estimated values for $\beta_1$, $\beta_2$, $D_1$, $D_2$, $\nu_1$, and $\nu_2$ from the third simulation study using mixtures of $t$-distributions.

<table>
<thead>
<tr>
<th>$g$</th>
<th>$\beta_g$</th>
<th>$\beta_{\hat{g}}$</th>
<th>$\nu_g$</th>
<th>$\nu_{\hat{g}}$</th>
<th>$\text{diag}{D_g}$</th>
<th>$\text{diag}{\hat{D}_g}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(2, -1)$^T$</td>
<td>(2.0, -1.01)$^T$</td>
<td>5</td>
<td>5.01</td>
<td>(0.3, 0.35, 0.3, 0.5)$^T$</td>
<td>(0.32, 0.32, 0.29, 0.53)$^T$</td>
</tr>
<tr>
<td>2</td>
<td>(1, 1)$^T$</td>
<td>(1.06, 0.99)$^T$</td>
<td>3</td>
<td>2.93</td>
<td>(1.15, 1.12, 1.2, 1.2)$^T$</td>
<td>(1.33, 1.15, 1.04, 1.29)$^T$</td>
</tr>
</tbody>
</table>

Table 5.5: True classifications (1,2) tabulated against predicted classifications (A,B,C) resulting from the Gaussian mixture model analysis of the data from the third simulation study.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>534</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>263</td>
<td>202</td>
</tr>
</tbody>
</table>

Perfect classifications were observed using the mixtures of multivariate $t$-distributions approach and the estimated parameters were very close to their true values (Table 5.4 and Appendix A.3.1). In this study, the same covariance structure as in simulation study 1 was used. The Gaussian approach has split the second group into two Gaussian components (B and C; cf., Table 5.5) and when we look at the parameter estimates (Table 5.6), especially the innovation variances, it becomes clear that component C contains observations from the heavy tail of the true group 2 (cf., $\text{diag}\{\hat{D}_3\}$, Table 5.6). The role of component C is further reinforced by inspection of the time courses (Figure 5.1), which clearly illustrate the tendency of mixtures of multivariate Gaussian distributions to overestimate the number of temporal classes when the underlying distribution is heavy tailed.
Table 5.6: True and estimated values of $\beta_1$, $\beta_2$, $D_1$, and $D_2$ from the Gaussian mixture models applied to data from the third simulation study.

<table>
<thead>
<tr>
<th>$g$</th>
<th>$\beta_g$</th>
<th>$\hat{\beta}_g$</th>
<th>$\text{diag}{D_g}$</th>
<th>$\text{diag}{\hat{D}_g}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$(2, -1)'$</td>
<td>$(2.01, -1.01)'$</td>
<td>$(0.3, 0.35, 0.3, 0.5)'$</td>
<td>$(0.50, 0.55, 0.43, 0.81)'$</td>
</tr>
<tr>
<td>2</td>
<td>$(1, 1)'$</td>
<td>$(1, 0.99)'$</td>
<td>$(1.15, 1.12, 1.2, 1.2)'$</td>
<td>$(0.75, 0.55, 0.47, 0.69)'$</td>
</tr>
<tr>
<td>3</td>
<td>$(1, 0.98)'$</td>
<td></td>
<td></td>
<td>$(5.04, 5.03, 4.24, 5.41)'$</td>
</tr>
</tbody>
</table>

Figure 5.1: The 1,000 expression profiles from the third simulation; the green lines represent the members of component C under the chosen Gaussian model.

5.3.2 Yeast sporulation time course data

The data

The term 'sporulation' describes the process whereby a certain type of diploid cell gives rise to haploid cells through meiosis. Chu et al. (1998) used 6,118 yeast genes to measure changes in gene expression during sporulation. Each expression was measured over seven time points, namely $(0, 0.5, 2.0, 5.0, 7.0, 9.0, 11.5)$. The role of clustering in time course analyses such as this one is important because the objective is to identify groups of genes that have similar expression patterns over
the course of the experiment. Recall that genes with similar expression patterns are said to co-express. Chu et al. (1998) eliminated about 80% of the genes in a pre-processing stage; they retained genes with the root mean square of \( \log_2 R \) greater than 1.13, where \( R \) is the ratio of each gene’s expression level at a given time point to its expression level at its vegetative state \( (t = 0) \). They pointed out that this is “essentially equivalent” to retaining genes that show an average 2.2-fold change across the entire time course or a 3-fold change at a particular time point. Chu et al. (1998) then chose forty genes as being representative of seven temporal patterns; the remaining genes were clustered into seven groups based on their correlation with these model profiles.

Wakefield et al. (2003) used a four-stage Bayesian hierarchical model to analyze these data. Before applying their hierarchical model, they used Bayes factors to reduce the number of genes from 6,118 to 1,104 to make their approach computationally feasible. They found that the number of temporal classes was probably between 11 and 14, with \( G = 12 \) being the most probable. McNicholas and Murphy (2010a) used a family of Gaussian mixture models with a modified Cholesky-decomposed covariance structure to model these data; they estimated the number of temporal patterns at \( G = 13 \). Although their approach allowed for flexible modelling of the covariance structure, they imposed a Gaussian component density that is \textit{ipso facto} a restriction. Furthermore, they did not engage in any modelling of the mean. Our modelling framework therefore represents a more flexible version of the approach of McNicholas and Murphy (2010a), utilizing a more robust component density and a linear model for the mean.
The issue of data treatment deserves further discussion. Wakefield et al. (2003) used six time points by dividing the values at each subsequent time point by the value at \( t = 0 \), while McNicholas and Murphy (2010a) used all seven time points. Figures 5.2 and 5.3 show the model expression profiles of Chu et al. (1998) for the seven and six-dimensional versions of the data, respectively. By inspection, the six-dimensional version will be more suitable when we model the mean using a linear model; this is most apparent for the metabolic and early classes, where the initial steep increase in the profiles for seven dimensions is clear. We followed the data reduction approach of Chu et al. (1998), which left us with 1,162 genes and 38 of their 40 model expression profiles (cf. Figures 5.2 and 5.3), and we used the six-dimensional data.

**Model-based clustering approach**

A family of eight mixture models was applied to these data, that is, mixtures of multivariate \( t \)-distributions with eight covariance structures (Table 5.1) and with linear models for the component means. Mixtures of multivariate \( t \)-distributions are difficult to fit (cf. Andrews and McNicholas, 2011a) and so we used 10 different random starting values for the \( \hat{z}_{ig} \), for each model for \( G = 1, \ldots, 14 \). Of the models that could be fit, the best model according to both the BIC and the ICL was a VVA model with \( G = 5 \). The estimated degrees of freedom for this model \( \hat{\nu} = (6.43, 9.68, 76.12, 12.80, 10.96) \) reveal that while a Gaussian distribution might be suitable for cluster 3, the other four clusters are well captured by a \( t \)-distribution with less than thirteen degrees of freedom. The fact that the BIC and ICL selected
Figure 5.2: The forty model expression profiles of Chu et al. (1998), viewed over all seven time points: 38 genes that met the selection criterion in grey and the other two genes in black.

the same model suggests a good deal of certainty in the assignment of genes into groups, i.e., that the posterior probabilities \( \hat{z}_{ig} \) are generally close to zero or one.

Inspection of the model profiles (Figure 5.3) shows clear differences in autoregressive structure amongst the genes. Interestingly, the chosen model does not impose a constraint on the autoregressive structure, which seems more realistic than the model with \( T_g = T \) that was chosen by the Gaussian mixture modelling approach of McNicholas and Murphy (2010a). The temporal patterns of our five clusters are distinct (Figure 5.4) and support the notion that the autoregressive structure is different for each group. The results from our analysis can be directly compared (Table 5.7) with the predicted classifications of McNicholas and Murphy (2010a), who suggested
13 temporal patterns. Recall that they used all 6,118 genes and only the classifications for the 1,162 genes that we used are shown in Table 5.7.

Table 5.7: Maximum \textit{a posteriori} classifications for our best model (1–5) cross-tabulated against those from McNicholas and Murphy (2010a).

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>8</td>
<td>24</td>
<td>14</td>
<td>81</td>
<td>13</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>140</td>
<td>73</td>
<td>47</td>
<td>7</td>
<td>4</td>
<td>10</td>
<td>6</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>76</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>1</td>
<td>11</td>
<td>97</td>
<td>127</td>
<td>97</td>
<td>90</td>
<td>20</td>
<td>31</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>15</td>
<td>66</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>84</td>
</tr>
</tbody>
</table>

While one could search for correspondence between the classifications — our group 2 and their groups A–D; our group 3 and their group E — they are in fact substantively different, as opposed to being similar classifications broken across
Figure 5.4: The temporal patterns for the five clusters selected by our best model on the yeast data.

different numbers of components. The adjusted Rand index (Rand, 1971; Hubert and Arabie, 1985) can be used to compare two classifications; a value of 1 indicates perfect agreement and the index has an expected value of 0 under random classification. The adjusted Rand index associated with Table 5.7 is 0.27, indicating little more than random correspondence between the two classifications. The reason for this difference is likely a combination of the following factors.

1. McNicholas and Murphy (2010a) used all 6,118 genes and so their results were influenced by noise.

2. Their best model had the constraint $T_g = T$, so that the same autoregressive
structure was imposed on each group.

3. We accounted for the existence of trends in many of the profiles by modelling the mean using a line; McNicholas and Murphy (2010a) did not.

4. We used multivariate \( t \) component densities whereas McNicholas and Murphy (2010a) used Gaussian component densities; four of our five component densities are clearly not Gaussian, i.e., we have \( \nu_g < 13 \) for \( g \in \{1, 2, 4, 5\} \).

This is not to suggest that our approach is necessarily better, but it certainly is different as are the resulting classifications. While McNicholas and Murphy (2010a) support the ascertainment of Wakefield et al. (2003) regarding the presence of 11–14 temporal classes in these data, our results suggest far fewer clusters.

The biological significance of our clusters was evaluated using information from The Gene Ontology Consortium (1999). Within our selected model, genes with similar or related functions were clustered together. Group 1 consists mainly of the genes responsible for metabolic and biosynthetic processes, whereas group 2 consists mostly of genes responsible for organization and localization of organelles and macromolecules. Group 3 generally comprises genes related to RNA processing, and groups 4 and 5 comprise genes related to cell fusion and cytoskeletal or organelle parts, respectively.

We compared our predicted classifications with the classifications from Chu et al. (1998) (Table 5.8). All the genes either fell into cluster 4 or cluster 5. Cluster 4 contains most of the metabolic and early 2 genes whereas cluster 5 contains most of the early 1 and middle genes, with the genes from the other classes being split across
Table 5.8: Maximum *a posteriori* classifications for our best model (4–5) cross-tabulated against the model expression profiles from Chu *et al.* (1998).

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Metabolic</th>
<th>Early 1</th>
<th>Early 2</th>
<th>Early-mid</th>
<th>Middle</th>
<th>Midlate</th>
<th>Late</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>5</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>6</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

clusters 4 and 5.

Model-based classification approach

In an attempt to increase the range of models that could be fitted, the analysis from Section 5.3.2 was repeated within the model-based classification framework (cf. Section 5.2.5), taking the 38 expression profiles of Chu *et al.* (1998) as having known classifications. Again, 10 random starting values were used for the $\hat{z}_{ig}$. Within this semi-supervised framework, four of the eight covariance structures could be fitted for all $G$; no one covariance structure could be fitted for all $G$ in the analysis in Section 5.3.2. However, despite the improvement in the amount of models that could be fitted, the best model in this semi-supervised framework had lower BIC and ICL values than the best model from our model-based clustering analysis. Of course, $G = 5$ was not an option in this model-based classification analysis because the 38 model profiles covered $G = 7$ components.

5.3.3 Bronchial HAHRma data

The data

The HAHRma data consist of time course responses of human bronchial cell line A549 to Interleukin 13 (IL13), a protein coded by the IL13 gene. IL13 is known
to up-regulate CD23 and MHC class II expression, promote switching of the IgE
isotype in a special kind of white blood cells known as B cells, and down-regulate
the production of pro-inflammatory cytokines and chemokines that aid in the defense
mechanism for white blood cells (Weizmann Institute of Science, 1996). The human
bronchial cells were exposed to IL13 and measurements on the expression levels of the
22,283 genes were taken at 0, 4, 12, and 24 hours after exposure by hybridization with
Affymetrix U133a chips. The data were pre-processed and only genes that showed
a 2-fold change in expression level were included in the analysis; this left 237 genes.
The expression levels at each time point for a particular gene were divided by the
level at $t = 0$ and only those three time points were used in the analysis.

**Model-based clustering**

Our family of eight mixture models was applied to these data; as with the
yeast data, we used 10 different random starting values for the $\hat{z}_{ig}$. Both the BIC
and the ICL selected a VEA model with $G = 3$ components; this model imposes
no constraints on the autoregressive structure but does suggest that the innovation
variances are similar across components. The estimated degrees of freedom $\nu =
(2.79, 3.20, 70.40)$ suggest that two of the three component densities are clearly not
Gaussian. Figure 5.5 suggests that the assumption of equal innovation variances
across clusters is reasonable for these data, but that the autoregressive structure is
different.

Looking at the three clusters (Figure 5.5) shows genes in cluster 1 are gener-
al up-regulated throughout the time course, genes in cluster 2 are down-regulated,
and genes in cluster 3 are almost all up-regulated after a short period of down-regulation.

5.4 Discussion

The yeast data that we used have been analyzed many times previously. We selected a model with $G = 5$ components and an unconstrained autoregressive structure; the latter seems reasonable by inspection of the data (Figure 5.4). The number of components selected is much closer to the findings of Mitchell (1994) and Chu et al. (1998), who suggest around 4–7 seven temporal patterns present in these data; this is contrary to Wakefield et al. (2003) and McNicholas and Murphy (2010a).
who found around 11–14 components. The five clusters that we discovered had very clear biological interpretations, as discussed in Section 5.3.2. For the HAHrma data, we found three clusters that had clear biological interpretations (Section 5.3.3). We also performed model-based classification on the yeast data. This semi-supervised approach has potential for further use in gene expression time course analysis; genes known to have the same function can be taken as belonging to the same component or, if appropriate, as belonging to one of a few components. Model-based discriminant analysis could also be used for this purpose (cf. Andrews and McNicholas, 2011b, Section 2.2).

An effective approach to the important biological problem of clustering gene expression time course data has been introduced. Specifically, a model-based approach was followed using mixtures of multivariate $t$-distributions where each component density had a modified Cholesky-decomposed covariance structure and a linear model for the mean. Imposing constraints on the covariance structure led to a family of eight mixtures models, including parsimonious models. While this covariance structure had previously been used for clustering time course data (McNicholas and Murphy, 2010a), neither the multivariate $t$-distribution nor a linear model for the mean had previously been used. Furthermore, our data analyses confirm that these two innovations had an impact; across both real data analyses, only two of a total of eight components had degrees of freedom above 13 and the trend in all clusters was clear (Figures 5.4 and 5.5). Simulation studies were used to demonstrate that our parameter estimation approach, via the EM algorithm, and model-selection using the BIC or ICL could recover group memberships and give parameter estimates close to
the true values. Notably, the BIC and the ICL chose the same model in all real and simulated data analyses; this suggests that the posterior probabilities $\hat{z}_{ig}$ were usually close to zero or to one.

Although this work represents a step forward in the model-based clustering of gene expression time course data, a number of questions emerge. With the modelling approach used herein, as with that of McNicholas and Murphy (2010a), the resulting temporal patterns can exhibit a symmetry about the mean in each component (cf. Figure 5.5, clusters 1 and 2)—that is, a gene with an early increase and a late decrease may well end up in the same group as a gene with an early decrease and a late increase. This is not a problem \textit{per se} because genes that are expressing (positively or negatively) at the same time might well belong in the same group. Furthermore, if one wanted to split a group to separate these types of patterns, it could be done very easily \textit{a posteriori}. This feature is a specific example of the general problem that a mixture component may not necessarily correspond to what one might practically consider to be a ‘cluster’.

In some recent model-based clustering and classification work using the multivariate $t$-distribution (Andrews and McNicholas, 2011a,b; Andrews \textit{et al}., 2011), models with constrained degrees of freedom ($\nu_g = \nu$) have been considered. We did not consider such models here because, from experience, we learned that such models might be selected even when they are obviously sub-optimal: we give an illustrative example of this point in Appendix A.3.2. This is a feature of the model-selection approach that we used herein; it is well known that the BIC or ICL can over-penalize the log-likelihood term for lack of parsimony. Considering how obviously different
the yeast gene expression profiles are, the fact that McNicholas and Murphy (2010a) chose a model with a constrained autoregressive parameter \( (T_g = T) \) could only be an artifact of the BIC. While the BIC is the most popular model selection criterion within the literature, it is not tailored for clustering or classification applications. This, together with the aforementioned tendency to over-penalize for lack of parsimony, makes the ongoing search for viable alternatives to the BIC an important area of endeavour.
Chapter 6

Conclusion

This thesis was an exposition of three recent pieces of work on mixture model-based clustering and classification. GPCM provides insight on geometrical representation of the data, whereas CWFA provides flexibility in modelling high dimensional data. Moreover, the mixture of $t$-distributions with Choleksy-decomposed covariance structure takes into account the correlations among the measurements making it ideal for longitudinal studies. Thus, the primary focus of the thesis was implementation of several decompositions of covariance matrix, each having its own merits and applicability in different scenarios. Different approaches for modelling and parameter estimation were utilized.

6.1 Variational Bayes approach for Gaussian parsimonious clustering models (GPCM)

6.1.1 Summary

Mixture model-based clustering is now commonly utilized within the family setting where a family arises as a result of decomposition followed by constraints on the decomposition of the covariance matrices. Gaussian parsimonious clustering
models, with eigen-decomposed covariance structures, are the most famous family of Gaussian mixture models within the literature, partly owing its popularity to the introduction of the R package mclust. The constraints of these covariance matrices provide a geometrical interpretation about the shape of the data. Deviating from the most popular EM-BIC rubric, we used variational Bayes approximations for parameter estimation and the deviance information criterion (DIC) for model selection. The variational Bayes approach alleviated some of the computational complexities associated with the EM algorithm by constructing a tight lower bound on the complex marginal likelihood and maximizing this lower bound by minimizing the associated Kullback-Leibler divergence. We also incorporated model-based classification, which is a semi-supervised clustering approach, in variational Bayes framework. Real and simulated data were used to compare our approach to the EM-BIC rubric.

6.1.2 Future work

Variational Bayes algorithm allows for simultaneous estimation of the parameters and the number of components. Despite the advantage of of the variational Bayes approach for simultaneously obtaining the number of components and the parameter estimation, a model selection criterion needs to be utilized while selecting the covariance structure. We used the DIC for the selection of the covariance structure and, as can be seen from the simulation studies, the correct structure can be selected using the DIC in conjunction with the variational Bayes approach. That said, it may well be the case that another criterion is more suitable for selecting the member of a family of models (i.e., the covariance structure).
This approach could easily be extended for other decomposition of the covariance matrices as well e.g. modified Cholesky-decomposed covariance matrix, mixtures of factor analysers. The simultaneous selection of number of components reduces the computational demand of the mixture of factor analysers specially in the case of high dimensional data where each covariance structure needs to be evaluated at possible number of components and number of latent factor.

6.2 Cluster-Weighted Factor Analysers

6.2.1 Summary

Cluster-weighted modeling is a flexible statistical framework for modeling local relationships in heterogeneous populations on the basis of weighted combinations of local models. In particular, this technique models the joint density $p(x, y)$ of a random vector $X$ and a random variable $Y$, which can be considered the input and output variables, respectively. It can be shown that, under suitable assumptions, CWM includes finite mixtures of regressions as special cases. We extended linear Gaussian cluster-weighted models to include an underlying latent factor structure for the input variables. This resulted in a covariance matrix that followed the structure of the PGMMs for the input variable. Utilizing the constraints by McNicholas and Murphy (2008) on the covariance structure of the input variables along with an additional constraint on the variance of the output variable, we introduced a novel family of models known as parsimonious cluster-weighted factor analyzers. This allowed the model the flexibility of modeling high dimensional datasets. Numerical results based
on both simulated and real data were provided.

6.2.2 Future work

For $q > 1$, $\Lambda$ is unidentifiable as the model is still satisfied even when the latent factor $u_i$ is replaced by $Hu_i$ and $\Lambda$ by $\Lambda H'$, where $H$ is any orthogonal matrix of order $q$ (McLachlan and Peel, 2000a). This results in an infinite number of possibilities for the factor loading matrix $\Lambda$. Even though this does not affect the clustering algorithm, because $\Lambda\Lambda'$ is unique, interpretation of estimated $\Lambda$ is not informative. A future avenue of research is to explore further constraints on the factor loading matrix to ensure a uniquely defined factor loading matrix $\Lambda$.

Also, while the BIC was able to identify the correct model, the choice of a convenient model selection criterion for these models is still an open question. Some future work will be devoted to the search for good model selection criteria for these models such as the ICL.

Also, here we assume that number of factors are same across group which might be too restrictive. However, that also increases the number of models that needs to be fitted resulting in huge computational burden. That also brings the need to explore other approaches like variational Bayes which will significantly reduce the number of models that needs to be fitted as mentioned in Section 6.1.2.
6.3 Mixture of $t$-distributions with modified Cholesky-decomposed
covariance structure

6.3.1 Summary

Clustering gene expression time course data is an important problem in bioinformatics because understanding which genes behave similarly can lead to the discovery of important biological information. Statistically, the problem of clustering time course data is a special case of the more general problem of clustering longitudinal data. In this paper, a very general and flexible model-based technique was used to cluster longitudinal data. Mixtures of multivariate $t$-distributions were utilized, with a linear model for the mean and a modified Cholesky-decomposed covariance structure. Constraints were placed upon the covariance structure, leading to a novel family of mixture models, including parsimonious models. In addition to model-based clustering, these models were also used for model-based classification, i.e., semi-supervised clustering. Parameters, including the component degrees of freedom, were estimated using an expectation-maximization algorithm and two different approaches to model selection are considered. The models were applied to simulated data to illustrate their efficacy; this included a comparison with their Gaussian analogues — the use of these Gaussian analogues with a linear model for the mean was novel in itself. Our family of multivariate $t$ mixture models was then applied to two real gene expression time course data sets and the results were discussed.
6.3.2 Future work

The parameter estimation difficulties, reflected by our inability to fit all models in either data analysis over 10 random starts, are also the subject of ongoing work. One approach we have tried is to use discrete degrees of freedom, in a fashion similar to that suggested by Besag et al. (1995). We repeated the analysis of Section 5.3.2 but at each iteration selected the value of the $\hat{\nu}_g$ that gave the best solution to Equation 5.4 from amongst the values \{2, 3, 4, 5, 6, 7, 8, 9, 10, 14, 18, 32\}. Unfortunately, this led to no improvement in the number of models that could be fitted, or in the fit of the models, over the results presented in Section 5.3.2. An approach that is the subject of ongoing work by the authors and others is the use of variational Bayes approximations for parameter estimation and selection of the number of mixture components. Initial work on their use in mixture model parameter estimation (cf. Corduneanu and Bishop, 2001; Ueda and Ghahramani, 2002; McGrory and Titterington, 2007) indicates this may be a promising direction for future research.

While the focus of this paper has been the analysis of gene expression data, where time courses are typically short, the methodology introduced can be applied to any longitudinal data set. Moreover, longer time course data sets would benefit from this clustering approach because the constraints (Table 5.1) greatly reduce the number of parameters to be estimated. In other applications, one may wish to relax the constrained nature of the linear model for the mean. Herein, the mean of each component was modelled as a line. However, a more general linear model of the form $Q_g \beta_g$ could be used for each component, where $Q_g$ is a $p \times q$ design matrix and $\beta_g$
is a $q \times 1$ vector of parameters. In such a framework, one would have the option of fitting more or less parsimonious models by imposing, or not, the constraints $Q_g = Q$ and $\beta_g = \beta$.

Non-parametric approaches for time course gene expression data are also emerging in the literature (Luan and Li, 2003; Ma et al., 2006), where the mean gene expression profiles are modelled as linear combinations of spline bases. Another interesting paradigm is noise-robust soft clustering (Futschik and Carlisle, 2005) which as opposed to the hard partitioning of the data (one gene into only one cluster), generates the internal clusters for genes and an overall global cluster structure for relationships between clusters. Incorporating these novel ideas into model-based clustering using mixture models might open up a completely new avenue for further research.
Bibliography

Airoldi, J. and Hoffmann, R. (1984a). Age variation in voles (Microtus californicus, M. ochrogaster) and its significance for systematic studies. Occasional papers of the Museum of Natural History 111, University of Kansas, Lawrence, KS.


Appendix A

A.1 Parameter Estimation and Model Selection for a Family of GPCM using Variational Bayes Approximations

A.1.1 Mathematical Details

Model EEV

The mixing proportions were assigned a Dirichlet prior distribution, such that

\[ q(\pi) = \text{Dir}(\rho, \alpha^{(0)}_1, \ldots, \alpha^{(0)}_G). \]

For the mean, a Gaussian distribution conditional on the covariance matrix was used, such that

\[ q(\mu | \lambda, A, D_1, \ldots, D_G) = \prod_{g=1}^{G} \phi(\mu_g; m^{(0)}_g, (\beta^{(0)}_g)^{-1} \lambda D_g A D_g'). \]

For the parameters of the covariance matrix, the following priors were used: the kth diagonal elements of \((\lambda A)^{-1}\) were assigned a Gamma \((a_k^{(0)}, b_k^{(0)})\) distribution and \(D_g\) was assigned a matrix von Mises-Fisher \((C_g^{(0)})\) distribution. By setting \(\tau = (\lambda A)^{-1}\),
its prior can be written as

\[ p_\tau(\tau) \propto \prod_{k=1}^{K} \frac{a_k^{(0)}}{\tau_k^2} - 1 \exp \left\{ - \frac{b_k^{(0)}}{2} \tau_k \right\}. \]

The matrix \( D \) has a density as defined by Downs (1972):

\[ p_D(D) = \prod_{g=1}^{G} a(C_g^{(0)}) \exp \left( \text{tr}(C_g^{(0)}D_g') \right). \]

for \( D_g \in O(n, p) \), where \( O(n, p) \) is the Stiefel manifold of \( n \times p \) matrices.

The joint distribution of \( \mu, \lambda, A, \) and \( D \) becomes

\[
p(\mu, \tau, D) \propto \prod_{g=1}^{G} |\beta_g^{(0)}\tau|^{\frac{1}{2}} \exp \left\{ \frac{-(\mu_g - m_g^{(0)})\beta_g^{(0)}D_g'\tau D_g(\mu_g - m_g^{(0)})'}{2} \right\} \\
\times \exp \left\{ \text{tr}(C_g^{(0)}D_g') \right\} \prod_{k=1}^{K} \frac{a_k^{(0)}}{\tau_k^2 - 1} \exp \left\{ - \frac{b_k^{(0)}}{2} \tau_k \right\}.
\]

The likelihood of the data can be written as

\[
\mathcal{L}(\mu, \tau, D \mid y_1, \ldots, y_n) \propto |\tau|^{-\frac{n^2 + n}{2}} \exp \left\{ \frac{-\sum_{i=1}^{n} z_{ig}(y - \mu_g)D_g'\tau D_g(y - \mu_g)'}{2} \right\}.
\]

Therefore, the joint posterior distribution of \( \mu, \lambda, A, \) and \( D \) is

\[
p(\mu, \tau, D \mid y_1, \ldots, y_n) \propto p(\mu, \tau, D) \times \mathcal{L}(\mu, \tau, D \mid y_1, \ldots, y_n).
\]
Thus, the posterior distribution of mean becomes

\[ q_\mu(\mu \mid \tau, D_1, \ldots, D_G) = \prod_{g=1}^{G} N_p(\mu_g; \mathbf{m}_g, (\beta_g D'_g \tau D_g)^{-1}), \]

where \( \beta_g = \beta_g^{(0)} + \sum_{i=1}^{n} \hat{z}_{ig} \) and

\[ \mathbf{m}_g = \frac{\beta_g^{(0)} \mathbf{m}_g^{(0)}}{\beta_g} + \sum_{i=1}^{n} \frac{\hat{z}_{ig} \mathbf{y}_i}{\beta_g}. \]

The posterior distribution for the \( k \)th diagonal element of \( \tau = (\lambda A)^{-1} \) becomes

\[ q_{\tau}(\tau_k) = \text{Gamma}(a_k, b_k) \]

where

\[ a_k = a_k^{(0)} + \sum_{g=1}^{G} \sum_{i=1}^{n} \hat{z}_{ig} \times d \]

and

\[ b_k = b_k^{(0)} + \sum_{g=1}^{G} \left( \sum_{i=1}^{n} \hat{z}_{ig} y_{ik}^2 + \beta_g^{(0)} m_{gk}^2 - \beta_g m_{gk}^2 \right). \]

For the posterior distribution for the \( D_g \), we have

\[ q(D_g | y, \mu_g, \tau) \propto \exp \left\{ \text{tr} \left( \frac{-(\mu_g - \mathbf{m}_g^{(0)}) \beta_g^{(0)} D'_g \tau D_g (\mu_g - \mathbf{m}_g^{(0)}){'}}{2} \right) \right\} \]

\[ \exp \left\{ \text{tr} \left( \frac{-\sum_{i=1}^{n} \hat{z}_{ig} (y - \mu_g) D'_g \tau D_g (y - \mu_g)}{2} + C_g^{(0)} D_g' \right) \right\}, \]

which is the functional form of matrix Bingham-von Mises-Fisher distribution, such that

\[ \exp \left\{ \text{tr}(Q_g D_g P D'_g + R_g D_g') \right\}, \]
where

\[ Q_g = -\sum_{i=1}^{n} z_{ig}(y - \mu_g)(y - \mu_g) + (\mu_g - m_g(0))^\beta_g(\mu_g - m_g(0))', \]

\[ P = \tau_g, \text{ and } R_g = C_g^{(0)}. \]

**Posterior for \( D_g \) in the VEV Model**

Similarly, the posterior distribution of \( D_g \) for the VEV model has the form

\[
q(D_g | y, \mu_g, \tau_g) \propto \exp \left\{ \text{tr} \left( \frac{-(\mu_g - m_g(0))^\beta_g(0)D'_g\tau_gD_g(\mu_g - m_g(0))'}{2} \right) \right\} \\
\exp \left\{ \text{tr} \left( \frac{-\sum_{i=1}^{n} z_{ig}(y - \mu_g)D'_g\tau_gD_g(y - \mu_g)'}{2} + C_g^{(0)}D'_g \right) \right\},
\]

which is a matrix Bingham-von Mises-Fisher distribution, such that

\[
\exp \left\{ \text{tr}(Q_gD_gP_gD'_g + R_gD'_g) \right\}
\]

where

\[ Q_g = -(\sum_{i=1}^{n} z_{ig}(y - \mu_g)(y - \mu_g) + (\mu_g - m_g(0))^\beta_g(\mu_g - m_g(0))'), \]

\[ P_g = \tau_g, \text{ and } R_g = C_g^{(0)}. \]

**A.1.2 Tables**

**Posterior Distributions**
<table>
<thead>
<tr>
<th>Model</th>
<th>Posterior Distributions</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>EII</td>
<td>Gamma $(a, b)$</td>
<td>\begin{align*} a &amp;= a^{(0)} + \sum_{g=1}^{G} \sum_{i=1}^{n} \hat{z}<em>{ig} \times d \ b &amp;= b^{(0)} + \sum</em>{g=1}^{G} \sum_{i=1}^{n} \hat{z}_{ig} y_i' y_i + \beta_g^{(0)} m_g^{(0)T} m^{(0)} - \beta_g m_g' m \end{align*}</td>
</tr>
<tr>
<td>VII</td>
<td>Gamma $(a_g, b_g)$</td>
<td>\begin{align*} a_g &amp;= a_g^{(0)} + \sum_{i=1}^{n} \hat{z}<em>{ig} \times d \ b_g &amp;= b_g^{(0)} + \sum</em>{i=1}^{n} \hat{z}_{ig} y_i' y_i + \beta_g^{(0)} m_g^{(0)T} m^{(0)} - \beta_g m_g' m \end{align*}</td>
</tr>
<tr>
<td>EEI</td>
<td>Gamma $(a_k, b_k)$</td>
<td>\begin{align*} a_k &amp;= a_k^{(0)} + \sum_{g=1}^{G} \sum_{i=1}^{n} \hat{z}<em>{ig} \ b_k &amp;= b_k^{(0)} + \sum</em>{g=1}^{G} \sum_{i=1}^{n} \hat{z}_{ig} y_i^2 + \beta_g^{(0)} m_g^{(0)T} m^{(0)} - \beta_g m_g' m \end{align*}</td>
</tr>
<tr>
<td>VEI</td>
<td>Gamma $(a_g, b_g)$</td>
<td>\begin{align*} a_g &amp;= a_g^{(0)} + \sum_{i=1}^{n} \hat{z}<em>{ig} \times d \ b_g &amp;= b_g^{(0)} + \sum</em>{i=1}^{n} \hat{z}_{ig} y_i' y_i + \beta_g^{(0)} m_g^{(0)T} m^{(0)} - \beta_g m_g' m \end{align*}</td>
</tr>
<tr>
<td></td>
<td>Gamma $(a_l, b_e)$</td>
<td>\begin{align*} a_l &amp;= a_l^{(0)} + \sum_{g=1}^{G} \sum_{i=1}^{n} \hat{z}<em>{ig} \ b_e &amp;= b_e^{(0)} + \sum</em>{g=1}^{G} \sum_{i=1}^{n} \hat{z}_{ig} y_i^2 + \beta_g^{(0)} m_g^{(0)T} m^{(0)} - \beta_g m_g' m \end{align*}</td>
</tr>
<tr>
<td>EVI</td>
<td>Gamma $(a, b)$</td>
<td>\begin{align*} a &amp;= a^{(0)} + \sum_{g=1}^{G} \sum_{i=1}^{n} \hat{z}<em>{ig} \times d \ b &amp;= b^{(0)} + \sum</em>{g=1}^{G} \sum_{i=1}^{n} \hat{z}_{ig} y_i' y_i + \beta_g^{(0)} m_g^{(0)T} m^{(0)} - \beta_g m_g' m \end{align*}</td>
</tr>
<tr>
<td></td>
<td>Gamma $(a_{lgk}, b_{egk})$</td>
<td>\begin{align*} a_{lgk} &amp;= a_{lgk}^{(0)} + \sum_{i=1}^{n} \hat{z}<em>{ig} \ b</em>{egk} &amp;= b_{egk}^{(0)} + \sum_{i=1}^{n} \hat{z}_{ig} y_i^2 \end{align*}</td>
</tr>
</tbody>
</table>
\[
+\beta_g^{(0)} m_{g k}^{(0)^2} - \beta_g m_{g k}^2
\]

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Parameters</th>
<th>Formulas</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>VVI</strong> Gamma ((a_g, b_g))</td>
<td>(a_g = a_g^{(0)} + \sum_{i=1}^n \hat{z}_{ig} \times d)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(b_g = b_g^{(0)} + \sum_{i=1}^n \hat{z}_{ig}y_i' y_i)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+(\beta_g^{(0)} m_{g}^{(0)^T} m^{(0)} - \beta_g m_g' m)</td>
<td></td>
</tr>
</tbody>
</table>

**Gamma \((al_{gk}, be_{gk})\)**

\[
al_{gk} = al_{gk}^{(0)} + \sum_{i=1}^n \hat{z}_{ig}
\]

\[
be_{gk} = be_{gk}^{(0)} + \sum_{i=1}^n \hat{z}_{ig} y_{ik}^2
\]

+\(\beta_g^{(0)} m_{g k}^{(0)^2} - \beta_g m_{g k}^2\)

<table>
<thead>
<tr>
<th><strong>EEE</strong> Wishart ((v, \Sigma^{-1}))</th>
<th>(v = v^{(0)} + \sum_{g=1}^G \sum_{i=1}^n \hat{z}_{ig})</th>
<th>(\Sigma^{-1} = \Sigma^{(0)-1} + \sum_{g=1}^G (\sum_{i=1}^n \hat{z}_{ig} y_i'y_i))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+(\beta_g^{(0)} m_{g}^{(0)^T} m^{(0)} - \beta_g m_g' m)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>VEE</strong> Gamma ((a_g, b_g))</th>
<th>(a_g = a_g^{(0)} + \sum_{i=1}^n \hat{z}_{ig} \times d)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(b_g = b_g^{(0)} + \sum_{i=1}^n \hat{z}_{ig}y_i' y_i)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+(\beta_g^{(0)} m_{g}^{(0)^T} m^{(0)} - \beta_g m_g' m)</td>
<td></td>
</tr>
</tbody>
</table>

**Wishart \((v, \Sigma)\)**

\[
v = v^{(0)} + \sum_{g=1}^G \sum_{i=1}^n \hat{z}_{ig}\]

\[
\Sigma = \Sigma^{(0)} + \sum_{g=1}^G (\sum_{i=1}^n \hat{z}_{ig} y_i'y_i)
\]

+\(\beta_g^{(0)} m_{g}^{(0)^T} m^{(0)} - \beta_g m_g' m\)

<table>
<thead>
<tr>
<th><strong>EEV</strong> Gamma ((a_k, b_k))</th>
<th>(a_k = a_k^{(0)} + \sum_{g=1}^G \sum_{i=1}^n \hat{z}_{ig} \times d)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(b_k = b_k^{(0)} + \sum_{g=1}^G (\sum_{i=1}^n \hat{z}<em>{ig} y</em>{ik}^2)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+(\beta_g^{(0)} m_{g k}^2 - \beta_g m_{g k}^2)</td>
<td></td>
</tr>
</tbody>
</table>

**Matrix Bingham-von**

See Appendix A.1.1

| **Mises-Fisher \((P, Q_g, R)\)\) | \(\Sigma = \Sigma^{(0)} + \sum_{g=1}^G (\sum_{i=1}^n \hat{z}_{ig} y_i'y_i)
\]
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+(\beta_g^{(0)} m_{g}^{(0)^T} m^{(0)} - \beta_g m_g' m)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>VEV</strong> Gamma ((a_g, b_g))</th>
<th>(a_g = a_g^{(0)} + \sum_{i=1}^n \hat{z}_{ig} \times d)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(b_g = b_g^{(0)} + \sum_{i=1}^n \hat{z}_{ig}y_i' y_i)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+(\beta_g^{(0)} m_{g}^{(0)^T} m^{(0)} - \beta_g m_g' m)</td>
<td></td>
</tr>
</tbody>
</table>
Gamma \((a, b)\)

\[ a_l = a_l^{(0)} + \sum_{g=1}^{G} \sum_{i=1}^{n} \hat{z}_{ig} \]

\[ b_{ek} = b_{ek}^{(0)} + \sum_{g=1}^{G} (\sum_{i=1}^{n} \hat{z}_{ig} y_{ik}^2 + \beta_g^{(0)} m_{gk}^{(0)2} - \beta_g m_{gk}^2) \]

matrix Bingham-von Mises-Fisher \((P_g, Q_g, R)\) See Appendix A.1.1

Mises-Fisher \((P_g, Q_g, R)\)

<table>
<thead>
<tr>
<th></th>
<th>EVV</th>
<th>Wishart ((v_g, \Sigma^{-1}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>(a = a^{(0)} + \sum_{g=1}^{G} \sum_{i=1}^{n} \hat{z}_{ig} \times d)</td>
<td>(v_g = v_g^{(0)} + \sum_{i=1}^{n} \hat{z}_{ig})</td>
</tr>
<tr>
<td>(b)</td>
<td>(b = b^{(0)} + \sum_{g=1}^{G} (\sum_{i=1}^{n} \hat{z}_{ig} y_i' y_i + \beta_g^{(0)} m_g^{(0)T} m_g^{(0)} - \beta_g m_g m_g'))</td>
<td>(\Sigma_g^{-1} = \Sigma_g^{(0)-1} + \sum_{i=1}^{n} \hat{z}_{ig} y_i y_i' + \beta_g^{(0)} m_g^{(0)T} m_g^{(0)} - \beta_g m_g m_g')</td>
</tr>
</tbody>
</table>

Wishart \((v_g, \Sigma^{-1})\)

<table>
<thead>
<tr>
<th></th>
<th>Wishart ((v_g, \Sigma^{-1}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(v_g)</td>
<td>(v_g = v_g^{(0)} + \sum_{i=1}^{n} \hat{z}_{ig})</td>
</tr>
<tr>
<td>(\Sigma_g^{-1})</td>
<td>(\Sigma_g^{-1} = \Sigma_g^{(0)-1} + \sum_{i=1}^{n} \hat{z}_{ig} y_i y_i' + \beta_g^{(0)} m_g^{(0)T} m_g^{(0)} - \beta_g m_g m_g')</td>
</tr>
</tbody>
</table>

Posterior Expected Values

A.2 A family of Cluster-Weighted factor analysers

A.2.1 Mathematical Details

Details on Conditional Distribution of \(Y|X, u\)

To compute the distribution of \(Y|X, u\), we begin by recalling that if \(Z \sim N_q(m, \Gamma)\) is a random vector with values in \(\mathbb{R}^q\) and if \(Z\) is partitioned as \(Z = \)
\((Z_1', Z_2')', \) where \(Z_1\) takes values in \(\mathbb{R}^{q_1}\) and \(Z_2\) in \(\mathbb{R}^{q_2} = \mathbb{R}^{q-q_1}\), then we can write
\[
m = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} \quad \text{and} \quad \Gamma = \begin{pmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{pmatrix}.
\]

Now, because \(Z\) has a multivariate normal distribution, \(Z_1|Z_2 = z_2\) and \(Z_2\) are statistically independent with \(Z_1|Z_2 = z_2 \sim N_{q_1}(m_{1|2}, \Gamma_{1|2})\) and \(Z_2 \sim N_{q_2}(m_2, \Gamma_{22})\), where
\[
m_{1|2} = m_1 + \Gamma_{12} \Gamma_{22}^{-1}(z_2 - m_2) \quad \text{and} \quad \Gamma_{1|2} = \Gamma_{11} - \Gamma_{12} \Gamma_{22}^{-1} \Gamma_{21}.
\]

Therefore, setting \(Z = (Z_1', Z_2')', \) where \(Z_1 = Y\) and \(Z_2 = (X', U')', \) gives \(m_1 = \beta_0 + \beta_1 \mu\) and \(m_2 = (\mu', 0')'\), with the elements of matrix (4.6) given by
\[
\Gamma_{11} = \beta_1 \Sigma + \sigma_\varepsilon^2, \quad \Gamma_{22} = \begin{pmatrix} \Sigma & \Lambda \\ \Lambda' & I_q \end{pmatrix}, \quad \Gamma_{12} = \begin{pmatrix} \beta_1 \Sigma & \beta_1 \Lambda \end{pmatrix}.
\]

It follows that \(Y|x, u\) is Gaussian with mean \(m_{y|x, u} = \mathbb{E}(Y|x, u)\) and variance \(\sigma_{y|x, u}^2 = \text{Var}(Y|x, u)\), in accordance with the formulae in (A.1). Because the inverse matrix of \(\Gamma_{22}\) is required in (A.1), a formula for the inverse of a partitioned matrix is utilized:
\[
\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & (D - CA^{-1}B)^{-1} \end{pmatrix}.
\]
Again, writing $\Sigma = \Lambda \Lambda' + \Psi$, we have
\[
\Gamma^{-1}_{22} = \begin{pmatrix} \Sigma & \Lambda \\ \Lambda' & I_q \end{pmatrix}^{-1} = \begin{pmatrix} \Psi^{-1} & -\Sigma^{-1}\Lambda (I_q - \Lambda'\Sigma^{-1}\Lambda)^{-1} \\ -\Lambda'\Psi^{-1} & (I_q - \Lambda'\Sigma^{-1}\Lambda)^{-1} \end{pmatrix}.
\]

Moreover, according to the Woodbury identity (Woodbury, 1950), $\Sigma^{-1} = (\Lambda \Lambda' + \Psi)^{-1} = \Psi^{-1} - \Psi^{-1}\Lambda (I_q + \Lambda'\Psi^{-1}\Lambda)^{-1}\Lambda'\Psi^{-1}$. Now,
\[
\Gamma_{12}\Gamma^{-1}_{22} = \begin{pmatrix} \beta_1' \Sigma & \beta_1' \Lambda \end{pmatrix} \begin{pmatrix} \Psi^{-1} & -\Sigma^{-1}\Lambda (I_q - \Lambda'\Sigma^{-1}\Lambda)^{-1} \\ -\Lambda'\Psi^{-1} & (I_q - \Lambda'\Sigma^{-1}\Lambda)^{-1} \end{pmatrix} = \begin{pmatrix} \beta_1' & 0 \end{pmatrix}.
\]

Finally, according to (A.1), we have
\[
m_{y|x,u} = m_1 + \Gamma_{12}\Gamma^{-1}_{22} (z_2 - m_2) = (\beta_0 + \beta_1' \mu) + \begin{pmatrix} \beta_1' & 0 \end{pmatrix} \begin{pmatrix} x - \mu \\ u - 0 \end{pmatrix} = \beta_0 + \beta_1' x,
\]
\[
s^2_{y|x,u} = \Gamma_{11} - \Gamma_{12}\Gamma^{-1}_{22}\Gamma_{21} = \beta_1' \Sigma\beta_1 + \sigma^2_\varepsilon - \begin{pmatrix} \beta_1' & 0 \end{pmatrix} \begin{pmatrix} \Sigma\beta_1 \\ \Lambda\beta_1 \end{pmatrix} = \sigma^2_\varepsilon.
\]

Details on the EM-Algorithm

**First Cycle.** Here, $\theta_1 = \{(\pi_g, \beta_g, \mu_g, \sigma^2_g), g = 1, \ldots, G\}$, where the missing data are the unobserved group labels $Z = (z'_1, \ldots, z'_n)$. The complete data likelihood is
\[
L_1(\theta_1) = \prod_{i=1}^{n} \prod_{g=1}^{G} \left[ \phi \left( y_i|x_i, \beta_g, \sigma^2_g \right) \phi_d \left( x_i; \mu_g, \Sigma_g \right) \pi_g \right]^{z_{ig}}, \tag{A.2}
\]
where

\[
\phi(y_i|x_i; \beta, \sigma_g^2) = \frac{1}{\sqrt{2\pi\sigma_g^2}} \exp \left\{ -\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma_g^2} \right\}
\]

\[
\phi_d(x_i; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{d/2}} \exp \left\{ -\frac{1}{2} (x_i - \mu)\Sigma^{-1}(x_i - \mu) \right\}.
\]

Consider the complete-data log-likelihood

\[
\mathcal{L}_{c1}(\theta_1) = \sum_{i=1}^n \sum_{g=1}^G z_{ig} \ln \left[ \phi(y_i|x_i; \beta, \sigma_g^2) \phi_d(x_i; \mu, \Sigma) \pi_g \right]
\]

\[
= -\frac{n}{2} (d+1) \ln 2\pi - \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G z_{ig} \ln \sigma_g^2 - \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G z_{ig} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{\sigma_g^2} +
\]

\[
- \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G z_{ig} \ln |\Sigma_g| - \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G z_{ig} (x_i - \mu)^\prime \Sigma_g^{-1}(x_i - \mu) + \sum_{g=1}^G n_g \ln \pi_g,
\]

where \(n_g = \sum_{i=1}^n z_{ig}\). Because \(\Sigma_g = \Lambda \Lambda_g' + \Psi_g\), we get

\[
\mathcal{L}_{c1}(\theta_1) = -\frac{n}{2} (d+1) \ln 2\pi - \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G z_{ig} \ln \sigma_g^2
\]

\[
- \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G z_{ig} \left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma_g^2}\right)^2 - \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G z_{ig} \ln |\Lambda \Lambda_g' + \Psi_g|
\]

\[
- \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G z_{ig} \text{tr} \left\{ (x_i - \mu)^\prime (x_i - \mu) (\Lambda_g \Lambda_g' + \Psi_g)^{-1} \right\} + \sum_{g=1}^G n_g \ln \pi_g.
\]

(A.3)

The E-step on the first cycle of the \((k+1)\)-th iteration requires the calculation of \(Q_1(\theta_1; \theta^{(k)}) = \mathbb{E}_{\theta^{(k)}}[\mathcal{L}_c(\theta_1)| (X, y)]\), which is the expected complete-data log-likelihood given the data \((X, y)\) and using the current estimate \(\theta^{(k)}\) for \(\theta\). In practice,
it requires calculating \( \mathbb{E}_{\theta^{(k)}} \{ Z_{ig} | (X, y) \} \); this step is achieved by replacing each \( z_{ig} \) by \( z_{ig}^{(k+1/2)} \), where

\[
\begin{align*}
\hat{z}_{ig}^{(k+1)} &= \frac{\phi \left( y_i | \mu_{ig}, \beta_{ig}^{(k)}, \sigma_g^{2(k)} \right) \phi_d \left( x_i | \mu_g^{(k)} \right) \pi_g^{(k)}}{\sum_{g=1}^G \phi \left( y_i | \mu_{ig}^{(k)}, \beta_{ig}^{(k)}, \sigma_g^{2(k)} \right) \phi_d \left( x_i | \mu_g^{(k)} \right) \pi_g^{(k)}}. \quad (A.4)
\end{align*}
\]

On the M-step, the maximization of this complete-data log-likelihood yields

\[
\begin{align*}
\pi_g^{(k+1)} &= \frac{\sum_{i=1}^n z_{ig}^{(k+1)}}{n} \\
\mu_g^{(k+1)} &= \frac{1}{n_g} \sum_{i=1}^n z_{ig}^{(k+1)} x_i \\
\beta_{1g}^{(k+1)} &= \left[ \frac{1}{n_g} \sum_{i=1}^n z_{ig}^{(k+1)} y_i \left( x_i - \mu_{ig}^{(k+1)} \right) \right] \left[ \frac{1}{n_g} \sum_{i=1}^n z_{ig}^{(k+1)} x_i' x_i - \mu_{ig}^{(k+1)} \mu_{ig}^{(k+1)} \right]^{-1} \\
\beta_{0g}^{(k+1)} &= \frac{\sum_{i=1}^n z_{ig}^{(k+1)} y_i}{n_g} - \beta_{1g}^{(k+1)} \frac{\sum_{i=1}^n z_{ig}^{(k+1)} x_i}{n_g} \\
\sigma_g^{2(k+1)} &= \frac{1}{n_g} \sum_{i=1}^n z_{ig}^{(k+1)} \left\{ y_i - \left( \beta_{0g}^{(k+1)} + \beta_{1g}^{(k+1)} x_i \right) \right\}^2. \quad (A.5)
\end{align*}
\]

where \( n_g^{(k+1)} = \sum_{i=1}^n z_{ig}^{(k+1)} \). Following the notation in ?, we set \( \theta^{(k+1/2)} = (\theta_1^{(k+1)'}, \theta_2^{(k)'})' \).

**Second Cycle.** Here, \( \theta_2 = \{ \Sigma_g, g = 1, \ldots, G \} = \{ (\Lambda_g, \Psi_g), g = 1, \ldots, G \} \), where the missing data are the unobserved group labels \( Z \) and the latent factors \( U \). There-
fore, the complete data likelihood is

\[
L_{c2}(\theta_2) = \prod_{i=1}^{n} \prod_{g=1}^{G} \left[ \phi \left( y_i \mid x_i, u_{ig} ; \beta_g^{(k+1)}, \sigma_g^{2(k+1)} \right) \phi_d \left( x_i \mid u_{ig} ; \mu_g^{(k+1)}, \Sigma_g \right) \phi_g(u_{ig}) \pi_g^{(k+1)} \right]^{z_{ig}} \\
= \prod_{i=1}^{n} \prod_{g=1}^{G} \left[ \phi \left( y_i \mid x_i ; \beta_g^{(k+1)}, \sigma_g^{2(k+1)} \right) \phi_d \left( x_i \mid u_{ig} ; \mu_g^{(k+1)}, \Lambda_g, \Psi_g \right) \phi_g(u_{ig}) \pi_g^{(k+1)} \right]^{z_{ig}} ,
\]

(A.6)
because \( Y \) is conditionally independent of \( U \) given \( X = x \) and

\[
\phi_d \left( x_i \mid u_{ig} ; \mu_g^{(k+1)}, \Psi_g \right) = \frac{1}{2\pi |\Psi_g|^{1/2}} \exp \left\{ -\frac{1}{2} (x_i - \mu_g^{(k+1)} - \Lambda_g u_{ig})' \Psi_g^{-1} (x_i - \mu_g^{(k+1)} - \Lambda_g u_{ig}) \right\} \\
\phi_g(u_{ig}) = \frac{1}{(2\pi)^{q/2}} \exp \left\{ -\frac{1}{2} u_{ig}' u_{ig} \right\}.
\]

Hence, the complete data log-likelihood is

\[
L_{c2}(\theta_2) = -\frac{n(d + q + 1)}{2} \ln 2\pi - \frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \ln \sigma_g^{2(k+1)} \\
- \frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \frac{(y_i - \beta_{0g}) - \beta_{1g}^{(k+1)} x_i)^2}{2\sigma_g^2} \\
+ \sum_{g=1}^{G} n_g \ln \pi_g + \frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \ln |\Psi_g^{-1}| \\
\frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \tr \left\{ (x_i - \mu_g^{(k+1)} - \Lambda_g u_{ig}) (x_i - \mu_g^{(k+1)} - \Lambda_g u_{ig})' \Psi_g^{-1} \right\} ,
\]

(A.7)

where we set \( S_g^{(k+1)} = (1/n_g^{(k+1)}) \sum_{i=1}^{n} z_{ig}^{(k+1)} (x_i - \mu_g^{(k+1)}) (x_i - \mu_g^{(k+1)})' \). The E-step on

the second cycle of the \((k+1)\)-th iteration requires the calculation of \( Q_2(\theta_2 ; \theta^{(k+1/2)}) = \)

\[ \mathbb{E}_{\theta^{(k+1/2)}} \{ \mathcal{L}_{c2}(\theta) \| (X, y) \}. \]

This implies calculating the following conditional expectations:

\[ \mathbb{E}_{\theta^{(k+1/2)}} \{ Z_{ig} \| X, y, U_{ig} \}, \mathbb{E}_{\theta^{(k+1/2)}} \{ Z_{ig} U_{ig} \| X, y \}, \text{ and } \mathbb{E}_{\theta^{(k+1/2)}} \{ Z_{ig} U_{ig} U'_{ig} \| X, y \}. \]

Based on (4.7), these are given by

\[
\mathbb{E}_{\theta^{(k+1/2)}} [Z_{ig} U_{ig} \| X, y] = z_{ig}^{(k+1)} \gamma_g^{(k)} (x_i - \mu_g^{(k+1)}),
\]

\[
\mathbb{E}_{\theta^{(k+1/2)}} [Z_{ig} U_{ig} U'_{ig} \| X, y] = z_{ig}^{(k+1)} \{ I_q - \gamma_g^{(k)} \Lambda_g^{(k)} + \gamma_g^{(k)} S_g \gamma_g^{(k)} \} = z_{ig}^{(k+1)} \Theta_g^{(k)},
\]

where

\[
\gamma_g^{(k)} = \Lambda_g^{(k)} (\Lambda_g^{(k)} + \Psi_g^{(k)})^{-1} \quad \text{(A.9)}
\]

\[
\Theta_g^{(k)} = I_q - \gamma_g^{(k)} \Lambda_g^{(k)} + \gamma_g^{(k)} S_g^{(k+1)} \gamma_g^{(k)} \quad \text{(A.10)}
\]

Thus, the \( g \)-th term of the expected complete-data log-likelihood \( Q_2(\theta_2; \theta^{(k+1/2)}) \)

then becomes

\[
Q_2(\Lambda_g, \Psi_g; \theta^{(k+1/2)}) = C(\theta_1^{(k+1)}) + \frac{1}{2} n_g^{(k+1)} \ln |\Psi_g^{-1}| - \frac{1}{2} n_g^{(k+1)} \text{tr} \{ S_g^{(k+1)} \Psi_g^{-1} \}
\]

\[
+ n_g^{(k+1)} \text{tr} \{ \Lambda_g \gamma_g^{(k)} S_g^{(k+1)} \Psi_g^{-1} \} - \frac{1}{2} n_g^{(k+1)} \text{tr} \left\{ \Lambda_g' \Psi_g^{-1} \Lambda_g \Theta_g^{(k)} \right\},
\]

where \( C(\theta_1^{(k+1)}) \) denotes the terms in (A.8) that do not depend on \( \theta_2 \). Then (A.11)
is maximized for \((\hat{\Lambda}, \hat{\Psi})\) satisfying

\[
\frac{\partial Q_2}{\partial \Lambda_g} = n_g^{(k+1)} \Psi_g^{-1} S_g^{(k+1)} \gamma_g^{(k)'} - n_g^{(k+1)} \Psi_g^{-1} \Lambda_g \Theta_g^{(k)} = 0 \tag{A.12}
\]

\[
\frac{\partial Q_2}{\partial \Psi_g^{-1}} = \frac{1}{2} n_g^{(k+1)} \Psi_g - \frac{1}{2} n_g^{(k+1)} S_g^{(k+1)} + n_g^{(k+1)} S_g^{(k+1)'} \gamma_g^{(k)'} \Lambda_g' - \frac{1}{2} n_g^{(k+1)} \Lambda_g \Theta_g^{(k)} \Lambda_g' = 0.
\tag{A.13}
\]

Therefore,

\[
S_g^{(k+1)} \gamma_g^{(k)'} - \Lambda_g \Theta_g^{(k)} = 0 \tag{A.14}
\]

\[
\Psi_g - S_g^{(k+1)} + 2 S_g^{(k+1)'} \gamma_g^{(k)'} \Lambda_g' - \Lambda_g \Theta_g^{(k)} \Lambda_g' = 0. \tag{A.15}
\]

From (A.14), we get

\[
\hat{\Lambda}_g = S_g^{(k+1)} \gamma_g^{(k)} \Theta_g^{-1}, \tag{A.16}
\]

and substituting in (A.15) we get

\[
\Psi_g - S_g^{(k+1)} + 2 S_g^{(k+1)'} \gamma_g^{(k)'} (S_g^{(k+1)} \gamma_g^{(k)} \Theta_g^{-1})' - (S_g \hat{\gamma}_g' \Theta_g^{-1}) \Theta_g (S_g \hat{\gamma}_g' \Theta_g^{-1})' = 0.
\]

which yields

\[
\hat{\Psi}_g = \text{diag}\left\{S_g^{(k+1)} - \hat{\Lambda}_g \hat{\gamma}_g' S_g^{(k+1)}\right\}.
\tag{A.17}
\]

Details on Constraint on Covariance of X

1. **Model UUU**: no constraint is assumed.

2. **Model UUC**: \(\Psi_g = \psi_g I_d\), where the parameter \(\psi_g\) is updated according to
3. **Model UCU**: $\Psi_g = \Psi$, where the matrix $\Psi$ is updated according to (4.22).

4. **Model UCC**: $\Psi_g = \psi I_d$. By combining (4.19) and (4.22) we get:

$$
\hat{\psi} = \frac{1}{p} \sum_{g=1}^{G} \frac{n_g^{(k+1)}}{n} \text{tr}\{S_g^{(k+1)} - \hat{\Lambda}_g \gamma_g^{(k)} S_g^{(k+1)}\} = \frac{1}{p} \sum_{g=1}^{G} \pi_g^{(k+1)} \text{tr}\{S_g^{(k+1)} - \hat{\Lambda}_g \gamma_g^{(k)} S_g^{(k+1)}\}.
$$

(A.18)

Thus,

$$
\psi^+ = \frac{1}{p} \sum_{g=1}^{G} \pi_g^{(k+1)} \text{tr}\{S_g^{(k+1)} - \Lambda_g^+ \gamma_g S_g^{(k+1)}\}
$$

(A.19)

$$
\gamma_g^+ = \Lambda_g^+ (\Lambda_g^+ \Lambda_g^+ + \psi^+ I_d)^{-1},
$$

(A.20)

with $\Theta_g^+$ being computed according to (4.16).

5. **Model CUU**: $\Lambda_g = \Lambda$, where the matrix $\Lambda$ is updated according to (4.26). In this case, $\Psi_g$ is estimated directly from (A.15) and thus

$$
\Psi_g^+ = \text{diag}\{S_g^{(k+1)} - 2\Lambda^+ \gamma_g S_g^{(k+1)} + \Lambda^+ \Theta_g \Lambda^+\},
$$

(A.21)

with $\gamma_g$ and $\Theta_g^+$ being computed according to (4.27) and (4.28), respectively.

6. **Model CUC**: $\Lambda_g = \Lambda$ and $\Psi_g = \psi_g I_d$. In this case, equation (4.25), for $\Psi_g = \psi_g I_d$, yields

$$
\sum_{g=1}^{G} \frac{\partial Q_2(\Lambda, \psi_g; \theta^{(k+1/2)})}{\partial \Lambda} = \sum_{g=1}^{G} n_g^{(k+1)} \psi_g^{-1} S_g^{(k+1)} \gamma_g^{(k)} - \sum_{g=1}^{G} n_g^{(k+1)} \psi_g^{-1} \Theta_g^{(k)} = 0.
$$
and afterwards

\[ \hat{\Lambda} = \left( \sum_{g=1}^{G} n_{g}^{(k+1)} S_{g}^{(k+1)} \gamma_{g}^{(k)} \right) \left( \sum_{g=1}^{G} n_{g}^{(k+1)} \Lambda \right)^{-1}, \tag{A.22} \]

with \( \gamma_{g}^{(k)} = \Lambda^{(k)} \Lambda^{(k)} + \psi_{g}^{(k)} I_{d} \). Moreover, from

\[
\frac{\partial Q_2(\Lambda, \psi_{g}; \theta^{(k+1/2)})}{\partial \psi_{g}^{-1}} = \frac{p}{2} \psi_{g} - \frac{n_{g}^{(k+1)}}{2} \left[ \text{tr}\{S_{g}^{(k+1)}\} - 2 \text{tr}\{S_{g}^{(k+1)} \gamma_{g}^{(k)} \Lambda\} + \text{tr}\{\Lambda \Theta_{g}^{(k+1)} \Lambda\} \right]
\]

\[ = 0 \]

we get

\[ \hat{\psi}_{g} = \frac{1}{p} \text{tr}\{S_{g}^{(k+1)}\} - 2 \hat{\Lambda} \gamma_{g}^{(k)} S_{g} + \hat{\Lambda} \Theta_{g}(\hat{\Lambda})'. \]

Thus,

\[ \Lambda^{+} = \left( \sum_{g=1}^{G} n_{g}^{(k+1)} S_{g}^{(k+1)} \gamma_{g}^{(k)} \right) \left( \sum_{g=1}^{G} n_{g}^{(k+1)} \Lambda \right)^{-1}, \tag{A.23} \]

\[ \psi_{g}^{+} = \frac{1}{p} \text{tr}\{S_{g}^{(k+1)}\} - 2 \Lambda^{+} \gamma_{g}^{(k)} S_{g} + \Lambda^{+} \Theta \Lambda^{+}, \tag{A.24} \]

\[ \gamma_{g}^{+} = \Lambda^{+} \Lambda^{+} + \psi_{g}^{+} I_{d}^{-1}. \tag{A.25} \]

with \( \Theta_{g}^{+} \) being computed according to (4.28).

7. **Model CCU**: \( \Lambda_{g} = \Lambda \) and \( \Psi_{g} = \Psi \), so that \( \gamma^{(k)} = \Lambda^{(k)} \Lambda^{(k)} + \Psi^{(k)} \).
Setting $\Psi_g = \Psi$ in (4.25), we get

$$
\frac{\partial Q_2(\Lambda, \Psi; \theta^{(k+1/2)})}{\partial \Lambda} = \sum_{g=1}^{G} n_g^{(k+1)} \Psi^{-1} \left[ S_g^{(k+1)} \gamma' - \Lambda \Theta_g^{(k)} \right]
$$

$$
= \Psi^{-1} \left[ \gamma'(k) \sum_{g=1}^{G} n_g^{(k+1)} S_g^{(k+1)} - \Lambda \sum_{g=1}^{G} n_g^{(k+1)} \Theta_g^{(k)} \right]
$$

$$
= \Psi^{-1} \left[ \gamma'(k) S^{(k+1)} - \Lambda \Theta^{(k)} \right] = 0.
$$

where

$$
S^{(k+1)} = \sum_{g=1}^{G} \pi_g^{(k+1)} S_g^{(k+1)}
$$

$$
\Theta^{(k)} = \sum_{g=1}^{G} \pi_g^{(k+1)} \Theta_g^{(k)} = I_q - \gamma'(k) \Lambda^{(k)} + \gamma'(k) S^{(k+1)} \gamma(k).
$$

Thus,

$$
\hat{\Lambda} = S^{(k+1)} \gamma'(k) (\Theta^{(k)})^{-1}.
$$

Moreover, setting $\Lambda_g = \Lambda$ in (4.21), we get

$$
\hat{\Psi} = \text{diag}\{ S^{(k+1)} - \Lambda \gamma'(k) S^{(k+1)} \}.
$$

(A.26)

Hence,

$$
\Lambda^+ = S^{(k+1)} \gamma' \Theta^{-1}
$$

(A.27)

$$
\Psi^+ = \text{diag}\{ S^{(k+1)} - \Lambda^+ \gamma' S^{(k+1)} \}
$$

(A.28)

$$
\gamma_g^+ = \Lambda'^+ (\Lambda^+ \Lambda'^+ + \Psi^+)^{-1},
$$

(A.29)
with $\Theta_g^+$ being computed according to (4.28).

8. **Model CCC**: $\Lambda_g = \Lambda$ and $\Psi_g = \psi I_d$, so that $\gamma^{(k)} = \Lambda^{(k)} (\Lambda^{(k)} + \psi^{(k)})^{-1}$. Here, the estimated loading matrix is again (7), while the isotropic term obtained from (A.18) for $\Lambda_g = \Lambda$ is

$$\hat{\psi} = \frac{1}{p} \text{tr}\{S^{(k+1)} - \hat{\Lambda} \gamma^{(k)} S^{(k+1)}\},$$

with $\gamma^{(k)} = \Lambda^{(k)} (\Lambda^{(k)} + \psi^{(k)})^{-1}$. Hence,

$$\psi^+ = \frac{1}{p} \text{tr}\{S^{(k+1)} - \Lambda^{+} \gamma S^{(k+1)}\} \quad (A.30)$$

$$\gamma^+ = \Lambda^+ (\Lambda^+ + \psi^+ I_d)^{-1}. \quad (A.31)$$

with $\Lambda^+$ and $\Theta_g^+$ being computed according to (A.27) and (4.28), respectively.

### A.2.2 Covariance Matrix

**Example 1**

Because the loading matrix is not unique, we compared the covariance matrices used for data simulation to the estimated covariance matrices. The values for
\( \Sigma_g \) and \( \dot{\Sigma}_g \) were

\[
\Sigma_1 = \begin{bmatrix}
103.36 & 103.07 & 101.37 & 79.41 & 105.66 \\
103.08 & 119.39 & 110.23 & 85.97 & 115.47 \\
101.37 & 110.23 & 129.77 & 106.08 & 118.50 \\
79.41 & 85.97 & 106.08 & 101.46 & 95.21 \\
105.66 & 115.47 & 118.50 & 95.21 & 121.63
\end{bmatrix},
\]

\[
\dot{\Sigma}_1 = \begin{bmatrix}
107.59 & 114.55 & 110.42 & 87.29 & 114.43 \\
114.55 & 139.40 & 127.06 & 100.09 & 132.06 \\
110.42 & 127.06 & 146.31 & 122.92 & 134.12 \\
87.29 & 100.09 & 122.92 & 117.97 & 110.09 \\
114.43 & 132.06 & 134.12 & 110.09 & 135.66
\end{bmatrix},
\]

\[
\Sigma_2 = \begin{bmatrix}
34.25 & 15.16 & 17.81 & 22.39 & 14.62 \\
15.16 & 17.01 & 11.42 & 13.98 & 8.95 \\
17.81 & 11.42 & 17.62 & 16.12 & 10.45 \\
14.62 & 8.95 & 10.45 & 13.11 & 10.19
\end{bmatrix},
\]

and

\[
\dot{\Sigma}_2 = \begin{bmatrix}
22.16 & 7.44 & 13.71 & 12.89 & 10.12 \\
7.44 & 11.25 & 7.59 & 8.05 & 5.48 \\
12.89 & 8.05 & 13.53 & 22.00 & 9.41 \\
10.12 & 5.48 & 10.13 & 9.41 & 8.63
\end{bmatrix}.
\]
Example 2

\[
\Sigma_1 = \begin{bmatrix}
10.41 & 3.61 & 4.07 & 4.48 & 5.71 \\
3.61 & 7.83 & 2.88 & 3.18 & 4.03 \\
4.07 & 2.88 & 8.67 & 3.81 & 4.64 \\
4.48 & 3.18 & 3.81 & 9.61 & 5.17 \\
5.71 & 4.04 & 4.64 & 5.17 & 11.73
\end{bmatrix},
\]

\[
\hat{\Sigma}_1 = \begin{bmatrix}
8.86 & 3.89 & 5.06 & 3.84 & 5.72 \\
3.89 & 7.23 & 3.59 & 1.79 & 4.04 \\
5.06 & 3.59 & 8.44 & 3.85 & 5.50 \\
3.84 & 1.79 & 3.85 & 7.74 & 4.38 \\
5.72 & 4.04 & 5.50 & 4.38 & 9.81
\end{bmatrix},
\]

\[
\Sigma_2 = \begin{bmatrix}
103.36 & 103.07 & 101.37 & 79.41 & 105.66 \\
103.08 & 122.1 & 110.23 & 85.97 & 115.47 \\
101.37 & 110.23 & 134.33 & 106.08 & 118.50 \\
79.41 & 85.97 & 106.08 & 102.73 & 95.21 \\
105.66 & 115.47 & 118.50 & 95.21 & 129.21
\end{bmatrix},
\]

\[
\hat{\Sigma}_2 = \begin{bmatrix}
106.17 & 100.46 & 93.18 & 73.81 & 105.01 \\
100.46 & 113.71 & 92.97 & 72.22 & 107.88 \\
93.18 & 92.97 & 108.25 & 83.08 & 102.36 \\
73.81 & 72.22 & 83.08 & 80.09 & 81.85 \\
105.01 & 107.88 & 102.36 & 81.85 & 122.59
\end{bmatrix},
\]
A.3 Clustering Gene Expression Time Course Data Using Mixtures of Multivariate t-Distributions

A.3.1 Estimation of $T_1$, $T_2$, $\hat{T}_1$, and $\hat{T}_2$ from the second simulation study

$\Sigma_3 = \begin{bmatrix}
25.19 & 15.16 & 17.81 & 22.39 & 14.62 \\
15.16 & 10.67 & 11.42 & 13.98 & 8.95 \\
17.81 & 11.42 & 13.12 & 16.12 & 10.45 \\
22.39 & 13.98 & 16.12 & 20.31 & 13.11 \\
14.62 & 8.95 & 10.45 & 13.11 & 8.70
\end{bmatrix}$

and

$\hat{\Sigma}_3 = \begin{bmatrix}
32.47 & 19.91 & 23.06 & 28.78 & 18.80 \\
19.91 & 14.10 & 14.96 & 18.25 & 11.66 \\
23.06 & 14.96 & 16.95 & 20.77 & 13.45 \\
28.78 & 18.25 & 20.77 & 25.95 & 16.77 \\
18.80 & 11.66 & 13.45 & 16.77 & 11.10
\end{bmatrix}$

$T_1$, $T_2$, $\hat{T}_1$, and $\hat{T}_2$ from the second simulation study

$T_1 = \begin{bmatrix}
1.00 \\
0.20 & 1.00 \\
0.05 & 0.30 & 1.00 \\
0.01 & 0.20 & 0.35 & 1.00
\end{bmatrix}$, $T_2 = \begin{bmatrix}
1.00 \\
0.75 & 1.00 \\
0.25 & 0.75 & 1.00 \\
0.05 & 0.25 & 0.75 & 1.00
\end{bmatrix}$
\[ \hat{T}_1 = \begin{bmatrix} 1.00 \\ 0.42 & 1.00 \\ 0.12 & 0.47 & 1.00 \\ 0.07 & 0.20 & 0.58 & 1.00 \end{bmatrix}, \quad \hat{T}_2 = \begin{bmatrix} 1.00 \\ 0.86 & 1.00 \\ 0.12 & 0.46 & 1.00 \\ 0.07 & 0.20 & 0.58 & 1.00 \end{bmatrix}. \]

\( \hat{T}_1 \) and \( \hat{T}_2 \) from the third simulation study using \( t \)-distribution

\[ \hat{T}_1 = \begin{bmatrix} 1.00 \\ 0.73 & 1.00 \\ 0.20 & 0.70 & 1.00 \\ 0.06 & 0.18 & 0.65 & 1.00 \end{bmatrix}, \quad \hat{T}_2 = \begin{bmatrix} 1.00 \\ 1.18 & 1.00 \\ 0.07 & 0.34 & 1.00 \\ -0.02 & 0.22 & 0.43 & 1.00 \end{bmatrix}. \]

\( \hat{T}_1, \hat{T}_2 \) and \( \hat{T}_3 \) from the third simulation study using Gaussian distribution

\[ \hat{T}_1 = \begin{bmatrix} 1.00 \\ 0.72 & 1.00 \\ 0.22 & 0.72 & 1.00 \\ 0.02 & 0.18 & 0.65 & 1.00 \end{bmatrix}, \quad \hat{T}_2 = \begin{bmatrix} 1.00 \\ 0.19 & 1.00 \\ 0.11 & 0.34 & 1.00 \\ -0.02 & 0.24 & 0.54 & 1.00 \end{bmatrix}, \quad \hat{T}_3 = \begin{bmatrix} 1.00 \\ 0.12 & 1.00 \\ -0.09 & 0.24 & 1.00 \\ -0.17 & 0.22 & 0.31 & 1.00 \end{bmatrix}. \]
A.3.2 Illustration of Constraining Degrees of Freedom

The degrees of freedom parameter can also be constrained to be equal across groups; i.e., we could set $\nu_g = \nu$ for $g = 1, \ldots, G$. The resulting models will be somewhat more parsimonious but, as illustrated by Andrews and McNicholas (2011a,b), there is also the possibility that an overall estimate of degrees of freedom will result in a better model fit and superior clustering performance than separate component-specific estimates. In a model with $\nu_g = \nu$, we have

$$ l_{2c}(\nu) = \sum_{g=1}^{G} \sum_{i=1}^{n} z_{ig} \left[ -\log \Gamma\left(\frac{\nu}{2}\right) + \frac{\nu}{2} \log \left(\frac{\nu}{2}\right) + \frac{\nu}{2} \left(\log u_{ig} - u_{ig}\right) - \log u_{ig}\right]. $$

Replacing the $z_{ig}$ and the $u_{ig}$ by their respective expected values (equations 5.2 and 5.3), differentiating the resulting function with respect to $\nu$ and setting it equal to zero amounts to solving the equation

$$ 1 - \varphi \left(\frac{\nu_{\text{new}}}{2}\right) + \frac{1}{n} \sum_{g=1}^{G} \sum_{i=1}^{n} \hat{z}_{ig} (\log \hat{u}_{ig} - \hat{u}_{ig}) + \log \left(\frac{\nu_{\text{new}}}{2}\right) + \varphi \left(\frac{\nu_{\text{old}} + p}{2}\right) - \log \left(\frac{\nu_{\text{old}} + p}{2}\right) = 0, $$

for $\nu_{\text{new}}$, where $\nu_{\text{old}}$ is the previous estimate of $\nu$ and, again, $\varphi(\cdot)$ is the digamma function.

In terms of the analyses carried out in this paper, we did not run models with constrained degrees of freedom. We took the view that our sample sizes were large enough to get sensible estimates of the component degrees of freedom and that to entertain constrained degrees of freedom would be to throw away some of the advantages of flexibility in modelling that our approach brings. In the results of our two real data analyses, it was clear that one cluster in each case required a much larger value for degrees of freedom than the others. An effective averaging of degrees of freedom.
freedom in these analyses would have led to different results. An illustrative example follows, based on the yeast data.

Figure A.1: The temporal patterns for the four clusters selected by our best model with constrained degrees of freedom on the yeast data.

Constraining the degrees of freedom to be equal across groups and then running our eight mixture models on the yeast data set within the model-based clustering framework for $G = 1, \ldots, 14$, resulted in the selection of a model with $G = 4$ components. Upon visual examination of individual components (Figure A.1), the algorithm seemed to capture the trend in the data. However, genes in component 3 might be better represented by two distinct clusters; upon closer inspection, they each appear to exhibit one of two distinct expression patterns (Figure A.2). A comparison of the classifications using models with constrained and unconstrained degrees of freedom showed that component 3 of the constrained model is effectively a mixture of components 3 and 1 from the unconstrained model (cf. Section 5.3.2). Out of the
79 observations assigned to component 3 in Section 5.3.2, 78 were in component 3 in the analysis with constrained degrees of freedom. The other 97 observations from component 3 of the constrained model were in component 1 in Section 5.3.2 and the remaining 6 were in component 2. This suggests that the four groups obtained by imposing the constraint on the degrees of freedom are a result of merging groups with clearly different correlation structures. Most concerning, however, is that the model with constrained degrees of freedom had a higher BIC than the model from Section 5.3.2. Based on experiences like this one, we did not consider models with constrained degrees of freedom in this paper.

![Cluster 3](image)

Figure A.2: Two apparently distinct temporal patterns for component 3 of our best model with constrained degrees of freedom on the yeast data.
Table A.2: Posterior expected value of the parameters of the eigen-decomposed covariance matrix.

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameters</th>
<th>Expected Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>EII</td>
<td>$\lambda I$</td>
<td>$\mathbb{E}[(\lambda)^{-1}] = \frac{d}{b}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\mathbb{E}[\log</td>
</tr>
<tr>
<td>VII</td>
<td>$\lambda_g I$</td>
<td>$\mathbb{E}[(\lambda_g)^{-1}] = \frac{b}{b_g}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\mathbb{E}[\log</td>
</tr>
<tr>
<td>EEI</td>
<td>$\lambda A$</td>
<td>$\mathbb{E}[(\lambda A)<em>{k,k}^{-1}] = \frac{b</em>{2k}}{b_k}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\mathbb{E}[\log</td>
</tr>
<tr>
<td>VEI</td>
<td>$\lambda_g A$</td>
<td>$\mathbb{E}[\lambda_g^{-1}] = \frac{b_g}{b}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\mathbb{E}[\log</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\mathbb{E}[(c A^{-1})<em>{k,k}] = \frac{a_k}{b</em>{2k}}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\mathbb{E}[\log</td>
</tr>
<tr>
<td>EVI</td>
<td>$\lambda A_g$</td>
<td>$\mathbb{E}[\lambda^{-1}] = \frac{a}{b}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\mathbb{E}[\log</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\mathbb{E}[(c_g A^{-1})<em>{k,k}] = \frac{a</em>{2k}}{b_{2k}}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\mathbb{E}[\log</td>
</tr>
<tr>
<td>VVI</td>
<td>$\lambda_g A_g$</td>
<td>$\mathbb{E}[\lambda_g^{-1}] = \frac{b_g}{b}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\mathbb{E}[\log</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\mathbb{E}[(c_g A^{-1})<em>{k,k}] = \frac{a</em>{2k}}{b_{2k}}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\mathbb{E}[\log</td>
</tr>
<tr>
<td>EEE</td>
<td>$\lambda DAD'$</td>
<td>$\mathbb{E}[\log</td>
</tr>
<tr>
<td>VEE</td>
<td>$\lambda_g DAD'$</td>
<td>$\mathbb{E}[\lambda_g^{-1}] = \frac{b_g}{b}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\mathbb{E}[\log</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\mathbb{E}[(DAD')^{-1}] = \frac{b}{v}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\mathbb{E}[\log</td>
</tr>
<tr>
<td>EEV</td>
<td>$\lambda D_g AD'$</td>
<td>$\mathbb{E}[(\lambda A)<em>{k,k}] = \frac{b</em>{2k}}{b_k}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\mathbb{E}[\log</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\mathbb{E}[(\lambda D_g AD')^{-1}</td>
</tr>
<tr>
<td>VEV</td>
<td>$\lambda_g D_g AD'$</td>
<td>$\mathbb{E}[\lambda_g^{-1}] = \frac{b_g}{b}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\mathbb{E}[\log</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\mathbb{E}[(c A^{-1})<em>{k,k}] = \frac{a</em>{2k}}{b_{2k}}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\mathbb{E}[\log</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\mathbb{E}[(\lambda_g^{-1} D_g AD_g')^{-1}</td>
</tr>
<tr>
<td>EVV</td>
<td>$\lambda D_g A_g D_g'$</td>
<td>$\mathbb{E}[\lambda] = \frac{b}{v}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\mathbb{E}[\log</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\mathbb{E}[(D_g A_g D_g')^{-1}] = \frac{v}{v_g}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\mathbb{E}[\log</td>
</tr>
<tr>
<td>VVV</td>
<td>$\lambda_g D_g A_g D_g'$</td>
<td>$\mathbb{E}[\lambda_g^{-1}] = \frac{b_g}{b}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\mathbb{E}[\log</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\mathbb{E}[(D_g A_g D_g')^{-1}] = \frac{v}{v_g}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\mathbb{E}[\log</td>
</tr>
</tbody>
</table>