Longitudinal Clustering via Mixtures of Multivariate Power Exponential Distributions

LONGITUDINAL CLUSTERING VIA MIXTURES OF MULTIVARIATE POWER EXPONENTIAL DISTRIBUTIONS

BY

NIDHI PATEL, B.Sc.

A THESIS

SUBMITTED TO THE DEPARTMENT OF MATHEMATICS & STATISTICS

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

Master of Science

Master of Science (2016) McMaster University

(Statistics) Hamilton, Ontario, Canada

TITLE:	Longitudinal Clustering via Mixtures of Multivariate Power Exponential Distributions
AUTHOR:	Nidhi Patel B.Sc., (Mathematics and Statistics) McMaster University, Hamilton,Canada
SUPERVISOR:	Dr. Paul D. McNicholas
NUMBER OF PAGES:	ix, 45

To my family and friends

Abstract

A mixture model approach for clustering longitudinal data is introduced. The approach, which is based on mixtures of multivariate power exponential distributions, allows for varying tail-weight and peakedness in data. In the longitudinal setting, this corresponds to more or less concentration around the most central time course in a component. The models utilize a modified Cholesky decomposition of the component scale matrices. The associated maximum likelihood estimators are derived via a generalized expectation-maximization algorithm.

Acknowledgements

This work is partly supported by the Canada Research Chairs program. This work is done in collaboration with Dr. Utkarsh J. Dang and Dr. Paul D. McNicholas. I would like to thank them for their support and guidance. I would also like to thank Dr. Roman Viveros-Aguilera and Dr. Ben Bolker for serving on my examination committee.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Longitudinal data arise from a set of measurements taken repeatedly over time. Measurements taken on each subject are correlated, so careful consideration needs to be taken when modeling the correlation structures. Longitudinal studies enable one to observe changes and patterns at an individual level over time. These observed changes and patterns in the data may naturally form groups or clusters that give more insight into the data. When groups are formed in a way where subjects are a lot more similar within a group than between groups and true memberships are unknown, then this is known as cluster analysis.

Herein, model-based clustering is used, which is a technique where data is clustered by utilizing finite mixture models. The general form of the finite mixture density is defined as

$$f(\mathbf{x}|\boldsymbol{\Theta}) = \sum_{g=1}^{G} \pi_g \phi(\mathbf{x}|\boldsymbol{\Theta}_g), \tag{1.1}$$

where $\phi(\mathbf{x}|\boldsymbol{\Theta}_g)$ are the component densities, $\boldsymbol{\Theta}_g$ is a vector of the parameters, G is the total number of groups and π_g is the probability of belonging to group g, where $\pi_g \in$

$(0, 1]$ and $\sum_{g=1}^{G} \pi_g = 1$. For clustering purposes, the component densities are usually taken to be of the same type. Each unimodal component of the mixture models represents a group or cluster (McNicholas, 2016a). It is crucial for the component to be unimodal because if it is not, then probably the wrong mixture distribution is being fit or not enough components are being used (McNicholas, 2016b).

There has been an increasing interest in model-based clustering. Multivariate Gaussian mixture models have been commonly used for model-based clustering. Examples of papers that use multivariate Gaussian mixture models for clustering are: Bouveyron *et al.* (2007), Browne and McNicholas (2014a), Browne and McNicholas (2014b), Celeux and Govaert (1995), Fraley and Raftery (2002), McLachlan and Peel (2000), and McNicholas and Murphy (2008). Due to the mathematical tractability of the multivariate Gaussian distribution it is commonly used, but it is not robust to tail weight and peakedness. McLachlan and Peel (1998) and Peel and McLachlan (2000) proposed the multivariate t-distribution as an alternative to the multivariate Gaussian distribution because it has a heavier tails compared to the Gaussian distribution. Examples of other papers that use multivariate t mixture models for clustering are: Andrews and McNicholas (2011), Andrews *et al.* (2011), Andrews and McNicholas (2012), and Lin *et al.* (2014). The option of constraining the degrees of freedom across all clusters was explored by Andrews and McNicholas (2011), to see if it leads to an improved classification performance. The multivariate t-distribution works well for clustering data with a heavier tail weight, but cannot handle data with a lighter tail weight than the Gaussian distribution. Dang *et al.* (2015) proposed the multivariate power exponential (MPE) distribution as an alternative to the multivariate Gaussian and t-distributions because of its robustness to tail weight (light or

heavy) and peakedness (thinner or flatter).

The majority of the literature focuses on classical model-based clustering rather than model-based clustering for longitudinal data. McNicholas and Murphy (2010) introduces a family of Gaussian mixture models with covariance structures that are specifically designed to be used for clustering longitudinal data. For longitudinal data, a covariance structure that takes into account the relationship between measurements at different time points is crucial. This is done by utilizing the modified Cholesky decomposition for the covariance matrix $\mathbf{\Sigma}$. This decomposition uses the fact that the covariance matrix, $\mathbf{\Sigma}$, can be decomposed using the relation $\mathbf{T\Sigma T'} = \mathbf{D}$, where $\mathbf{T}$ is a unique lower unit triangular matrix and $\mathbf{D}$ is a unique diagonal matrix with strictly positive entries (Pourahmadi, 1999, 2000). The relation, $\mathbf{T\Sigma T'} = \mathbf{D}$, can also be written as

$$\mathbf{\Sigma}^{-1} = \mathbf{T'D}^{-1}\mathbf{T}, \tag{1.2}$$

which provides convenience for when this decomposition is used for mixture models. The matrices $\mathbf{T}$ and $\mathbf{D}$ are interpreted as generalized autoregressive parameters and innovation variances, respectively (Pourahmadi, 1999). Therefore, the linear least squares predicator of Y_t, based on $Y_{t-1}, ..., Y_1$, can be written as

$$\hat{Y}_t = \mu_t + \sum_{s=1}^{t-1} (-\phi_{ts}) (Y_s - \mu_s) + \sqrt{d_t}\epsilon_t, \tag{1.3}$$

where $\epsilon_t \sim \mathcal{N}(0, 1)$, ϕ_{ts} are the sub-diagonal elements of $\mathbf{T}$ and d_t are the diagonal elements of $\mathbf{D}$ (McNicholas and Murphy, 2010). Imposing constraints on the component densities allows for a family of mixture models to be created. Let $\mathbf{T}_g$ and $\mathbf{D}_g$ be the autoregressive parameters and innovation variances for a specific group

g, respectively. McNicholas and Murphy (2010) imposed various constraints on the covariance structure, specifically the matrices $\mathbf{T}_g$ and $\mathbf{D}_g$. There was an option of constraining either $\mathbf{T}_g$ or $\mathbf{D}_g$, or both to be equal across all groups. In other words, there is an option to restrict the autoregressive structure or the noise or the entire covariance structure to be the same for all groups. Constraining $\mathbf{T}_g$ to be equal across all groups, i.e. $\mathbf{T}_g = \mathbf{T}$, means that the correlation structures are the same for all the groups. Constraining $\mathbf{D}_g$ to be equal across all the groups, i.e. $\mathbf{D}_g = \mathbf{D}$, means that the variability at each time point is the same for each group. There is also an option to impose the isotropic constraint, $\mathbf{D}_g = \delta_g \mathbf{I}_p$ (cf. Tipping and Bishop, 1999), which means the variability is the same at all time points in component g. McNicholas and Murphy (2010) used the above constraints that resulted in a family of eight mixture models, which is shown in Table 1.1. McNicholas and Subedi (2012) explored the idea of clustering longitudinal data using the multivariate t-distribution. The modified Cholesky decomposed covariance structure is also used and the same constraints are possible with the Gaussian mixture models. As mentioned above, the idea of constraining the degrees of freedom has been explored, but McNicholas and

Table 1.1: The nomenclature, covariance structures and number of covariance parameters for Gaussian and t mixture models

Model	$\mathbf{T}_g$	$\mathbf{D}_g$	$\mathbf{D}_g$	Number of Covariance Parameters
EEA	Equal	Equal	Anisotropic	$p(p-1)/2+p$
VVA	Variable	Variable	Anisotropic	$Gp(p-1)/2+Gp$
VEA	Variable	Equal	Anisotropic	$Gp(p-1)/2+p$
EVA	Equal	Variable	Anisotropic	$p(p-1)/2+Gp$
VVI	Variable	Variable	Isotropic	$Gp(p-1)/2+G$
VEI	Variable	Equal	Isotropic	$Gp(p-1)/2+1$
EVI	Equal	Variable	Isotropic	$p(p-1)/2+G$
EEI	Equal	Equal	Isotropic	$p(p-1)/2+1$

Subedi (2012) does not impose this constraint. Therefore, the family of t mixture models consists of eight models and are the same as the models in Table 1.1. This paper aims to utilize the MPE distribution for clustering longitudinal data.

The MPE distribution is also known as the multivariate generalized Gaussian distribution. The shape parameter, β, controls the distribution's tail weight and peakedness, so two different kinds of distributions can be obtained. A leptokurtic distribution is associated with a thinner peak and heavy tails compared to the Gaussian distribution, which is obtained when $0 < \beta < 1$. A platykurtic distribution is associated with flatter peak and thin tails compared to the Gaussian distribution, which is obtained when $\beta > 1$. Due to the shape parameter, the MPE distribution is very flexible and can produce common distributions from the exponential family. When $\beta = 0.5$ and $\beta = 1$, the MPE distribution is equivalent to the Laplace distribution and the Gaussian distribution, respectively. The MPE distribution converges to the multivariate uniform distribution when $\beta \to \infty$. Figure 1.1 displays density plots for different β values and it shows the flexibility of the MPE distribution.

In Chapter 2, we look into mixtures of the MPE model for longitudinal data and the relevant statistical inferences. In Chapter 3, longitudinal clustering is applied to real datasets and compared to the performances of the mixtures of the multivariate Gaussian and mixtures of the multivariate t-distributions. This chapter also looks at the option of constraining the sub-diagonals of $\mathbf{T}_g$ and this is investigated by applying it to a real dataset. In Chapter 4, model-based classification is introduced and applied to a real dataset. In Chapter 5, the results are summarized and future works are discussed.

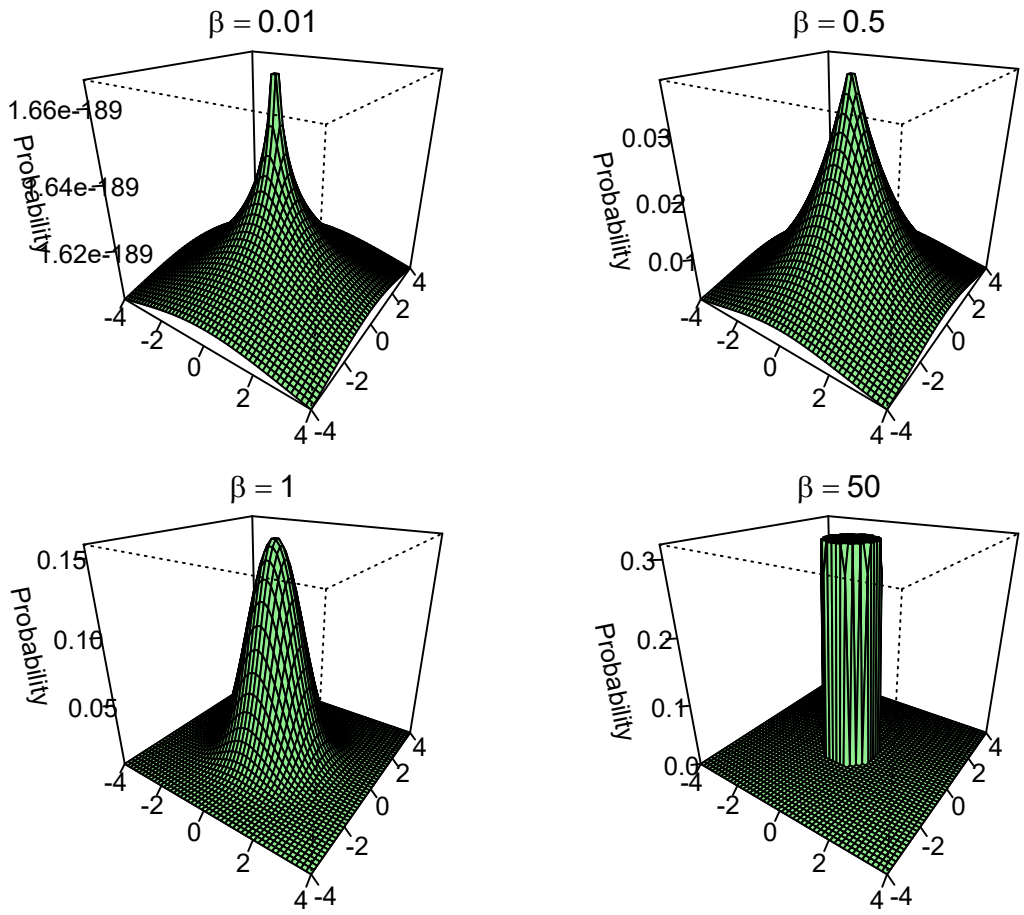

Figure 1.1: Density plots of bivariate MPE distribution for different β values.

Chapter 2

Methodology

2.1 The Model

For a p-dimensional random vector $\mathbf{X}$, the MPE distribution's density has the form of

$$g(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, r, s) = c_p \left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}} \exp\left\{-\frac{r}{2^s}\delta(\mathbf{x})^s\right\}, \tag{2.1}$$

where

$$c_p = \frac{s\Gamma\left(\frac{p}{2}\right)}{(2\pi)^{p/2}\Gamma\left(\frac{p}{2s}\right)} r^{p/(2s)},$$

$\delta(\mathbf{x}) := \delta(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$, $\boldsymbol{\mu}$ is the location parameter, $\boldsymbol{\Sigma}$ is a positive-definite scale matrix, $r, s > 0$ and $\mathbf{x}$ is a random vector (Landsman and Valdez, 2003). This form of the density was not used by Dang *et al.* (2015) because of identifiability issues with $\boldsymbol{\Sigma}$ and r, so a reparametrized form of the density was used, which was given by Gómez *et al.* (1998a). The reparametrization of (2.1) is done by taking $r = 2^{\beta-1}$ and $s = \beta$, where β is a shape parameter that determines the kurtosis. Take $\mathbf{X}$ to be a random vector with p dimensions that follows the MPE distribution and

the density has the form

$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \beta) = k\,|\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\delta(\mathbf{x})^{\beta}\right\}, \tag{2.2}$$

where

$$k = \frac{p\Gamma\left(\frac{p}{2}\right)}{(\pi^{p/2})\Gamma\left(1+\frac{p}{2\beta}\right)2^{1+\frac{p}{2\beta}}},$$

and the rest of the parameters are the same as in (2.1). When the parameterized MPE density (2.2) is replaced in the mixture model density (1.1) it gives the density for mixtures of MPE, which is defined as

$$h(\mathbf{x}|\boldsymbol{\Theta}) = \sum_{g=1}^{G} \pi_g f(\mathbf{x}|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \beta_g), \tag{2.3}$$

where $f(\mathbf{x}|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \beta_g)$ is the g^{th} group's density with its respective parameters. The modified Cholesky decomposition is used for the covariance structure, in order to correctly model longitudinal data. The mixtures of MPE model's density (2.3) with the modified Cholesky decomposed covariance structure for a specific group is

$$f(x_i|\boldsymbol{\mu}_g, \mathbf{T}_g, \mathbf{D}_g, \beta_g) = k_g\,|\mathbf{D}_g|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left[(\mathbf{x}_i-\boldsymbol{\mu}_g)'\mathbf{T}_g'\mathbf{D}_g^{-1}\mathbf{T}_g(\mathbf{x}_i-\boldsymbol{\mu}_g)\right]^{\beta_g}\right\}, \tag{2.4}$$

where $\mathbf{T}_g'\mathbf{D}_g^{-1}\mathbf{T}_g = \boldsymbol{\Sigma}_g^{-1}$, $\mathbf{T}_g$ and $\mathbf{D}_g$ are $p \times p$ matrices and

$$k_g = \frac{p\Gamma\left(\frac{p}{2}\right)}{(\pi^{p/2})\Gamma\left(1+\frac{p}{2\beta_g}\right)2^{1+\frac{p}{2\beta_g}}}.$$

Like the families of Gaussian and t mixture models, the MPE distribution also has the option of constraining $\mathbf{T}$, $\mathbf{D}$ or $\boldsymbol{\Sigma}$, which would lead to the same family of eight mixture models from Table 1.1. However, the MPE distribution has an additional

Table 2.1: MPE mixture model family with its respective nomenclature and number of covariance parameters

Model	$\mathbf{T}_g$	$\mathbf{D}_g$	$\mathbf{D}_g$	β_g	Number of Covariance Parameters
EEAE	Equal	Equal	Anisotropic	Equal	$p(p-1)/2+p$
VVAE	Variable	Variable	Anisotropic	Equal	$Gp(p-1)/2+Gp$
VEAE	Variable	Equal	Anisotropic	Equal	$Gp(p-1)/2+p$
EVAE	Equal	Variable	Anisotropic	Equal	$p(p-1)/2+Gp$
VVIE	Variable	Variable	Isotropic	Equal	$Gp(p-1)/2+G$
VEIE	Variable	Equal	Isotropic	Equal	$Gp(p-1)/2+1$
EVIE	Equal	Variable	Isotropic	Equal	$p(p-1)/2+G$
EEIE	Equal	Equal	Isotropic	Equal	$p(p-1)/2+1$
EEAV	Equal	Equal	Anisotropic	Variable	$p(p-1)/2+p$
VVAV	Variable	Variable	Anisotropic	Variable	$Gp(p-1)/2+Gp$
VEAV	Variable	Equal	Anisotropic	Variable	$Gp(p-1)/2+p$
EVAV	Equal	Variable	Anisotropic	Variable	$p(p-1)/2+Gp$
VVIV	Variable	Variable	Isotropic	Variable	$Gp(p-1)/2+G$
VEIV	Variable	Equal	Isotropic	Variable	$Gp(p-1)/2+1$
EVIV	Equal	Variable	Isotropic	Variable	$p(p-1)/2+G$
EEIV	Equal	Equal	Isotropic	Variable	$p(p-1)/2+1$

shape parameter, β, so there is an option of constraining β_g to be equal across all the groups, i.e. $\beta_g = \beta$ (Dang *et al.*, 2015). Combining all of the above constraints there are a total of 16 MPE mixture models. The nomenclature and number of covariance parameters of the 16 models are given in Table 2.1. From this family of mixture models the fully unconstrained model is the VVAV and the fully constrained model is the EEIE.

2.2 Model Fitting

McNicholas and Murphy (2010) and McNicholas and Subedi (2012) used the expectation-maximization (EM) algorithm to fit the models, but the MPE mixture models require the use of the generalized EM (GEM; Dempster *et al.*, 1977) algorithm as

seen in Dang *et al.* (2015). The EM algorithm is an iterative method based on the complete-data likelihood, where at each iteration the expected value of the complete-data log-likelihood is maximized and this results in parameter updates (Dempster *et al.*, 1977). There also is an expectation-conditional-maximization (ECM) algorithm, which is like the EM algorithm, but the maximization (M) step is replaced with several conditional maximization (CM) steps (Meng and Rubin, 1993). For the ECM, the expected complete-data log-likelihood is maximized at each iteration like the EM algorithm, but if it only increases at each iteration instead of maximizing, then it is known as the GEM algorithm.

The likelihood for p-dimensional $\mathbf{x}_1, \ldots, \mathbf{x}_N$ from a MPE mixture model with a modified Cholesky decomposed covariance structure is

$$L_0(\boldsymbol{\Theta}) = \prod_{i=1}^{N} \sum_{g=1}^{G} \pi_g k_g |\mathbf{D}_g|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \left(\mathbf{m}_{ig}' \mathbf{T}_g' \mathbf{D}_g^{-1} \mathbf{T}_g \mathbf{m}_{ig} \right)^{\beta_g} \right\}, \tag{2.5}$$

where $\mathbf{m}_{ig} = \mathbf{x}_i - \boldsymbol{\mu}_g$ and N is the number of observations. Now $\mathbf{z}_i = (z_{i1}, \ldots, z_{iG})$ is the group's label vector such that $z_{ig} = 1$ if $\mathbf{x}_i$ belongs in group g and $z_{ig} = 0$ if it does not belong. The complete-data combines the known data, $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, with the missing data, $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$. The complete-data log-likelihood for the MPE mixture model is

$$\mathcal{L}_c(\boldsymbol{\Theta}) = \sum_{i=1}^{N} \sum_{g=1}^{G} z_{ig} \log \left[\pi_g k_g |\mathbf{D}_g|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \left(\mathbf{m}_{ig}' \mathbf{T}_g' \mathbf{D}_g^{-1} \mathbf{T}_g \mathbf{m}_{ig} \right)^{\beta_g} \right\} \right]. \tag{2.6}$$

The expectation (E) step for the GEM algorithm requires the expected complete-data

log-likelihood to be calculated. The expected complete-data log-likelihood is given by

$$\mathbf{Q}(\boldsymbol{\Theta}) = \sum_{i=1}^{N}\sum_{g=1}^{G} \frac{\hat{z}_{ig}}{2}\left[\log|\mathbf{D}_g|^{-1} - \left(\mathbf{m}_{ig}'\mathbf{T}_g'\mathbf{D}_g^{-1}\mathbf{T}_g\mathbf{m}_{ig}\right)^{\beta_g}\right], \tag{2.7}$$

where the expected values, $\hat{z}_{ig}$, are given by

$$\hat{z}_{ig} = \frac{\hat{\pi}_g f\left(\mathbf{x}_i|\hat{\boldsymbol{\mu}}_g, \hat{\mathbf{D}}_g, \hat{\mathbf{T}}_g, \hat{\beta}_g\right)}{\sum_{h=1}^{G}\hat{\pi}_h f\left(\mathbf{x}_i|\hat{\boldsymbol{\mu}}_h, \hat{\mathbf{D}}_h, \hat{\mathbf{T}}_h, \hat{\beta}_h\right)}, \tag{2.8}$$

for $i = 1, \ldots, N$ and $g = 1, \ldots, G$.

GEM's M-step consists of several CM steps, where the expected value of the complete-data log-likelihood with respect to its parameters, $\boldsymbol{\Theta}$, is maximized. $\hat{\pi}_g$ is the probability of belonging to group g. This is the only update that has a closed form solution, which is

$$\hat{\pi}_g = \frac{n_g}{N}, \tag{2.9}$$

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

The updates for $\hat{\boldsymbol{\mu}}_g$, $\hat{\beta}_g$, $\hat{\mathbf{D}}_g$ and $\hat{\mathbf{T}}_g$ are not possible in closed form, so other methods need to be utilized. For updating $\hat{\boldsymbol{\mu}}_g$ the Newton-Raphson method had to be implemented. The update is

$$\hat{\boldsymbol{\mu}}_g^{\text{new}} = \hat{\boldsymbol{\mu}}_g^{\text{old}} - \frac{\partial Q/\partial\boldsymbol{\mu}_g}{\partial^2 Q/\partial\boldsymbol{\mu}_g\partial\boldsymbol{\mu}_g'}, \tag{2.10}$$

where $\hat{\boldsymbol{\mu}}_g^{\text{old}}$ is the update for $\boldsymbol{\mu}_g$ from the last iteration and $\hat{\boldsymbol{\mu}}_g^{\text{new}}$ is the $\boldsymbol{\mu}_g$ update

from the current iteration. The partial derivatives are defined as

$$\frac{\partial Q}{\partial \boldsymbol{\mu}_g} = \hat{\beta}_g \sum_{i=1}^{N} \hat{z}_{ig} \delta_{ig}(\mathbf{x}_i)^{\hat{\beta}_g - 1} \hat{\boldsymbol{\Sigma}}_g^{-1} \mathbf{m}_{ig}, \tag{2.11}$$

$$\frac{\partial^2 Q}{\partial \boldsymbol{\mu}_g \partial \boldsymbol{\mu}_g'} = \hat{\beta}_g \sum_{i=1}^{N} \hat{z}_{ig} \left[-\delta_{ig}(\mathbf{x}_i)^{\hat{\beta}_g - 1} \hat{\boldsymbol{\Sigma}}_g^{-1} + (\hat{\beta}_g - 1) \delta_{ig}(\mathbf{x}_i)^{\hat{\beta}_g - 2} \hat{\boldsymbol{\Sigma}}_g^{-1} \mathbf{m}_{ig} (-2\hat{\boldsymbol{\Sigma}}_g^{-1} \mathbf{m}_{ig})' \right], \tag{2.12}$$

where

$$\delta_{ig}(\mathbf{x}_i) := (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)' \hat{\boldsymbol{\Sigma}}_g^{-1} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)$$

and $\hat{\boldsymbol{\Sigma}}_g^{-1} = \hat{\mathbf{T}}_g' \hat{\mathbf{D}}_g^{-1} \hat{\mathbf{T}}_g$. The Newton-Raphson method is also used to update $\hat{\beta}_g$, but this update depends on whether $\hat{\beta}_g$ is constrained or unconstrained across all groups. The update is

$$\hat{\beta}_g^{\text{new}} = \hat{\beta}_g^{\text{old}} - \frac{\partial Q / \partial \beta_g}{\partial^2 Q / \partial \beta_g^2}, \tag{2.13}$$

where $\hat{\beta}_g^{\text{old}}$ is the update for β_g from the last iteration and $\hat{\beta}_g^{\text{new}}$ is the β_g update from the current iteration.

If $\hat{\beta}_g$ is unconstrained in the model then use,

$$\frac{\partial Q}{\partial \beta_g} = \frac{p n_g}{2\hat{\beta}_g^2} \psi\left(1 + \frac{p}{2\hat{\beta}_g}\right) + \frac{p n_g \log 2}{2\hat{\beta}_g^2} - \sum_{i=1}^{N} \frac{\hat{z}_{ig}}{2} \delta_{ig}(\mathbf{x}_i)^{\hat{\beta}_g} \log \delta_{ig}(\mathbf{x}_i) \tag{2.14}$$

$$\begin{aligned} \frac{\partial^2 Q}{\partial \beta_g^2} = {} & \frac{-p n_g}{\hat{\beta}_g^3} \psi\left(1 + \frac{p}{2\hat{\beta}_g}\right) - \frac{p^2 n_g}{4\hat{\beta}_g^4} \psi_1\left(1 + \frac{p}{2\hat{\beta}_g}\right) - \frac{p n_g \log 2}{\hat{\beta}_g^3} \\ & - \sum_{i=1}^{N} \frac{\hat{z}_{ig}}{2} \delta_{ig}(\mathbf{x}_i)^{\hat{\beta}_g} \left[\log \delta_{ig}(\mathbf{x}_i)\right]^2, \end{aligned} \tag{2.15}$$

where $\psi(\cdot)$ is the digamma function and $\psi_1(\cdot)$ is the trigamma function. The digamma function is defined as the logarithmic derivative of the gamma function, which is

defined as

$$\psi(z) = \frac{d}{dz} \log \Gamma(z)$$

(Sibuya, 1979). The trigamma function is the derivative of the digamma function, which is defined as

$$\psi_1(z) = \frac{d}{dz}\psi(z) = \frac{d^2}{dz^2} \log \Gamma(z)$$

(Sibuya, 1979). If $\hat{\beta}_g$ is constrained in the model then use,

$$\frac{\partial Q}{\partial \beta} = \frac{pN}{2\hat{\beta}^2}\psi\left(1 + \frac{p}{2\hat{\beta}}\right) + \frac{pN \log 2}{2\hat{\beta}^2} - \sum_{g=1}^{G}\sum_{i=1}^{N} \frac{\hat{z}_{ig}}{2}\delta_{ig}(\mathbf{x}_i)^{\hat{\beta}} \log \delta_{ig}(\mathbf{x}_i) \tag{2.16}$$

$$\begin{aligned}\frac{\partial^2 Q}{\partial \beta^2} = {} & \frac{-pN}{\hat{\beta}^3}\psi\left(1 + \frac{p}{2\hat{\beta}}\right) - \frac{p^2 N}{4\hat{\beta}^4}\psi_1\left(1 + \frac{p}{2\hat{\beta}}\right) - \frac{pN \log 2}{\hat{\beta}^3} \\ & - \sum_{g=1}^{G}\sum_{i=1}^{N} \frac{\hat{z}_{ig}}{2}\delta_{ig}(\mathbf{x}_i)^{\hat{\beta}} \left[\log \delta_{ig}(\mathbf{x}_i)\right]^2 .\end{aligned} \tag{2.17}$$

The updates for $\hat{\mathbf{D}}_g$ and $\hat{\mathbf{T}}_g$ are also impossible to obtain in closed form, so other methods need to be implemented. The update for $\hat{\mathbf{D}}_g$ relies on the convexity properties for maximization, so the minorization-maximization (MM) algorithm is utilized for the M-step. A particular MM algorithm allows for the expected complete-data log-likelihood to increase at each iteration instead of maximizing (Hunter and Lange, 2000). The MM algorithm is constructed by using the convexity of the objective function, which is a surrogate minorizing function (Browne and McNicholas, 2014b). Due to the behaviour of the MPE distribution, the $\hat{\mathbf{D}}_g$ update has to be done in two parts, which depends on the value of $\hat{\beta}_g$. The sub-diagonal elements of $\hat{\mathbf{T}}_g$ are updated by a built-in optimizer in R through the `optim` function. The L-BFGS-B method was used for the `optim` function, where the lower and upper bounds on the

variables was -100 and 100, respectively. The objective function used for the `optim` function is

$$\sum_{i=1}^{N} \frac{\hat{z}_{ig}}{2} \log |\mathbf{D}_g|^{-1} - \frac{\hat{z}_{ig}}{2} \operatorname{tr} \left\{ \left(\mathbf{m}_{ig}' \mathbf{T}_g' \mathbf{D}_g^{-1} \mathbf{T}_g \mathbf{m}_{ig} \right)^{\beta_g} \right\}. \tag{2.18}$$

The pseudo-code for the parameter estimations is:

1. Initialize $\hat{z}_{ig}$, $\hat{\beta}_g$, $\hat{\boldsymbol{\mu}}_g$ and $\hat{\boldsymbol{\Sigma}}_g$.
2. Compute $\hat{\pi}_g$ using (2.9).
3. Update $\hat{\beta}_g$ using (2.14) and (2.17) or (2.16) and (2.17), depending on whether $\hat{\beta}_g$ is unconstrained or constrained.
4. CM step: Update $\hat{\boldsymbol{\mu}}_g$ using (2.11) and (2.12).
5. CM step: Update $\hat{\mathbf{D}}_g$, depending on the scale structure.
6. CM step: Update $\hat{\mathbf{T}}_g$, using (2.18) and the `optim` function in R.
7. Update $\hat{z}_{ig}$ by using (2.8).
8. Update $\hat{\pi}_g$ using (2.9).
9. Check for convergence. If not converged, then go back to step 3.

2.3 $\hat{\mathbf{D}}_g$ update for the VVAV model

As mentioned in Section 2.2, the update for $\hat{\mathbf{D}}_g$ depends on $\hat{\beta}_g$, which results in two different updates for $\hat{\mathbf{D}}_g$. The first update is for when $\hat{\beta}_g \in (0, 1)$ and the second update is for when $\hat{\beta}_g \in [1, \infty)$.

The $\hat{\mathbf{D}}_g$ update for when $\hat{\beta}_g \in (0,1)$:

$$\begin{aligned}
Q(\boldsymbol{\Sigma}_g) &= \sum_{i=1}^{N}\sum_{g=1}^{G} \frac{\hat{z}_{ig}}{2}\log|\boldsymbol{\Sigma}_g|^{-1} - \frac{\hat{z}_{ig}}{2}\left((\mathbf{x}_i-\boldsymbol{\mu}_g)'\boldsymbol{\Sigma}_g^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_g)\right)^{\beta_g} \\
&= \sum_{i=1}^{N}\sum_{g=1}^{G} \frac{\hat{z}_{ig}}{2}\log|\mathbf{D}_g|^{-1} - \frac{\hat{z}_{ig}}{2}\operatorname{tr}\left((\mathbf{x}_i-\boldsymbol{\mu}_g)'\mathbf{T}_g'\mathbf{D}_g^{-1}\mathbf{T}_g(\mathbf{x}_i-\boldsymbol{\mu}_g)\right)^{\beta_g} \\
&= \sum_{i=1}^{N}\sum_{g=1}^{G} \frac{\hat{z}_{ig}}{2}\log|\mathbf{D}_g|^{-1} - \frac{\hat{z}_{ig}}{2}\operatorname{tr}\left(\mathbf{v}_{ig}'\mathbf{D}_g^{-1}\mathbf{v}_{ig}\right)^{\beta_g} \\
&= \sum_{i=1}^{N}\sum_{g=1}^{G} -\frac{\hat{z}_{ig}}{2}\log|\mathbf{D}_g| - \frac{\hat{z}_{ig}}{2}\operatorname{tr}\left(\mathbf{D}_g^{-1}\mathbf{V}_{ig}\right)^{\beta_g} \\
&= \sum_{i=1}^{N}\sum_{g=1}^{G} -\frac{\hat{z}_{ig}}{2}\log|\boldsymbol{\Lambda}_g^{-1}| - \frac{\hat{z}_{ig}}{2}\operatorname{tr}\left(\boldsymbol{\Lambda}_g\mathbf{V}_{ig}\right)^{\beta_g},
\end{aligned}$$

where $\mathbf{v}_{ig} = \mathbf{T}_g(\mathbf{x}_i - \boldsymbol{\mu}_g)$, $\mathbf{V}_{ig} = \mathbf{v}_{ig}\mathbf{v}_{ig}'$ and $\boldsymbol{\Lambda}_g = \mathbf{D}_g^{-1}$

Let x^{β_g} be equal to $\operatorname{tr}\left(\boldsymbol{\Lambda}_g\mathbf{V}_{ig}\right)^{\beta_g}$. By using the supporting hyperplane inequality for a concave function:

$$x^{\beta_g} \leq x_0^{\beta_g} + \beta_g x_0^{\beta_g-1}(x - x_0),$$

$$\text{i.e., } \operatorname{tr}\left(\boldsymbol{\Lambda}_g\mathbf{V}_{ig}\right)^{\beta_g} \leq \operatorname{tr}\left(\boldsymbol{\Lambda}_g^{(0)}\mathbf{V}_{ig}\right)^{\beta_g} + \beta_g \operatorname{tr}\left(\boldsymbol{\Lambda}_g^{(0)}\mathbf{V}_{ig}\right)^{\beta_g-1}\left[\operatorname{tr}\left(\boldsymbol{\Lambda}_g\mathbf{V}_{ig}\right) - \operatorname{tr}\left(\boldsymbol{\Lambda}_g^{(0)}\mathbf{V}_{ig}\right)\right]$$

Constructing a minorizer,

$$-\operatorname{tr}\left(\boldsymbol{\Lambda}_g\mathbf{V}_{ig}\right)^{\beta_g} \geq -\operatorname{tr}\left(\boldsymbol{\Lambda}_g^{(0)}\mathbf{V}_{ig}\right)^{\beta_g} - \beta_g \operatorname{tr}\left(\boldsymbol{\Lambda}_g^{(0)}\mathbf{V}_{ig}\right)^{\beta_g-1}\left[\operatorname{tr}\left(\boldsymbol{\Lambda}_g\mathbf{V}_{ig}\right) - \operatorname{tr}\left(\boldsymbol{\Lambda}_g^{(0)}\mathbf{V}_{ig}\right)\right]$$

$$\begin{aligned}
Q(\boldsymbol{\Sigma}_g) \geq \sum_{i=1}^{N}\sum_{g=1}^{G} &\frac{\hat{z}_{ig}}{2}\log|\boldsymbol{\Lambda}_g| - \frac{\hat{z}_{ig}}{2}\operatorname{tr}\left(\boldsymbol{\Lambda}_g^{(0)}\mathbf{V}_{ig}\right) - \frac{\hat{z}_{ig}}{2}\beta_g \operatorname{tr}\left(\boldsymbol{\Lambda}_g^{(0)}\mathbf{V}_{ig}\right)^{\beta_g-1}\operatorname{tr}\left(\boldsymbol{\Lambda}_g\mathbf{V}_{ig}\right) \\
&+ \frac{\hat{z}_{ig}}{2}\beta_g \operatorname{tr}\left(\boldsymbol{\Lambda}_g^{(0)}\mathbf{V}_{ig}\right)^{\beta_g-1}\operatorname{tr}\left(\boldsymbol{\Lambda}_g^{(0)}\mathbf{V}_{ig}\right)
\end{aligned}$$

Taking the derivative of $Q(\boldsymbol{\Sigma}_g)$ with respect to $\boldsymbol{\Lambda}_g$,

$$\frac{\partial Q}{\partial \boldsymbol{\Lambda}_g} = \sum_{i=1}^{N} \frac{\hat{z}_{ig}}{2} \left(\boldsymbol{\Lambda}_g'\right)^{-1} - \frac{\hat{z}_{ig}}{2} \beta_g \operatorname{tr}\left(\boldsymbol{\Lambda}_g^{(0)} \mathbf{V}_{ig}\right)^{\beta_g - 1} \mathbf{V}_{ig}' \tag{2.19}$$

Setting (2.19) to $\mathbf{0}$ gives,

$$\begin{aligned}
\mathbf{0} &= \frac{\hat{\boldsymbol{\Lambda}}_g^{-1}}{2} \sum_{i=1}^{N} \hat{z}_{ig} - \frac{\beta_g}{2} \sum_{i=1}^{N} \hat{z}_{ig} \operatorname{tr}\left(\boldsymbol{\Lambda}_g^{(0)} \mathbf{V}_{ig}\right)^{\beta_g - 1} \mathbf{V}_{ig} \\
-\frac{\hat{\boldsymbol{\Lambda}}_g^{-1}}{2} n_g &= -\frac{\beta_g}{2} \sum_{i=1}^{N} \hat{z}_{ig} \operatorname{tr}\left(\boldsymbol{\Lambda}_g^{(0)} \mathbf{V}_{ig}\right)^{\beta_g - 1} \mathbf{V}_{ig} \\
\hat{\boldsymbol{\Lambda}}_g^{-1} &= \frac{\beta_g}{n_g} \sum_{i=1}^{N} \hat{z}_{ig} \operatorname{tr}\left(\boldsymbol{\Lambda}_g^{(0)} \mathbf{V}_{ig}\right)^{\beta_g - 1} \mathbf{V}_{ig}
\end{aligned}$$

Replace $\hat{\boldsymbol{\Lambda}}_g^{-1}$ with $\hat{\mathbf{D}}_g$ to get the update for $\hat{\mathbf{D}}_g$,

$$\hat{\mathbf{D}}_g = \frac{\beta_g}{n_g} \sum_{i=1}^{N} \hat{z}_{ig} \operatorname{tr}\left(\mathbf{D}_g^{-1(0)} \mathbf{V}_{ig}\right)^{\beta_g - 1} \mathbf{V}_{ig}$$

The $\hat{\mathbf{D}}_g$ update for when $\hat{\beta}_g \in [1, \infty)$:

Let $\mathbf{D}_g = \boldsymbol{\Lambda}_g^{-1/\beta_g}$, so the modified Cholesky decomposed $\boldsymbol{\Sigma}_g^{-1} = \mathbf{T}_g \boldsymbol{\Lambda}_g^{1/\beta_g} \mathbf{T}_g'$.

From the first part,

$$\begin{aligned}
Q(\boldsymbol{\Sigma}_g) &= \sum_{i=1}^{N} \sum_{g=1}^{G} -\frac{\hat{z}_{ig}}{2} \log |\boldsymbol{\Lambda}_g^{-1/\beta_g}| - \frac{\hat{z}_{ig}}{2} \operatorname{tr}\left((\mathbf{x}_i - \boldsymbol{\mu}_g)' \mathbf{T}_g \boldsymbol{\Lambda}_g^{1/\beta_g} \mathbf{T}_g' (\mathbf{x}_i - \boldsymbol{\mu}_g)\right) \\
&= \sum_{i=1}^{N} \sum_{g=1}^{G} -\frac{\hat{z}_{ig}}{2} \log |\boldsymbol{\Lambda}_g^{-1/\beta_g}| - \frac{\hat{z}_{ig}}{2} \operatorname{tr}\left(\boldsymbol{\Lambda}_g^{1/\beta_g} \mathbf{v}_{ig} \mathbf{v}_{ig}'\right)
\end{aligned}$$

For $i = 1, \ldots, N$,

$$f(\boldsymbol{\lambda}_g) = \operatorname{tr}\left(\mathbf{v}_{ig}' \boldsymbol{\Lambda}_g^{1/\beta_g} \mathbf{v}_{ig}\right)^{\beta_g} = \left(\sum_{k=1}^{p} \lambda_{kg}^{1/\beta_g} v_{kg}^2\right)^{\beta_g},$$

where $\mathbf{\Lambda}_g = \text{diag}(\lambda_{g1}, \ldots, \lambda_{gp})$. This function is concave with respect to the eigenvalues $\boldsymbol{\lambda} = \{\lambda_{g1}, \ldots, \lambda_{gp}\}$.

$$\nabla f(\boldsymbol{\lambda}_g) = \beta_g \left(\sum_{k=1}^{p} \lambda_{kg}^{1/\beta_g} v_{kg}^2 \right)^{\beta_g - 1} \left(\left(\frac{v_{1g}^2}{\beta_g} \lambda_{1g}^{1/\beta_g - 1} \right), \ldots, \left(\frac{v_{kg}^2}{\beta_g} \lambda_{kg}^{1/\beta_g - 1} \right) \right)'$$

A surrogate function is constructed using,

$$\begin{aligned}
f(\boldsymbol{\lambda}_g) &\leq f(\boldsymbol{\lambda}_g^{(0)}) + (\nabla f(\boldsymbol{\lambda}_g^{(0)}))'(\boldsymbol{\lambda} - \boldsymbol{\lambda}^{(0)}) \\
&\leq \text{tr}\left(\mathbf{\Lambda}_g^{(0)1/\beta_g} \mathbf{V}_{ig} \right)^{\beta_g} + \text{tr}\left(\mathbf{v}_{ig}' \mathbf{\Lambda}_g^{(0)1/2\beta_g} \mathbf{\Lambda}_g^{(0)1/2\beta_g} \mathbf{v}_{ig} \right)^{\beta_g - 1} \\
&\quad \times \left(\left(v_{1g}^2 \lambda_{1g}^{(0)1/\beta_g - 1} \right), \ldots, \left(v_{kg}^2 \lambda_{kg}^{(0)1/\beta_g - 1} \right) \right)' \times \left(\left(\lambda_{1g} - \lambda_{1g}^{(0)} \right), \ldots, \left(\lambda_{kg} - \lambda_{kg}^{(0)} \right) \right)' \\
&\leq \text{tr}\left(\mathbf{\Lambda}_g^{(0)1/\beta_g} \mathbf{V}_{ig} \right)^{\beta_g} + \text{tr}\left(\mathbf{v}_{ig}' \mathbf{\Lambda}_g^{(0)1/2\beta_g} \mathbf{\Lambda}_g^{(0)1/2\beta_g} \mathbf{v}_{ig} \right)^{\beta_g - 1} \\
&\quad \times \left(\sum_{k=1}^{p} v_{kg}^2 \lambda_{kg} \lambda_{kg}^{(0)1/\beta_g - 1} - \sum_{k=1}^{p} v_{kg}^2 \lambda_{kg}^{(0)1/\beta_g} \right) \\
&\leq \text{tr}\left(\mathbf{\Lambda}_g^{(0)1/\beta_g} \mathbf{V}_{ig} \right)^{\beta_g} + \text{tr}\left(\mathbf{v}_{ig}' \mathbf{\Lambda}_g^{(0)1/2\beta_g} \mathbf{\Lambda}_g^{(0)1/2\beta_g} \mathbf{v}_{ig} \right)^{\beta_g - 1} \\
&\quad \times \text{tr}\left(\mathbf{v}_{ig}' \mathbf{\Lambda}_g^{(0)1/2\beta_g - 1/2} \mathbf{\Lambda}_g \mathbf{\Lambda}_g^{(0)1/2\beta_g - 1/2} \mathbf{v}_{ig} - \mathbf{v}_{ig}' \mathbf{\Lambda}_g^{(0)1/2\beta_g} \mathbf{\Lambda}_g^{(0)1/2\beta_g} \mathbf{v}_{ig} \right) \\
&\leq \text{tr}\left(\mathbf{\Lambda}_g^{(0)1/\beta_g} \mathbf{V}_{ig} \right)^{\beta_g} + \text{tr}\left(\mathbf{w}_{ig}' \mathbf{w}_{ig} \right)^{\beta_g - 1} \left(\text{tr}\left(\mathbf{w}_{ig}' \mathbf{\Lambda}_g^{(0)-1/2} \mathbf{\Lambda}_g \mathbf{\Lambda}^{(0)-1/2} \mathbf{w}_{ig} \right) - \text{tr}\left(\mathbf{w}_{ig}' \mathbf{w}_{ig} \right) \right) \\
&\leq \text{tr}\left(\mathbf{\Lambda}_g^{(0)1/\beta_g} \mathbf{V}_{ig} \right)^{\beta_g} + \text{tr}\left(\mathbf{\Lambda}_g \mathbf{\Lambda}_g^{(0)-1/2} \mathbf{W}_{ig}^{\beta_g} \mathbf{\Lambda}_g^{(0)-1/2} \right) - \text{tr}\left(\mathbf{W}_{ig}^{\beta_g} \right),
\end{aligned}$$

where $\mathbf{w}_{ig} = \mathbf{v}_{ig}' \mathbf{\Lambda}_g^{(0)1/2\beta_g}$ and $\mathbf{W}_{ig}^{\beta_g} = \mathbf{w}_{ig} \mathbf{w}_{ig}' (\mathbf{w}_{ig}' \mathbf{w}_{ig})^{\beta_g - 1}$.
Let $c = \text{tr}\left(\mathbf{\Lambda}_g^{(0)1/\beta_g} \mathbf{V}_{ig} \right)^{\beta_g} - \text{tr}\left(\mathbf{W}_{ig}^{\beta_g} \right)$, which are constants. Then, Q is maximized,

$$Q(\mathbf{\Sigma}_g) = \sum_{i=1}^{N} \sum_{g=1}^{G} -\frac{\hat{z}_{ig}}{2\beta_g} \log |\mathbf{\Lambda}_g| + \frac{\hat{z}_{ig}}{2} \left(c + \text{tr}\left(\mathbf{\Lambda}_g \mathbf{\Lambda}_g^{(0)-1/2} \mathbf{W}_{ig}^{\beta_g} \mathbf{\Lambda}_g^{(0)-1/2} \right) \right)$$

$$\frac{\partial Q}{\partial \mathbf{\Lambda}_g} = \sum_{i=1}^{N} -\frac{\hat{z}_{ig}}{2\beta_g} \mathbf{\Lambda}_g^{-1} + \sum_{i=1}^{N} \frac{\hat{z}_{ig}}{2} \mathbf{\Lambda}_g^{(0)-1/2} \mathbf{W}_{ig}^{\beta_g} \mathbf{\Lambda}_g^{(0)-1/2}$$

$$\frac{\hat{\boldsymbol{\Lambda}}_g^{-1}}{2\beta_g}\sum_{i=1}^{N}\hat{z}_{ig} = \frac{1}{2}\sum_{i=1}^{N}\hat{z}_{ig}\boldsymbol{\Lambda}_g^{(0)-1/2}\mathbf{W}_{ig}^{\beta_g}\boldsymbol{\Lambda}_g^{(0)-1/2}$$

$$\hat{\boldsymbol{\Lambda}}_g^{-1} = \frac{\beta_g}{n_g}\sum_{i=1}^{N}\hat{z}_{ig}\boldsymbol{\Lambda}_g^{(0)-1/2}\mathbf{W}_{ig}^{\beta_g}\boldsymbol{\Lambda}_g^{(0)-1/2}$$

$$\hat{\mathbf{D}}_g^{\beta_g} = \frac{\beta_g}{n_g}\sum_{i=1}^{N}\hat{z}_{ig}\mathbf{D}_g^{(0)\beta_g/2}\mathbf{W}_{ig}^{\beta_g}\mathbf{D}_g^{(0)\beta_g/2}$$

Therefore, the update for $\hat{\mathbf{D}}_g$ is

$$\hat{\mathbf{D}}_g = \left(\frac{\beta_g}{n_g}\sum_{i=1}^{N}\hat{z}_{ig}\mathbf{D}_g^{(0)\beta_g/2}\mathbf{W}_{ig}^{\beta_g}\mathbf{D}_g^{(0)\beta_g/2}\right)^{1/\beta_g}.$$

2.4 Convergence Criterion

The Aitken acceleration is utilized as the stopping criterion to determine the convergence of the EM algorithm, by providing an asymptotic estimate of the log-likelihood at each iteration. The Aitken acceleration at iteration m is given by

$$a^{(m)} = \frac{l^{(m+1)} - l^{(m)}}{l^{(m)} - l^{(m-1)}}, \tag{2.20}$$

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

$$l_\infty^{(m+1)} = l^{(m)} + \frac{1}{1-a^{(m)}}\left(l^{(m+1)} - l^{(m)}\right) \tag{2.21}$$

(Böhning *et al.*, 1994). The EM algorithm has converged if

$$l_\infty^{(m+1)} - l^{(m)} < \epsilon$$

(McNicholas *et al.*, 2010), where ϵ was taken to be 0.005 in the analysis herein.

2.5 Model Selection

Model selection is crucial for model-based clustering because multiple models are fit for a range of G values. The Bayesian information criterion (BIC; Schwarz, 1978) is used to select the best model. The BIC is written as

$$\text{BIC} = 2l(\mathbf{x}, \hat{\boldsymbol{\theta}}) - \rho \log N, \tag{2.22}$$

where $l(\mathbf{x}, \hat{\boldsymbol{\theta}})$ is the maximized log-likelihood, $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$, ρ is the number of free parameters in the model and N is the number of observations.

2.6 Performance Assessment

For the purposes of this paper, cluster analysis was only done on datasets that had known clusters, so the performance can be assessed and compared to other methods. The adjusted Rand Index (ARI; Rand, 1971; Hubert and Arabie, 1985) is used to assess the clustering performance, by comparing the true cluster memberships to the estimated ones given by the best chosen model. The expected value of the ARI under random classification is zero and the value of the ARI under perfect agreement is one.

Chapter 3

Illustrations

3.1 Growth Curves Data

This dataset consists of 20 pre-adolescent girls whose heights were measured yearly from ages six to ten. The girls were categorized based on their mother's height, where the categories were: short, medium and tall. There were six girls in the short group, seven girls in the medium group and seven girls in the tall group. This dataset was published in Verbeke and Lesaffre (1997). This paper did not mention what the exact height ranges were for the categories of the mother's height. The measurements taken of the girls' heights over time can be seen on the time series plot in Figure 3.1. From this plot, it seems that there is a relationship between the mother's height and the daughter's height, but the groups are close together overlapping. For the purpose of performing cluster analysis, the mothers' heights are treated as the true memberships for the girls, which are taken to be unknown *a priori*.

The family of the MPE mixtures models from Table 2.1 are fit to this data, which results in 16 different models. Each model is fit with 10 different random starting

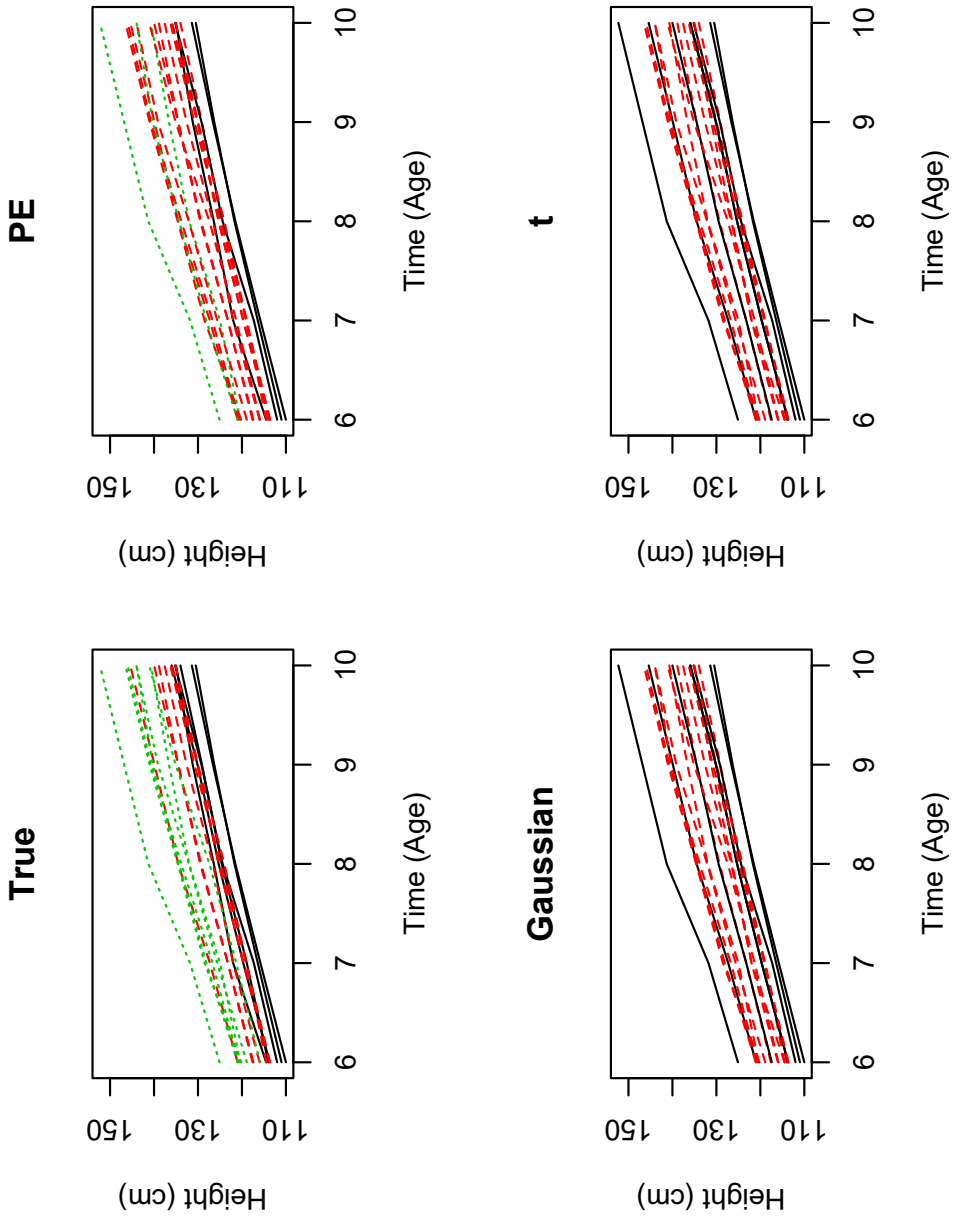

Figure 3.1: These time series plots show each girls' height between age 6 and 10. The top left plot shows the true group memberships which is grouped by mothers' height, where short is the black solid lines, medium is red dashed lines and tall is green dotted lines. The top right plot shows the estimated group memberships obtained from the PE mixture models. The bottom left plot shows the estimated memberships obtained from the Gaussian mixture models and the bottom right shows the estimated memberships from the t mixture models.

Table 3.1: True clusters cross-tabulated against estimated classifications for the growth curves data, using PE, Gaussian, and t-mixtures, respectively.

	PE			Gaussian		t	
	1	2	3	1	2	1	2
short	4	2	0	4	2	4	2
medium	0	7	0	3	4	3	4
tall	0	4	3	1	6	1	6

values for $\hat{z}_{ig}$, for $G = 1, \ldots, 5$. The best model chosen based on BIC is the EEAV model with $G = 3$. The number of clusters selected are the same as the true number of clusters, but there are some miss-classifications, which can be seen in Table 3.1. From this table it can be seen that the seven girls with a medium mother were all classified correctly, which tells us that medium height mothers have medium height daughters. Four girls with a short mother were correctly classified, but two girls from this group were incorrectly classified into the medium mothers group. This suggests that short height mothers usually have short height daughters, but there is a chance of them having a medium height daughter. Three girls with a tall mother were correctly classified, but four girls from this group were incorrectly classified into the medium mothers group. This suggests that tall height mothers can have tall or medium height daughters. The ARI associated to Table 3.1 is 0.24, which indicates that the classification between the true and estimated classifications is better than what would be expected under random classification. The estimated classifications can be seen on the times series plot in Figure 3.1. The EEAV model suggests that the autoregressive structure is the same for all three groups and the fact that the isotropic constraint was not imposed, suggests that the innovation variances differ among time points for a particular group. This model also suggests that the shape parameter is different for all three groups. $\hat{\boldsymbol{\beta}} = (2.01, 2.31, 2.17)'$, which suggests

a lighter-tailed distribution than the Gaussian, so the data might not be correctly modeled by a multivariate Gaussian distribution.

For comparison purposes the family of Gaussian mixture models and t mixture models, which were mentioned in Chapter 1, are fit to this data. Both families consist of eight models each, which are shown in Table 1.1. Each model is fit with 10 different random starting values for $\hat{z}_{ig}$ for $G = 1, \ldots, 5$. Longitudinal clustering via multivariate Gaussian and multivariate t-distribution was performed using the function `longclustEM` from the package `longclust` (McNicholas *et al.*, 2015) in R. The best model selected from both the Gaussian and t mixture models was the EVA model with $G = 2$. The cross-tabulation between the true and estimated classifications for both families can be seen in Table 3.1. Both families have the exact same estimated classifications, which seems reasonable because the estimated degrees of freedom $\hat{\boldsymbol{\nu}} = (86.29, 86.15)$ for t mixture models, which reflect the fact that both clusters are Gaussian. The ARI associated with Table 3.1 for the Gaussian and t mixture models is 0.05, which suggests the agreement between the true and estimated classifications are no better than random. The estimated classifications for both mixture models can be seen on the time series plot in Figure 3.1.

3.2 Growth of Sitka Spruce Trees Data

This dataset consists of 79 Sitka spruce trees where the growth was monitored for 674 days. This dataset was sourced from the `nlme` package (Pinheiro *et al.*, 2016) in R and was originally published in Diggle *et al.* (1994). The study objective that Diggle *et al.* (1994) had was to assess the effect of ozone pollution on the tree growth. This was done by measuring trees in plots with ozone exposure at 70 ppb or were under

control conditions. There were two ozone exposure plots with 27 trees in each and two control plots with 12 in one and 13 in the other. In the dataset these plots are referred to as plots 1,2,3, and 4, respectively. The tree's size was measured by taking the product of the tree's height and diameter squared. Diggle *et al.* (1994) used the logarithm of size for the analysis. The size of the trees was measured on 152, 174, 201, 227, 258, 469, 496, 528, 556, 579, 613, 639, and 674 days after the experiment began, which results in 13 time points. Due to the lack of time, only five time points were used for this analysis because this would shorten the computation time. The five time points used were, 152, 227, 496, 579, and 674 days after the experiment began. These measurements can be seen on the time series plot in Figure 3.2. From this plot it can be seen that the four groups are overlapped and might be a hard cluster analysis problem. For the purpose of performing cluster analysis, the four plots of land are treated as the true memberships for the trees, which are taken to be unknown *a priori*.

All 16 models from the family of the MPE mixture models in Table 2.1 are fit. Each model is fit with 10 different random starting values of $\hat{z}_{ig}$ for $G = 1, \ldots, 6$. The best model selected based on BIC is the EVIE model with $G = 3$. The number of clusters selected did not match up with the true number of clusters. The details of the differences between plots were not given other than if they were ozone exposed or not, so there might be extraneous variables in play that may show there is no distinctive differences between all four plots of land. The cross-tabulation between the true and estimated group memberships can be seen in Table 3.2. From this table it can be seen that most of the trees from plot 1 and 2 were clustered together as group 1, but this group also includes eight trees from plot 3. The estimated groups 2 and 3 consist

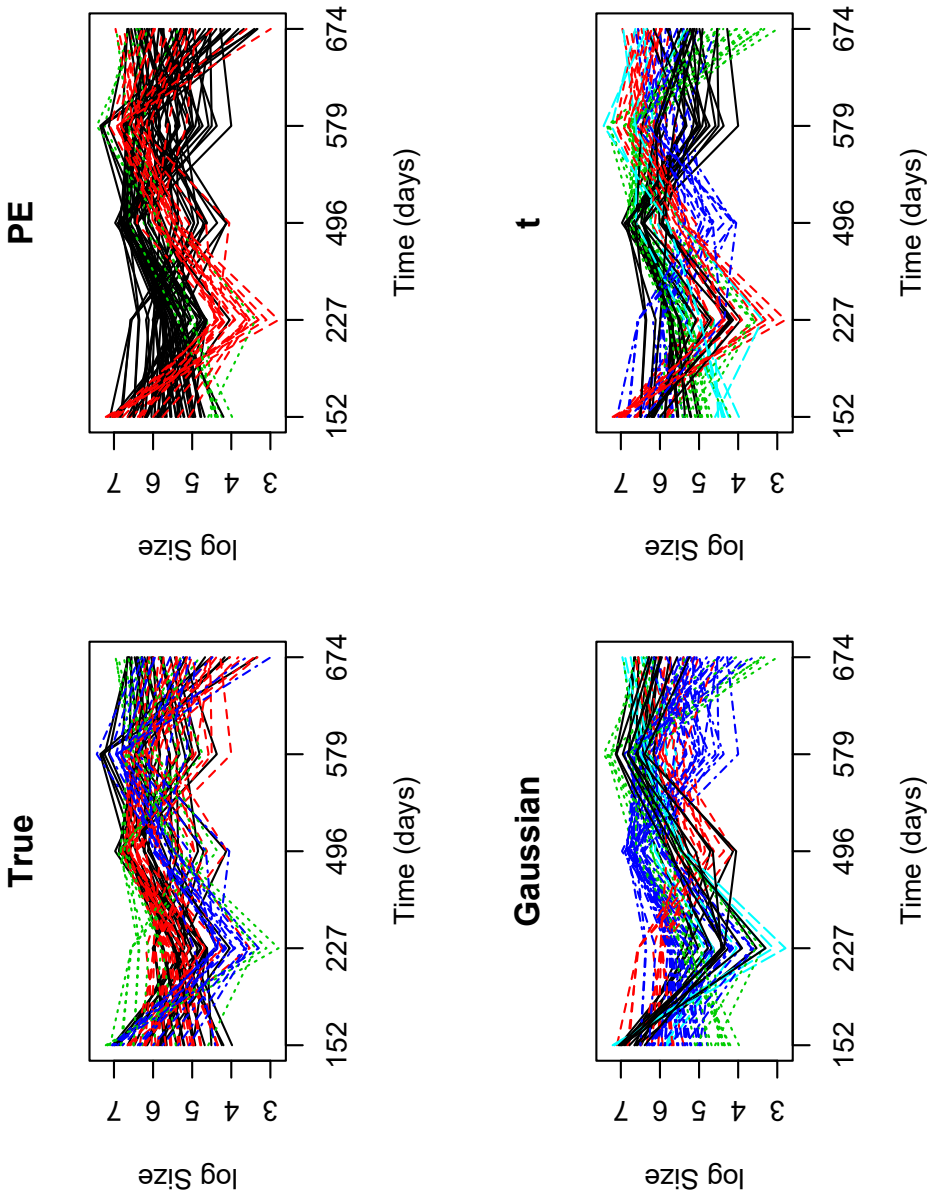

Figure 3.2: These time series plots show each trees' growth between 152 and 674 days. The top left plot shows the true group memberships which is grouped by land plots, where the black solid lines is the ozone exposed plot 1, red dashed lines is ozone exposed plot 2, green dotted lines is control conditions plot 1, and blue dashed/dots line is the control conditions plot 2. The top right plot shows the estimated group memberships obtained from the PE mixture models. The bottom left plot shows the estimated memberships obtained from the Gaussian mixture models and the bottom right shows the estimated memberships from the t mixture models.

Table 3.2: True clusters cross-tabulated against estimated classifications for the Sitka spruce tree data, using PE, Gaussian, and t-mixtures, respectively.

	PE			Gaussian					t				
	1	2	3	1	2	3	4	5	1	2	3	4	5
plot 1	23	1	3	6	3	6	9	3	6	7	6	5	3
plot 2	22	3	2	3	6	5	13	0	10	3	7	5	2
plot 3	8	3	1	0	3	1	5	3	3	3	2	3	1
plot 4	0	12	1	6	0	2	5	0	3	4	3	2	1

of trees from all four plots. The ARI associated to Table 3.2 is 0.21, which indicates that the classification between the true and estimated classifications is better than what would be expected under random classification. The estimated classifications can be seen on a time series plot in Figure 3.2. The EVIE model suggests that the autoregressive structure is the same for all three groups and the innovation variances are different for all three groups; however, because the isotropic constraint has been imposed the innovation variances are the same across all time points for a particular group. This model also suggests the shape parameter is the same for all three groups, i.e., $\hat{\beta}_g = \hat{\beta} = 1.42$, where the magnitude suggests it may be similarly modeled by a multivariate Gaussian distribution.

For comparison purposes the family of Gaussian mixture models and t mixture models are also fit to this data, as in Section 3.1. Each model is fit with 10 different random starting values of $\hat{z}_{ig}$ for $G = 1, \ldots, 6$. The best model selected from the Gaussian mixture models is the EVA model with $G = 5$ and from the t mixture models is the EEA model with $G = 5$. The cross-tabulations between the true and estimated classifications for both can be seen in Table 3.2. Based on this table the classification seems random because most of the estimated clusters consist of trees from each plot. The ARI values associated to this table for the Gaussian and t

mixture models are 0.01 and -0.02, respectively. The ARI values suggest that the classification is random. The estimated classifications for both Gaussian and t mixture models can be seen on the time series plots in Figure 3.2. The degrees of freedom $\hat{\boldsymbol{\nu}} = (96.91, 82.30, 80.13, 108.11, 67.63)$ for the t mixture models, which reflect the fact that the clusters are Gaussian. However, both families of mixture models did not select the same model and the estimated classifications were not exactly the same, as in Section 3.1.

3.3 Weight Loss Data

This dataset consists of 34 individuals where weight and self-esteem were monitored over three months. This dataset was sourced from the `car` package (Fox *et al.*, 2007) in R. Each individual was placed in one of three groups, control, diet, and diet+exercise. There are 12 individuals in the control group, 12 in the diet group and 10 in the diet+exercise group. Each individual's amount of weight lost in pounds was restored monthly for three months. Along with weight lost, the individual self-ranked their self esteem each month. For the purpose of this analysis only the individual's weight loss measurements and the group they belong to are considered. These measurements can be seen on the time series plot in Figure 3.3. From the plot it can be seen that there does not seem to be a distinctive difference between the control and diet group. This dataset has known group memberships *a priori*, but for the purpose of performing cluster analysis the group memberships are taken to be unknown.

All 16 models from the family of the MPE mixture models in Table 2.1 are fit. Each model is fit with 10 random starting values of $\hat{z}_{ig}$ for $G = 1, \ldots, 5$. The best model selected is the EVIE model with $G = 2$. The number of clusters selected did

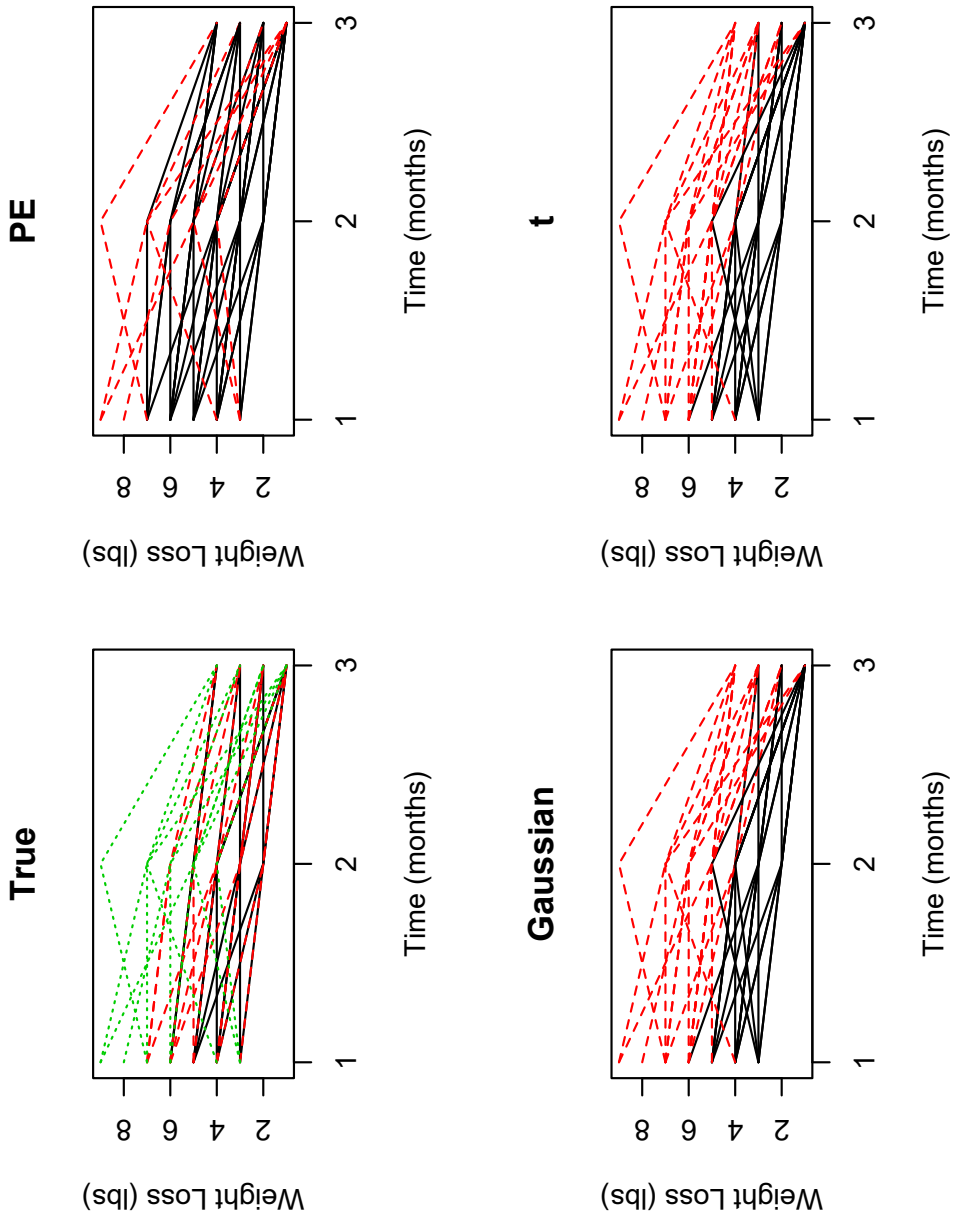

Figure 3.3: These time series plots show each individual's weight loss over 3 months. The top left plot shows the true group memberships, where the black solid lines is the control group, red dashed lines is the diet group, and the green dotted line is the diet+exercise group. The top right plot shows the estimated group memberships obtained from the PE mixture models. The bottom left plot shows the estimated memberships obtained from the Gaussian mixture models and the bottom right shows the estimated memberships from the t mixture models.

Table 3.3: True clusters cross-tabulated against estimated classifications for the weight loss data, using PE, Gaussian, and t-mixtures, respectively.

	PE		Gaussian		t	
	1	2	1	2	1	2
control	12	0	10	2	10	2
diet	12	0	6	6	6	6
diet+exercise	3	7	2	8	2	8

not match up with the true number of clusters, but from the plot in Figure 3.3 it was seen that groups 1 and 2 seem to follow a similar pattern over the three months. The cross-tabulation between the true and estimated group memberships can be seen in Table 3.3. The ARI associated with Table 3.3 is 0.25, which indicates that the classification between the true and estimated classifications is better than what would be expected under random classification. This table shows that the control and diet groups were combined into one group plus three individuals from the diet+exercise group. The other seven individuals were correctly classified into their diet+exercise group. The estimated classifications can be seen on a time series plot in Figure 3.3. The EVIE model suggests that the autoregressive structure is the same for both groups and the innovation variances are different for both groups, but because the isotropic constraint has been imposed the innovation variances are the same at each time point for a particular group. This model also suggests the shape parameter is the same for both groups, i.e. $\hat{\beta}_g = \hat{\beta} = 6.23$ and the magnitude of $\hat{\beta}$ suggests that the data cannot be correctly modeled by a multivariate Gaussian distribution.

For comparison purposes the family of Gaussian mixture models and t mixture models is fit to the weight loss data as well. Each of the models is fit with 10 random starting values of $\hat{z}_{ig}$, for $G = 1, \ldots, 5$. The best model chosen from the Gaussian and t mixture models is the EVI with $G = 2$. The cross-tabulation between the true

and estimated classifications for both families can be seen in Table 3.3. Both families have the exact same estimated classifications, which seem reasonable because the estimated degrees of freedom $\hat{\boldsymbol{\nu}} = (61.85, 53.88)$ for t mixture models, which reflect the fact that both clusters are Gaussian. The ARI associated with Table 3.3 for the Gaussian and t mixture models is 0.13, which suggests the agreement between the true and estimated classifications are close to being random. The estimated classifications for both mixture models can be seen on a time series plot in Figure 3.3. All three family of mixture models selects the covariance structure to be EVI with $G = 2$, but the MPE mixture model performs better.

3.4 Constraining Sub-diagonals of $\hat{\mathbf{T}}_\mathbf{g}$

3.4.1 Introduction

McNicholas and Murphy (2010) introduced the concept of constraining the sub-diagonals of the $\hat{\mathbf{T}}_g$ matrix. This can be utilized when the $\hat{\mathbf{T}}_g$ matrix consists of elements that are very small below a certain sub-diagonal, which may suggest autocorrelation over large time lags exists and can be removed by constraining certain sub-diagonals to be zero. Constraining $\hat{\mathbf{T}}_g$ to have zeros below the dth sub-diagonal implies an order d autoregressive structure within the framework of Equation (1.3). Let

$$\hat{\mathbf{T}}_g = \begin{bmatrix} 1.00 & 0 & 0 & 0 \\ -0.61 & 1.00 & 0 & 0 \\ -0.05 & 0.75 & 1.00 & 0 \\ 0.01 & -0.02 & -0.58 & 1.00 \end{bmatrix}.$$

For this example, it can be seen that below the first sub-diagonal the elements are very small, which indicates that some of the autoregressive parameters may not be needed and replacing them with a zero would lead to a more parsimonious class of models.

Consider the model VVAV and constrain the elements below the dth sub-diagonal to be zero, then the notation used for this is V_dVAV, where $d = 1, \ldots, p-1$. Note that, when $d = p-1$ this is equivalent to the full $\hat{\mathbf{T}}_g$ matrix.

3.4.2 Parameter Estimates

The sub-diagonal elements of $\hat{\mathbf{T}}_g$ are updated by a built-in optimizer in R through the `optim` function. The objective function used for the `optim` function is

$$\sum_{i=1}^{N} \frac{\hat{z}_{ig}}{2} \log |\mathbf{D}_g|^{-1} - \frac{\hat{z}_{ig}}{2} \operatorname{tr} \left\{ \left(\mathbf{m}'_{ig} \mathbf{T}'_g \mathbf{D}_g^{-1} \mathbf{T}_g \mathbf{m}_{ig} \right)^{\beta_g} \right\}, \tag{3.1}$$

where $\mathbf{T}_g$ is constrained to have zeros below the dth sub-diagonal.

3.4.3 Application to Growth Curves Data

The Growth Curves data is used for cluster analysis in Section 3.1 and is also an example where a $\hat{\mathbf{T}}_g$ with constraints on the sub-diagonal would be chosen over the full $\hat{\mathbf{T}}_g$. In Section 3.1 the MPE mixture model chose EEAV with $G = 3$ to be the best model. This dataset consists of five time points, so the E_dEAV model is fit for $d = 1, \ldots, 4$, where E_4EAV is the full model.

Table 3.4 shows the E_dEAV models fit with its respective BIC values and ΔBIC values. The difference between the dth sub-diagonal model's BIC and the full model's

BIC is refereed to as ΔBIC. The best model chosen based on BIC is E_2EAV model, meaning all the elements below the second sub-diagonal are zero. The estimated group memberships were similar to the full EEAV model and can be seen on the time series plot in Figure 3.4. E_2EAV model is the only one that has a positive ΔBIC value, where ΔBIC is approximately two. There are no concrete guidelines for measuring the strength of evidence against the model with the lower BIC value. However, Kass and Raftery (1995) give guidelines for measuring the strength of evidence and is summarized in Table 3.5. According to Table 3.5, E_2EAV model has positive evidence against the full model. The E_2EAV model does not give very strong evidence against the full model, but still gives similar estimated group memberships as the full model and provides with a more parsimonious model. Therefore, the E_2EAV model would be a better option than the full model.

Table 3.4: BIC values for the E_dEAV models fitted to the Growth Curves data.

Model	BIC	ΔBIC
E_1EAV	-380.68	-5.69
E_2EAV	-373.05	1.94
E_3EAV	-378.46	-3.47
E_4EAV	-374.99	0

Table 3.5: Guidelines for the strength of evidence against the model with lower BIC value.

ΔBIC	Evidence against lower BIC
0 to 2	Not worth more than a bare mention
2 to 6	Positive
6 to 10	Strong
>10	Very strong

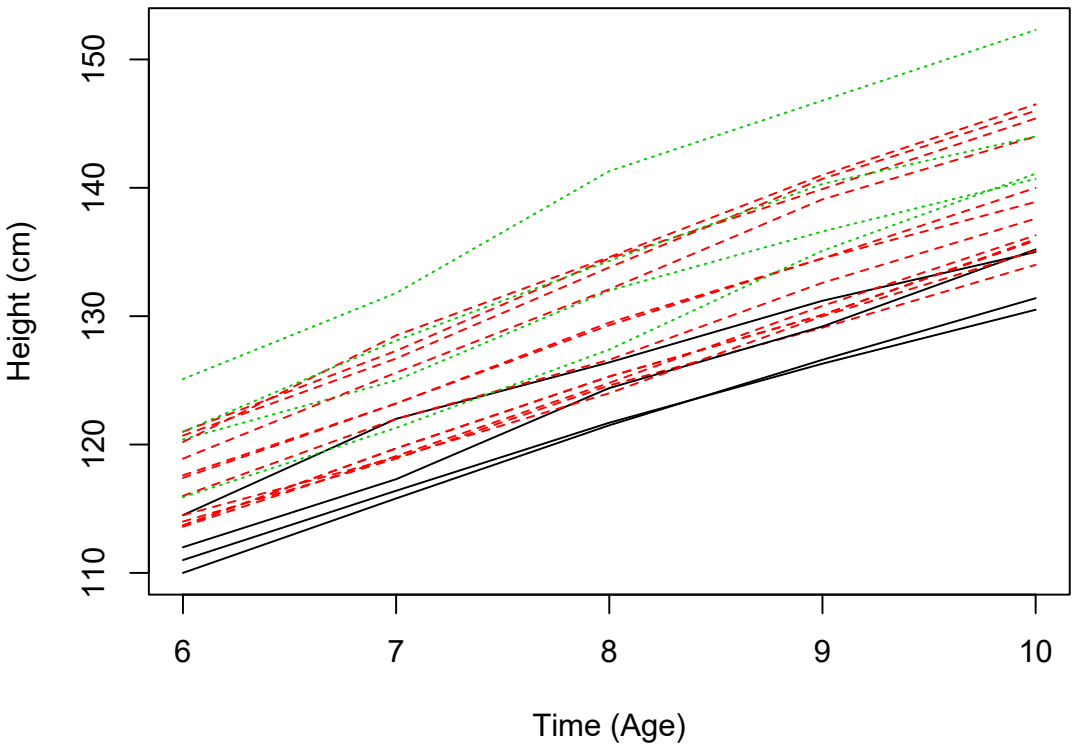

Figure 3.4: Estimated classifications for the chosen E_2EAV model from the family of MPE mixture models. Refer to Figure 3.1 for the time series plot with the true classifications.

Chapter 4

Model-based Classifications

4.1 Introduction

Model-based classification is when some of the group memberships are known in a semi-supervised fashion *a priori* and they are used to predict the other observations' group memberships. The parameters that result from the combination of observations with known and unknown group memberships give the classifications for the observations with unknown group memberships, which is also known as partial classification (Andrews *et al.*, 2011). For N observations, let k observations have known group memberships and $N-k$ observations have unknown group memberships. Let $\mathbf{x}_i$ be an observation, where its group membership is known for $i = 1, \ldots, k$ and unknown for $i = k+1, \ldots, N$. Following McNicholas (2010), order the data such that the first k points are labeled without the loss of generality. Then, the likelihood is

$$L(\boldsymbol{\Theta}|\mathbf{x}) = \prod_{i=1}^{k}\prod_{g=1}^{G}\left[\pi_g f\left(\mathbf{x}_i|\boldsymbol{\mu}_g, \mathbf{D}_g, \mathbf{T}_g, \beta_g\right)\right]^{z_{ig}} \prod_{j=k+1}^{N}\sum_{h=1}^{H}\pi_g f\left(\mathbf{x}_j|\boldsymbol{\mu}_h, \mathbf{D}_h, \mathbf{T}_h, \beta_h\right), \quad (4.1)$$

where $H \geq G$ and $f(\cdot)$ is Equation (2.4).

4.2 Application to Schizophrenia Data

This dataset consists of schizophrenic patients who are a part of a clinical trail to test for the effectiveness of certain drugs. The original dataset presented in Hedeker and Gibbons (1997) had three drugs being tested along with a placebo group, but it was found that all three drug groups responded similarly, so for the purposes of their analysis all three drug groups were combined. This led to only two groups, where patients were either taking a drug or were in the placebo group. These patients were observed for six weeks and measurements of the severity of their illness were recorded on weeks 0, 1, 3, and 6. The severity of the illness was measured on a one to seven scale, where one represents normal and seven is the most severe case of the illness. A major problem for clinical studies is that patients either drop out part way through or do not show up for a certain check up, leading to missing data. Hedeker and Gibbons (1997) used this dataset with the missing data, but Weiss (2005) used the dataset to compare the severity of the illness between patients who showed up for all the checkups and did not drop out versus patients who had missing data. For the purposes of this analysis, only patients with no missing data were included. There were 387 patients, where 294 are in the drug group and 93 are in the placebo group. The measurements taken on the patients' can be seen on the time series plot in Figure 4.1. For the purposes of performing classification, every third patient's group membership was taken to be unknown, which leads to approximately 33% of unknown group memberships.

The family of the MPE mixture models from Table 2.1 are fit to this data and

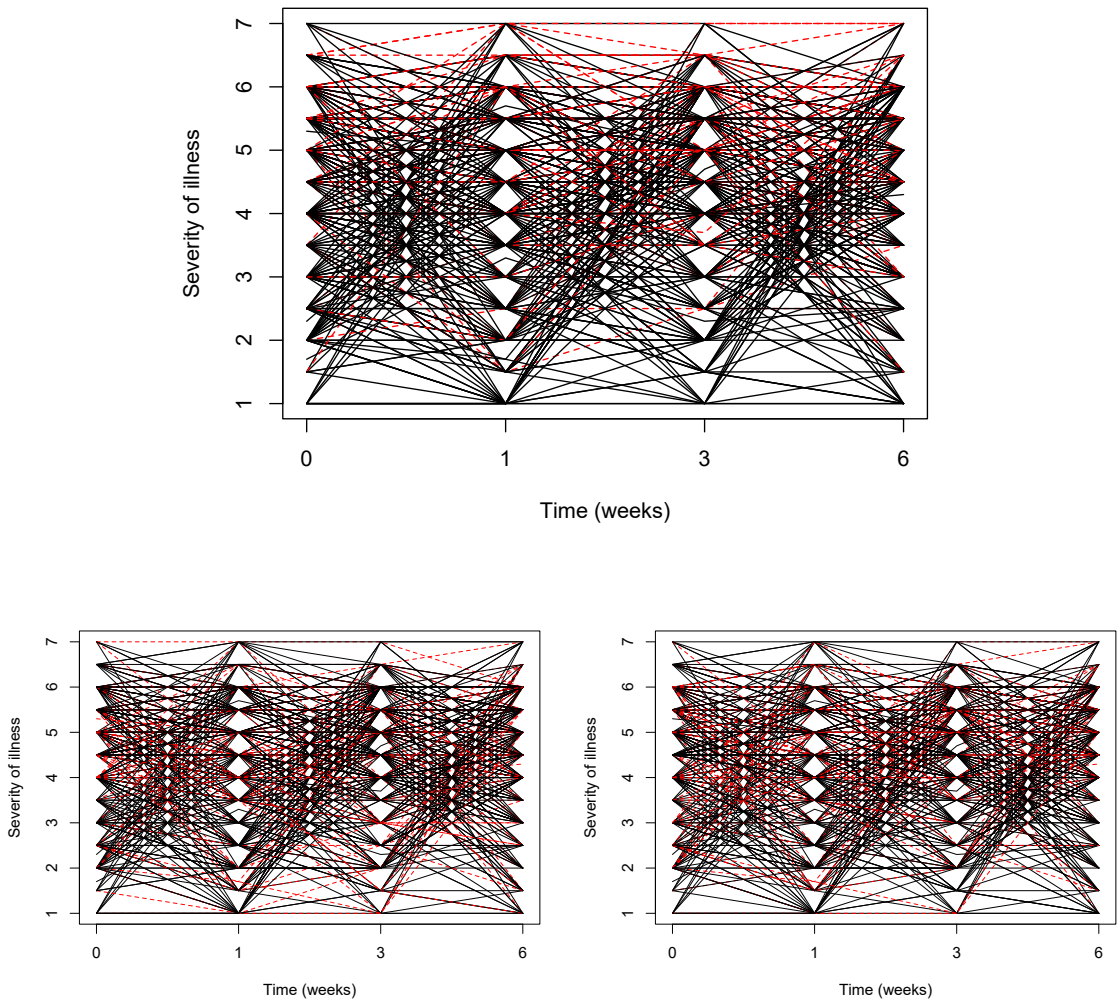

Figure 4.1: These time series plots show each patients' severity of illness over 6 weeks. The top time series plot shows the true memberships for patient, which are grouped by either drug or placebo group, where the drug group is the black solid lines and the placebo group is the red dashed lines. The bottom left time series plot shows the estimated group memberships while performing classification and the bottom right plot shows the estimated group memberships while performing clustering.

each model is fit with 10 different random starting values for $\hat{z}_{ig}$, for $G = 1, \ldots, 5$. The best model chosen based on BIC is the EVIE model with $G = 2$. The right number of clusters is chosen, but with some misclassifications. The ARI for the cross-tabulations between true and estimated classifications is 0.70, which means it performed pretty well but not with a perfect agreement. The EVIE model suggests that the autoregressive structure is the same for both groups and the innovation variances are different for both groups; however, because the isotropic constraint has been imposed the innovation variances are the same at each time point for a particular group. This model also suggests the shape parameter is the same for both groups.

Classification can perform better than clustering because some of the labels are known *a priori*. To show this is true, clustering is done by fitting all 16 models from Table 2.1 for 10 random starting values for $\hat{z}_{ig}$, for $G = 1, \ldots, 5$. The best model selected based was the EVIE with $G = 2$. Surprisingly, both classification and clustering selected the same G and the same model. However, the ARI for estimated clustering group memberships is 0.13, which suggests the agreement was only a bit more than random. This shows classification performed much than clustering, even though the same model and G was selected.

Chapter 5

Conclusion

Model-based clustering for longitudinal data using a Cholesky decomposed covariance structure has been previously explored with the use of Gaussian mixture models (McNicholas and Murphy, 2010) and t mixtures models (McNicholas and Subedi, 2012). This paper introduces a family of MPE mixture models for longitudinal clustering, which was introduced in Dang *et al.* (2015) for cluster analysis. The MPE distribution was chosen for its flexibility in dealing with different tail weights and peakedness. The modified Cholesky decomposition was utilized for the covariance structure because this decomposition accounts for the relationship between measurements at different time points. Various constraints can be placed on the modified Cholesky decomposed covariance structure and constraints can also be placed on the shape parameter, which led to a family of 16 mixture models. Real datasets were used to show that the family of MPE mixture models is a good alternative for non-Gaussian data, where the families of mixtures of Gaussian and mixtures of t do not perform well. Constraining the sub-diagonals of $\hat{\mathbf{T}}_g$ was further looked at and through the use of a real dataset was shown that constraining certain sub-diagonals to zero may lead to a better model

with a more parsimonious covariance structure. Model-based classification with the use of the MPE mixture models was investigated by applying it to a real dataset.

The family of MPE mixture models presented in this paper only works in the univariate context. Univariate longitudinal data is when each subject only has one measurement taken at each time point. Anderlucci *et al.* (2015) introduced a family of mixture models for clustering multivariate longitudinal data, which is when each subject has multiple measurements taken at each time point. This was an extension of the proposed family of multivariate Gaussian mixture models for clustering univariate longitudinal data in McNicholas and Murphy (2010). Anderlucci *et al.* (2015) used the matrix normal distribution for modeling the density required for their proposed model. Future work will include formulating a family of MPE mixture models for multivariate longitudinal data, which could use the matrix power exponential distribution (Gómez *et al.*, 1998b) for modeling the density. Model-based classification was looked at, so future work will include considering fractionally supervised classification (Vrbik and McNicholas, 2015).

Bibliography

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

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

Andrews, J. L. and McNicholas, P. D. (2012). Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions: The tEIGEN family. *Statistics and Computing*, **22**(5), 1021–1029.

Andrews, J. L., McNicholas, P. D., and Subedi, S. (2011). Model-based classification via mixtures of multivariate t-distributions. *Computational Statistics and Data Analysis*, **55**(1), 520–529.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., and Lindsay, B. (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics*, **46**, 373–388.

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

Browne, R. P. and McNicholas, P. D. (2014a). Estimating common principal components in high dimensions. *Advances in Data Analysis and Classification*, **8**(2), 217–226.

Browne, R. P. and McNicholas, P. D. (2014b). Orthogonal Stiefel manifold optimization for eigen-decomposed covariance parameter estimation in mixture models. *Statistics and Computing*, **24**(2), 203–210.

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

Dang, U. J., Browne, R. P., and McNicholas, P. D. (2015). Mixtures of multivariate power exponential distributions. *Biometrics*, **71**(4), 1081–1089.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, **39**(1), 1–38.

Diggle, P. J., Liang, K.-Y., and Zeger, S. L. (1994). *Analysis of Longitudinal Data*. Oxford University Press, Oxford.

Fox, J., Friendly, G. G., Graves, S., Heiberger, R., Monette, G., Nilsson, H., Ripley, B., Weisberg, S., Fox, M. J., and Suggests, M. (2007). The car package.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, **97**(458), 611–631.

Gómez, E., Gomez-Viilegas, M. A., and Marin, J. M. (1998a). A multivariate generalization of the power exponential family of distributions. *Communications in Statistics – Theory and Methods*, **27**(3), 589–600.

Gómez, E., Gomez-Viilegas, M., and Marin, J. (1998b). A multivariate generalization of the power exponential family of distributions. *Communications in Statistics-Theory and Methods*, **27**(3), 589–600.

Hedeker, D. and Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological methods*, **2**(1), 64.

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

Hunter, D. R. and Lange, K. (2000). Rejoinder to discussion of "Optimization transfer using surrogate objective functions". *Journal of Computational and Graphical Statistics*, **9**, 52–59.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**(430), 773–795.

Landsman, Z. M. and Valdez, E. A. (2003). Tail conditional expectations for elliptical distributions. *North American Actuarial Journal*, **7**(4), 55–71.

Lin, T.-I., McNicholas, P. D., and Hsiu, J. H. (2014). Capturing patterns via parsimonious t mixture models. *Statistics and Probability Letters*, **88**, 80–87.

McLachlan, G. J. and Peel, D. (1998). Robust cluster analysis via mixtures of multivariate t-distributions. In *Lecture Notes in Computer Science*, volume 1451, pages 658–666. Springer-Verlag, Berlin.

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

McNicholas, P. D. (2010). Model-based classification using latent Gaussian mixture models. *Journal of Statistical Planning and Inference*, **140**(5), 1175–1181.

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

McNicholas, P. D. (2016b). Model-based clustering. *Journal of Classification*, **33**(2).

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

McNicholas, P. D. and Murphy, T. B. (2010). Model-based clustering of longitudinal data. *The Canadian Journal of Statistics*, **38**(1), 153–168.

McNicholas, P. D. and Subedi, S. (2012). Clustering gene expression time course data using mixtures of multivariate t-distributions. *Journal of Statistical Planning and Inference*, **142**(5), 1114–1127.

McNicholas, P. D., Murphy, T. B., McDaid, A. F., and Frost, D. (2010). Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. *Computational Statistics and Data Analysis*, **54**(3), 711–723.

McNicholas, P. D., Jampani, R. K., and Subedi, S. (2015). *longclust: Model-Based Clustering and Classification for Longitudinal Data*. R package version 1.2.

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, **80**, 267–278.

Peel, D. and McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing*, **10**(4), 339–348.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2016). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-128.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika*, **86**(3), 677–690.

Pourahmadi, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika*, **87**(2), 425–435.

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

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

Sibuya, M. (1979). Generalized hypergeometric, digamma and trigamma distributions. *Annals of the Institute of Statistical Mathematics*, **31**(1), 373–390.

Tipping, M. E. and Bishop, C. M. (1999). Mixtures of probabilistic principal component analysers. *Neural Computation*, **11**(2), 443–482.

Verbeke, G. and Lesaffre, E. (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics & Data Analysis*, **23**(4), 541–556.

Vrbik, I. and McNicholas, P. D. (2015). Fractionally-supervised classification. *Journal of Classification*, **32**(3), 359–381.

Weiss, R. E. (2005). *Modeling Longitudinal Data: With 72 Figures*. Springer Science & Business Media.