Durham University

Durham E-Theses

Models for multivariate data with latent structures with application in regression and clustering

YINGJUAN ZHANG

How to cite:

ZHANG, YINGJUAN (2024) Models for multivariate data with latent structures with application in regression and clustering. Doctoral thesis, Durham University.

Use policy

The full-text may be used and/or reproduced, and given to third parties in any format or medium, without prior permission or charge, for personal research or study, educational, or not-for-profit purposes provided that:

- a full bibliographic reference is made to the original source
- a https://etheses.durham.ac.uk/id/eprint/15824/ is made to the metadata record in Durham E-Theses
- the full-text is not changed in any way

The full-text must not be sold in any format or medium without the formal permission of the copyright holders.

Please consult the full Durham E-Theses policy for further details.

Academic Support Office, The Palatine Centre, Durham University, Stockton Road, Durham, DH1 3LE
e-mail: e-theses.admin@durham.ac.uk Tel: +44 0191 334 6107
https://etheses.durham.ac.uk

Models for multivariate data with latent structures with application in regression and clustering

Yingjuan Zhang

Supervised by

Dr Jochen Einbeck and Dr Reza Drikvandi

A thesis submitted for the degree of Doctor of Philosophy

Department of Mathematical Sciences

November 25, 2024

Acknowledgments

As I reach the end of this journey, I am reminded that no journey is ever walked alone. I would like to begin by expressing my genuine thanks to those who have guided and supported me throughout my PhD journey.

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr Jochen Einbeck, for your invaluable guidance throughout my academic journey and for always being patient and supportive. This gratitude comes from the bottom of my heart.

I would also like to extend my thanks to my second supervisor, Dr Reza Drikvandi, for the valuable advice and insights you provided during my research. Additionally, I sincerely thank my examination committee, Dr Konstantinos Perrakis and Dr Ardo van den Hout, for your constructive feedback and insightful questions during my viva.

To my mum and dad, thank you for everything.

Abstract

A novel approach is proposed for analyzing clustered and highly correlated multivariate data where a one-dimensional latent structure, parametrized by a single random effect, is used to approximate the data. The estimation methodology makes use of a nonparametric maximum likelihood-type approach, where the random effect distribution is approximated by a discrete mixture, hence allowing for the use of the ECM algorithm for the estimation of all model parameters. We derive the estimators required for the subsequent ECM algorithm under various error variance parameterizations that may depend on the random effect. We extend the proposed model by including covariates, enabling regression of multivariate responses on these covariates, introducing another perspective for analyzing multivariate data whereas typically only one variable is taken as the response variable with the remaining variables constituting a multivariate space of predictors. Accounting for the multivariate response character has several inferential benefits including potentially reduced standard errors and increased powers especially for situations where the main concern is the effect of several correlated response variables on a set of predictors. We further extend this methodology to a two-level version to accommodate repeated measurements. Simulation studies are conducted to assess the accuracy of parameter estimators, the significance of choosing the correct mixture components, and the use of AIC and BIC as model selection criteria. Additionally, the impact of the random effect distribution is examined. Furthermore, several important inferential problems, including clustering using different techniques, projection, ranking, regression on covariates, and regression of an external response on the predicted latent variable, are considered and illustrated with real data examples.

Contents

Chapter 1

Introduction

1.1 Background

Multivariate data is ubiquitous across all fields, and they are rarely distributed homogeneously in space. In practice, one will often observe that the data reside on a latent linear subspace of a smaller dimension than itself, or that the data are concentrated into a certain number of clusters. From a statistical modelling point of view, these two concepts are usually dealt with in isolation or in succession, but not simultaneously. That is, often one will account for the lower 'intrinsic' dimensionality through methods such as principal component analysis, partial least squares, factor analysis (see, e.g., Krzanowski, 2000), etc., and then account for clustering in the resulting lower-dimensional space (for instance, by fitting a mixture model to the projections onto that space), or, less commonly, firstly partition the data into clusters and then apply separate compressions onto linear subspaces within each of them. In some situations (e.g. the price indexes for several goods, educational attainment scores on various abilities, or multiple psychological mental health indicators), such a set of variables are so strongly correlated that they can be considered as intrinsically one-dimensional, meaning that they can be considered to be generated by some latent one-dimensional linear subspace plus noise.

We propose a new approach which is firmly rooted in basic principles of statistical modelling, and hence allows versatile access to routine statistical tasks such as clustering or regression. The basic idea is to consider the approximating lower-dimensional subspace as a latent variable in a multivariate statistical model and to model this latent variable by a random effect. We will develop and implement this very general idea in a more specific framework, where we assume the low-dimensional structure to be a one-dimensional space, i.e. a straight line. We implement this idea through the mixture-based approach for the estimation of random effect

models, hence conveniently enabling clustering of observations along the latent linear subspace, and derive the estimators required for the ensuing ECM algorithm under several error variance parameterizations.

Another perspective of analyzing multivariate data is to single out one variable as the 'response' of interest, with the remaining variables constituting a multivariate space of predictors. This then gives rise to 'multivariate' regression methods including regularization techniques such as the Least Absolute Shrinkage and Selection Operator (LASSO; Hastie et al., 2015). However, there are many situations in which the space of responses should be considered multivariate by itself, for instance when considering the effect of some event (such as, a government intervention on the energy market) on a set of consumer price indexes, or the effect of an educational intervention on a set of outcome measures (such as reading, writing and numeracy). The use of multivariate response models is, however, not very widespread in statistical practice. This may be related to the circumstance that ready-to-use implementations are either only accessible via specialized software (such as SAS or Stata), or are equivalent to fitting separate univariate response models (such as R function `lm`). Accounting for the multivariate response character has several inferential benefits including potentially reduced standard errors and increased powers. This holds especially for the case when the data at hand possesses a repeated measures structure, such as pupils nested within schools. Accounting for the ensuing correlations of lower-level units (here pupils) is crucial in order to obtain correct standard errors of parameter estimates. Random effect models are an effective tool of achieving this, where all upper-level units share a common random effect, hence inducing the required correlations.

A natural way of achieving the multivariate response analysis mentioned above is to extend the model by including covariates. In this scenario, the multivariate data become a multivariate set of response variables, with the included covariates serving as predictors. Addressing the issue of repeated measures mentioned above requires a two-level extension of the model. This extension involves using a single random effect to account for correlations between observations at the same level, thereby ensuring that the estimated effects for several response dimensions remain interconnected.

As mentioned at the beginning of our introduction, our initial model approximates mul-

tivariate data through a one-dimensional space, akin to the basic idea of principal component analysis using only one principal component. Up to this point, the latent structure of the data has been assumed to be linear. However, for multivariate data with a non-linear latent structure, there exist non-linear extensions of principal components such as principal curves. These curves are defined as smooth curves parameterized by a one-dimensional variable, passing through the middle of multivariate data. As an extension of our one-level model, we propose a parametric non-linear model which can approximate multivariate data with non-linear latent structure based on the latent variable technique developed earlier for data with latent linear structures.

1.2 Data Sets

In this section, we will introduce some motivating case studies that inspired the models introduced in this dissertation, and demonstrate the type of data we are working with. The analysis of these datasets can be found in later sections. There are two types of multivariate data that we use: one-level data and two-level data. One-level data refers to data with no hierarchical structure, while two-level data refers to data with a two-level structure, such as repeated measures data or longitudinal data.

1.2.1 Faithful Data (One-level Data)

Let us begin with a simple one-level data set first, the faithful data in R package **MASS**. It is a two-dimensional data set with 272 observations and two variables: eruption time and the waiting time between two eruptions. The scatter plot is shown in Figure 1.1. We observe that there are two clusters and a positive correlation between these two variables. We will return to this example in Section 2.10.1 and illustrate there in detail how exactly this image translates into projections (dimension reduction) and clustering.

1.2.2 Soils Data (One-level Data)

The second dataset we consider is the Soils dataset in the R package **carData** (Fox et al., 2020). It consists of soil chemical characteristics. This data set contains 14 variables with 48

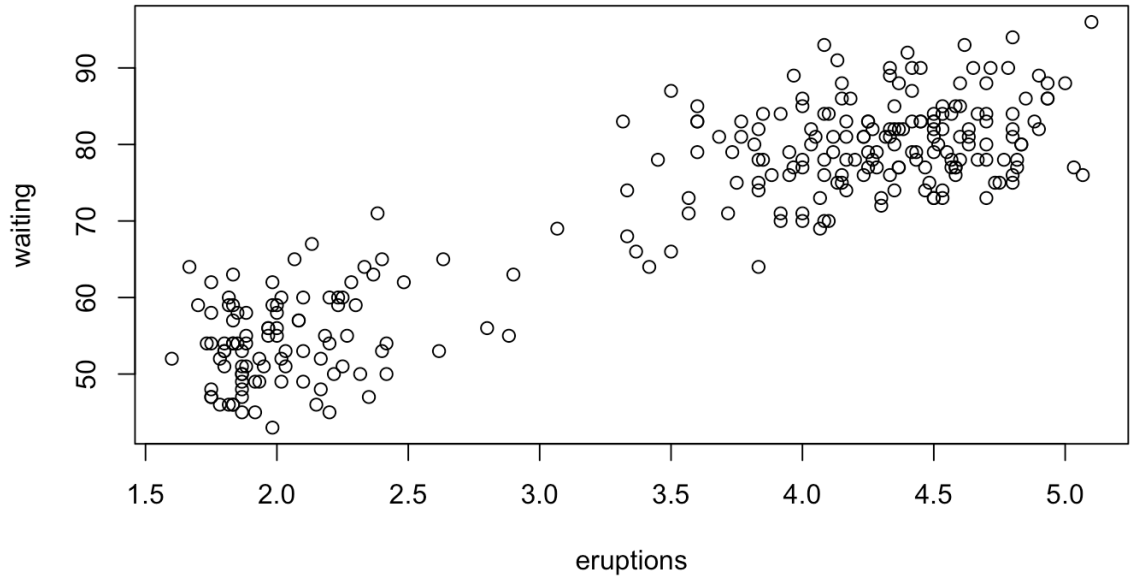

Figure 1.1: Faithful Data

observations. We focus on the 6 chemical element variables. We construct a data frame with $n = 48$ and six variables: Nitrogen, Phosphorous, Calcium, Magnesium, Potassium and Sodium (which are highly correlated, but do not all use the same units). Figure 1.2 shows the correlation between the six variables. The main application of this data set is dimension reduction, and in particular fitting regression models with the model-based scores as predictors and additional variables as response (the variable 'Density' (bulk density in gm/cm^3) is considered and used as the response). Detailed application can be found in Section 2.10.2.

1.2.3 Literacy Survey Data (One-level Data)

The third data set we consider is the IALS data available in R package **mult.latent.reg** (Zhang & Einbeck, 2024b). The data is obtained from the International Adult Literacy Survey (IALS), collected in 13 countries on Prose, Document, and Quantitative scales between 1994 and 1995. The data are reported as the percentage of individuals who could not reach a basic level of literacy in each country. For each country, there are two values for Prose, Document, and Quantitative: one for females and another for males. There is a two-level structure in the data, requiring a two-level model to fit it. However, in the analysis, we choose to focus only on the Prose data. We consider the prose attainment of males and females as a bivariate response for each of the 13 countries. Figure 1.3 shows the relationship of these two variables. Later in Section 2.10.3 we will use this data set to illustrate how our methodology can be effectively

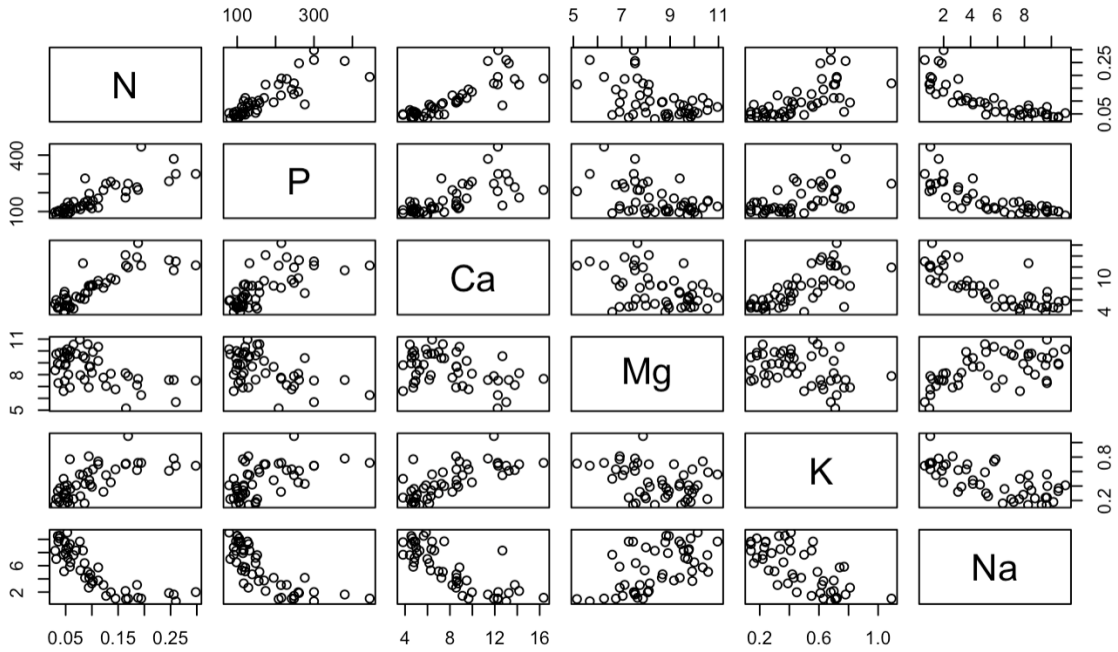

Figure 1.2: Soils data.

used in clustering and ranking. Furthermore, in Section 3.7.4, we consider all three outcome variables as multivariate responses and the gender variable as the covariate; this will require a two-level model.

1.2.4 Foetal Movement Data (One-level Data)

For the fourth data set, we consider a set of foetal movements data collected before and during the Covid-19 pandemic. The study, which was executed by researchers of the Neonatal Research Lab at Durham University, aims to analyse the effects of Covid on fetal development (Reissland et al., 2024). The data were recorded via 4D ultrasound scans from a total of 40 mothers (20 before Covid and 20 during Covid) at 32 weeks gestation, and consist of the number of movements each fetus carries out in relation to the recordable scan length. The ratio of these counts to scan length then form the response variables of interest, with the following five specific movements recorded during the 4D ultrasound scans: upper face movements, head movements, mouth movements, touch movements and eye blink. We are interested in the relationship of these five movements to the variable 'status', which indicates the period during which the data was collected ('pre-Covid' or 'during Covid'). This data set is available in R package **mult.latent.reg**. Figure 1.4 shows the correlation of these 5 variables. These five movements will be considered as a 5-variate response variable, and the status ('pre-Covid' or

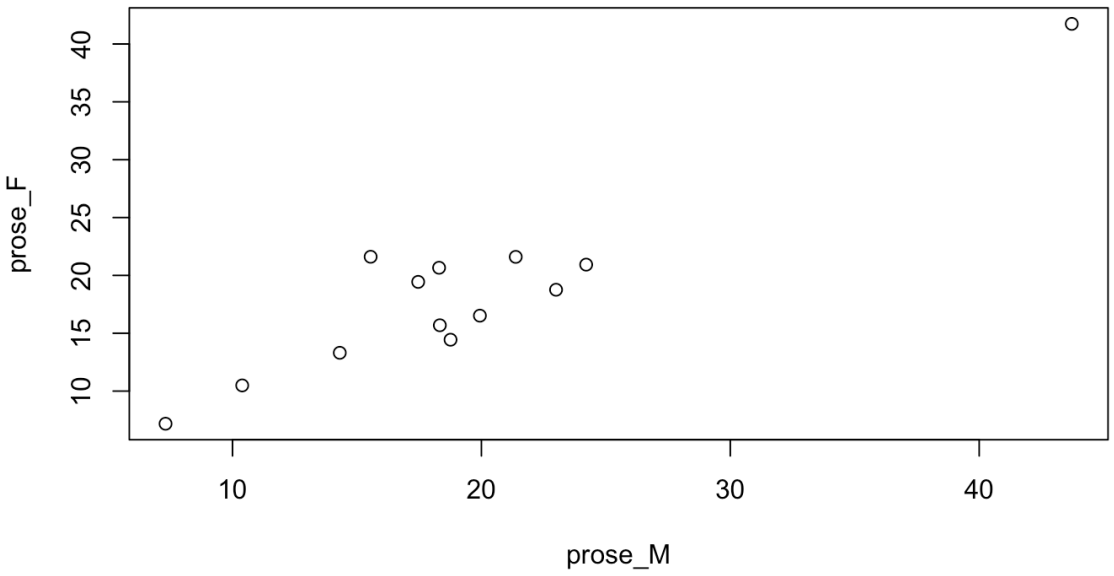

Figure 1.3: IALS data.

'during Covid') will be included as a binary covariate. We use this data to illustrate how our one-level model with covariate can be applied in estimating covariate coefficients, and we will also use a bootstrapped algorithm to obtain the standard errors, which can be found in Section 2.10.4.

1.2.5 Foetal twins' touch movements (Two-level Data)

In our third case study for the two-level model, we consider a data set collected for research on the effects of maternal mental health on prenatal movements in twins and singletons (Reissland et al., 2021). We here work with with slightly reduced data where the singletons are omitted. In the remaining twins' data, from 14 mothers who were pregnant with twins, 11 mothers were available for one scan and 3 were available for two scans, i.e. in total there are 34 observations. There are two touch movement types of the fetus recorded during the scans: self touch and other touch (this is the reason why we omit singletons from the data set: singletons can't touch the 'other' twin). Additionally, the mothers' mental health status was collected on three variables: depression, perceived stress scale and anxiety. The objective is to fit a bivariate response two-level model for self-touch and other touch, taking the correlation of the measurements of fetuses belonging to the same mother into account. Figure 1.5 shows the scatter plot of the two response variables symbolized by values of the upper-level variable 'mother'. We will return to this application in Section 3.7.1 where we compare the estimates of the coefficients and

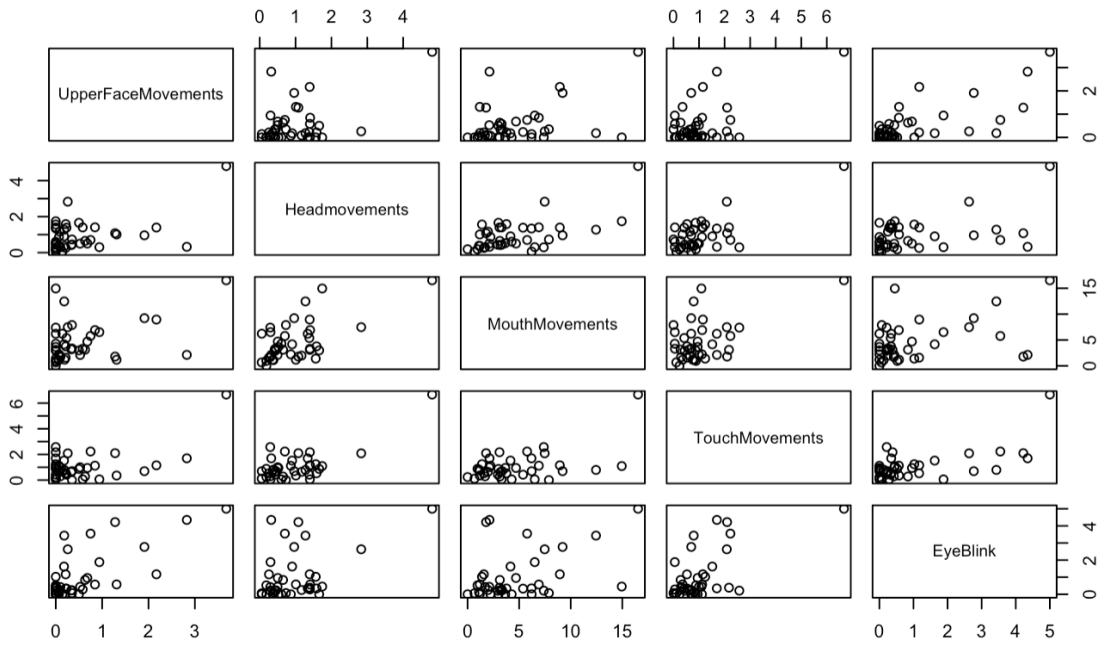

Figure 1.4: IALS data.

standard errors to the ones obtained through fitting separate linear models using the `lmer()` function.

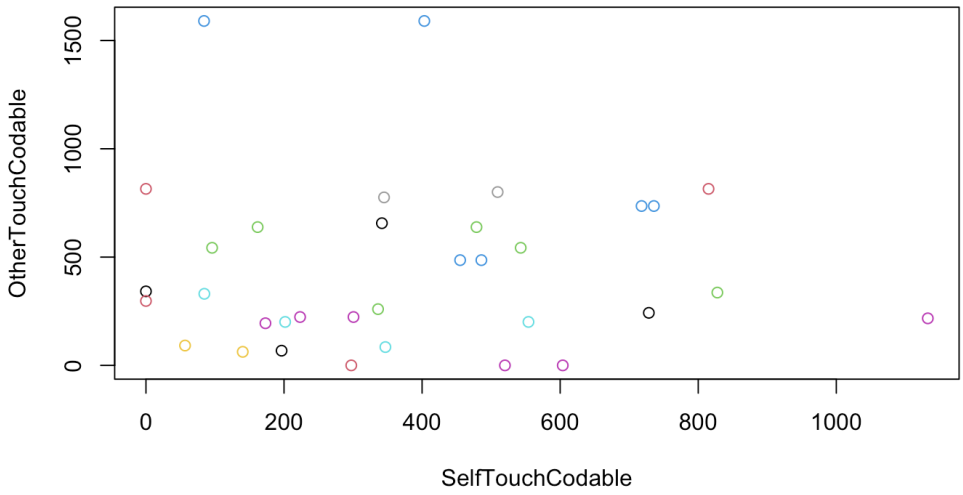

Figure 1.5: Feotal twins' touch movements data, colored by mothers.

1.2.6 Import and Export Data (Two-level Data)

In our first case study for the two-level model, we consider a data set concerning trade in goods and services, or more specifically the transactions in goods and services between residents and non-residents, measured in million USD. The data is extracted from the OECD website (Organisation for Economic Co-operation and Development, 2023b). The variables are given

as the country-wise percentages of imports and exports in relation to the overall GDP in each country. The dataset comprises data from 44 countries, and for our analysis we selected the time period between 2018 and 2022, during which a varying number of observations is available for different countries. Specifically, Australia, Japan, Korea, Mexico, New Zealand, Turkey, United States, China, and Colombia have four observations each, while India, Russia, and Brazil have three observations each. The remaining countries have five observations each. Figure 1.6 visualizes the data where the observations from same country has the same color. This could be considered as a multivariate repeated measures scenario, with unbalanced measurement occasions, and without covariates. We are particularly interested in clustering the countries with respect to their overall export/import activity relative to GDP size, taking within-country correlations across the repeated measurements into account. The detailed analysis of this data set can be found in Section 3.7.2.

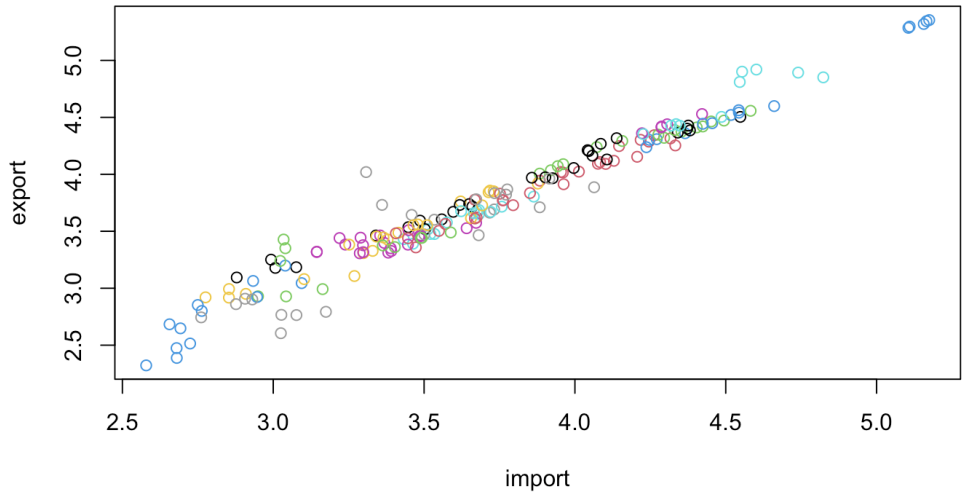

Figure 1.6: Import and export data, coloured by countries.

1.2.7 PIAAC survey of adult skills (Two-level Data)

In our second case study for the two-level model, we consider data from the Programme for the International Assessment of Adult Competencies (PIAAC) survey of adult skills, carried out in 2011 and 2012 by the OECD. The PIAAC survey was designed to assess the proficiency of adults in the key information-processing skills of literacy, numeracy and problem solving (in technology-rich environments). A definition of the three skill types is provided by Organisation for Economic Co-operation and Development (2023a): *Literacy* is the ability to

understand and use information from written texts in a variety of contexts to achieve goals and develop knowledge and potential, *Numeracy* is the ability to use, apply, interpret, and communicate mathematical information and ideas, and *Problem solving in technology-rich environments* refers to the ability to use technology to solve problems and accomplish complex tasks. The survey was designed to be valid cross-nationally and hence to allow comparisons of the adult skill levels between countries, enabling policy makers to identify target areas for improvement in terms of the the skill base of workers. The data used in the analysis is extracted from the PIAAC explorer (https://piaacdataexplorer.oecd.org/ide/idepiaac/) where one can specify to extract different combinations of up to three covariates. There are 28 countries and 17 sub-national regions on the data explorer. For our analysis, we use all three criteria with two covariates: gender and current work status (employee or self employed) for 28 countries and two sub-national entities: Australia, Austria, Canada, Chile, Czech Republic, Denmark, England (UK), Estonia, Finland, Flanders (Belgium), France, Germany, Greece, Hungary, Ireland, Israel, Italy, Japan, Mexico, Netherlands, New Zealand, Norway, Poland, Republic of Korea, Slovak Republic, Slovenia, Spain, Turkey, Sweden, United States. We consider these as 30 'countries' henceforth. More details about this survey can be found on the OECD website (Organisation for Economic Co-operation and Development, 2023a). Figure 1.7 shows the correlation between the three response variables, each plotted against the others and colored by the upper levels (countries). As in the previous example, we are interested in the clustering of countries in the presence of country-level correlations, with the focus shifting here towards the creation of league tables of countries from the posterior random effects. A secondary interest here is in the study of the effect of the covariates on the outcomes. The detailed application of this example can be found in Section 3.7.3.

1.2.8 Mussels Data (One-level Data with non-linear latent structure)

For the multivariate data with non-linear latent structure example, let us first consider the Mussels' muscles data (available in R package **dr**). This is a five-dimensional data set with 82 observations. The five variables are Shell height (denoted as H), Shell length (denoted as L), Shell mass (denoted as S) and Shell width (denoted as W) and Muscle mass (denoted as M).

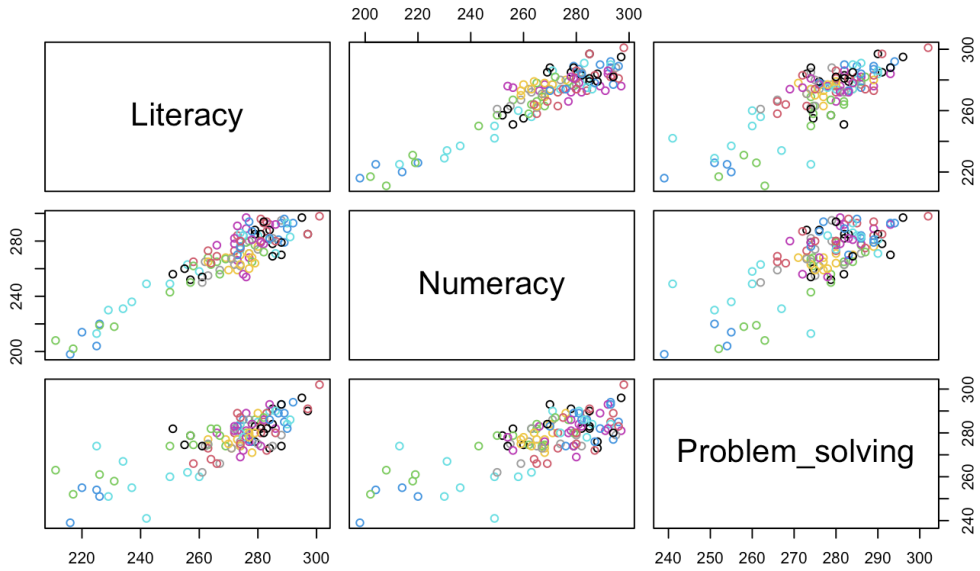

Figure 1.7: PIAAC data, coloured by countries.

The variable Muscle mass is collected as a response variable. Figure 1.8 shows the scatter plot of the four variables H, L, S and W. We will return to this data set in Section 4.1 and Section 4.6, where we focus on describing the data using a latent structure.

1.2.9 Traffic Data (One-level Data with non-linear latent structure)

In our last case study, we consider a data set (with non-linear latent structure) of the speed and flow recorded on Line 5 of a Californian Freeway. The data is available in R package **LPCM** (Einbeck & Evers, 2024). This data set consists of 444 observations on 4 variables; we consider two variables: Lane5Flow and Lane5Speed, Figure 1.9 shows the scatterplot of these two variables. To represent such a data set will require a nonlinear model, at least a quadratic model to capture the shape of the data. We will return to this example in Section 4.7 where we compare our quadratic model, in terms of fitting curves and projections, to principal curves.

1.3 Relationship with Published Papers

The content relating to the one-level model, including the model, simulation, and applications, overlaps with both the journal paper 'A Versatile Model for Clustered and Highly Correlated Multivariate Data' (Zhang & Einbeck, 2024d) and the conference paper 'Simultaneous Linear Dimension Reduction and Clustering with Flexible Variance Matrices' (Zhang & Einbeck, 2022)

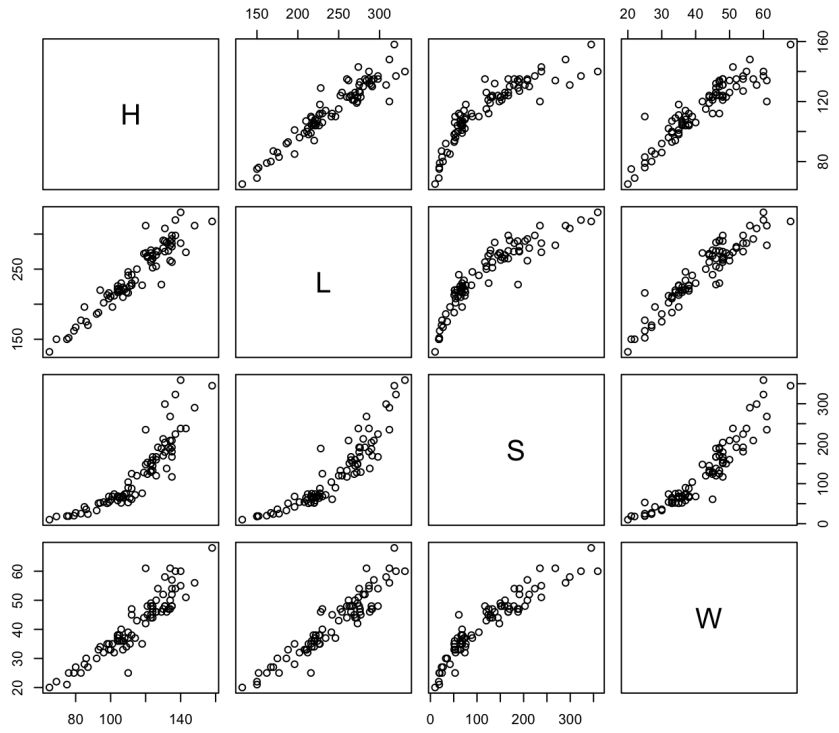

Figure 1.8: Mussels data

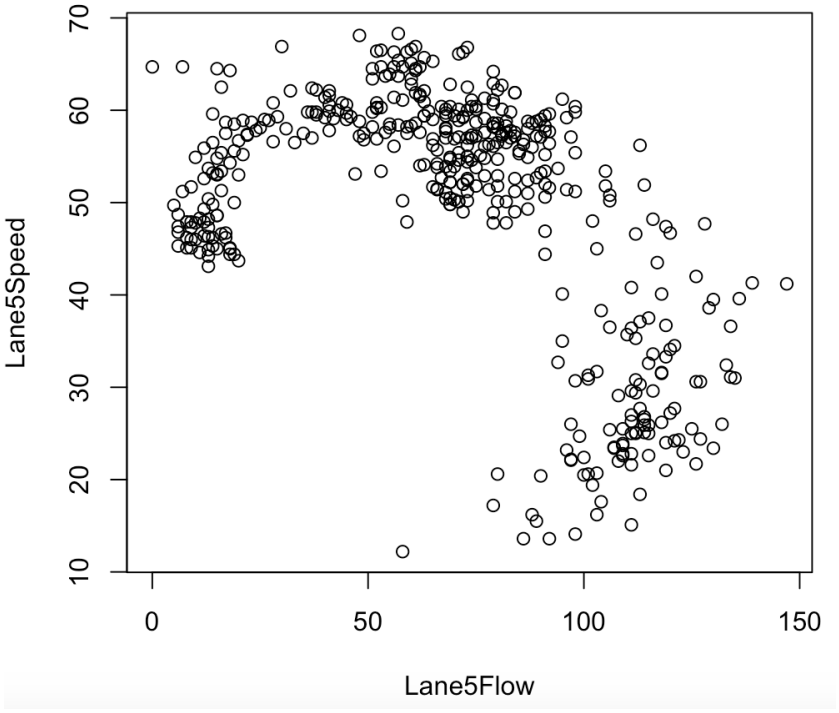

Figure 1.9: Graph showing the scatterplot of the speed-flow data.

in the proceedings of IWSM2022, Trieste. The content of the two-level model, including the model, simulation, and applications, is excerpted from 'A Two-level Multivariate Response Model for Data with Latent Structures' (submitted) (Zhang et al., 2024), the conference paper 'A Multilevel Multivariate Response Model for Data with Latent Structures' (Zhang et al., 2023) in the proceedings of IWSM2023, Dortmund, and a presentation titled 'A Multivariate Response Model for Data with Correlation Structures' at CMStatistics 2023. The section regarding the R package **mult.latent.reg** (Zhang & Einbeck, 2024b) is similar to the conference paper 'R package **mult.latent.reg** for Multivariate Response Scenarios with Latent Structures' (Zhang & Einbeck, 2024c) in the proceedings of IWSM2024, Durham, and the R package help files. The content regarding the non-linear model is similar to 'Directed clustering of multivariate data based on linear or quadratic latent variable models' (Zhang & Einbeck, 2024a).

1.4 Outline of the Dissertation

In Chapter 2, we give details of the one-level model in Section 2.1, 2.2 and 2.3, focusing on the the nonparametric maximum likelihood procedure to estimate the parameters of the model, yielding an ECM algorithm which also automatically estimates masses, mass points, and posterior probabilities of data points being associated with those. In the following Section 2.4, we give detailed derivations of estimators for the parameters used in the ECM algorithm. We solve the identifiability problem in Section 2.5 and in Section 2.6 we give the choice of starting values for the ECM algorithm that has been used in the following simulations and applications. In Section 2.7, we consider extension of the proposed framework allowing for covariates along with a bootstrap approach for the computation of standard errors; In Section 2.8, we conduct simulations that illustrate the accuracy of the proposed estimation methodology. In Section 2.9, we will lay down the clustering and projection operations explicitly. Applications to several real data scenarios are given in Section 2.10, which we also use to illustrate the main application pillars of clustering, dimension reduction, ranking, and regression, explicitly. In Section 2.11 we will discuss two existing methodologies that are somewhat related to our proposed models.

In Chapter 3 we focus on the two-level model. In Section 3.1, 3.2 and 3.3, we introduce the proposed two-level model for multivariate response data, we present an ECM algorithm for

the proposed model, resembling the nonparametric maximum likelihood method for variance component models. The derivations of the estimators for the parameters can be found in Appendix B. Section 3.4 presents the calculation of the intraclass correlation. Section 3.5 shows simulation results that demonstrate the performance and accuracy of the implementation of our proposed model and the robustness of our approach when considering different distributions of the random effect. Section 3.6 gives additional inferential aspects for the two-level model. Section 3.7 provides real data examples that illustrate the main applications of our model, including the fitting of a multivariate response model resulting in reduced standard errors, the construction of league tables, and the clustering of upper-level units based on the fitted model. In Section 3.8, we provide an additional application of both the one-level model and the two level-model in level reduction.

We move on to the non-linear model in Chapter 4. We begin with the general form of a non-linear model in Section 4.1, which is parameterized by a set of basis functions. Then, we consider a simple quadratic form to investigate this model. In Section 4.2, we provide details of the ECM algorithm for the quadratic model, and in Section 4.3, we offer the derivations of the estimators. A simulation study to assess the accuracy of the parameter estimators is presented in Section 4.5. Applications and comparisons with principal curves are given in Section 4.7, where both simulated and real data are used.

In Chapter 5 we will introduce the R package **multiple.latent.reg** that implements methodology for the estimation of multivariate response models with random effects on one or two levels.

Chapter 2

One-level Models

2.1 One-level Model (without covariates)

Let us consider a scenario where the multivariate data $x_i \in \mathbb{R}^m$ are noisy observations scattered along the one-dimensional space $\alpha + \beta z$, where $\alpha, \beta \in \mathbb{R}^m$, and $z \in \mathbb{R}$ is an unobserved instance of a (latent) variable Z. Then the observed data $x_i = (x_{i1}, \ldots, x_{im})^T, i = 1, \ldots, n$, are assumed to be generated from the following random effect model,

$$x_i = \alpha + \beta z_i + \varepsilon_i, \tag{2.1}$$

where $\varepsilon_i \sim N(0, \Sigma_i)$ is m-variate Gaussian noise with a positive definite variance matrix $\Sigma_i \equiv \Sigma(z_i) \in \mathbb{R}^{m \times m}$. It is clear that a model with observation specific $m \times m$ variance matrices is heavily overparametrized, and we will never contemplate fitting this model in full generality. We still provide this general notation in Equation (2.1) as it contains all practically relevant special cases that will be naturally arising, including, of course, the homoscedastic case $\Sigma_i = \Sigma$, $i = 1, \ldots, n$.

For the estimation of the random effect distribution along this line we use the nonparametric maximum likelihood approach, which amounts to representing this distribution by a set of discrete mass points (mixture centres) with some corresponding masses (mixture probabilities). While this may look like a restrictive assumption, it is actually more flexible than the application of a Gaussian random effect, as it allows for multi-modalities in the distributions of the latent variable. Indeed, the mixture character of this approach allows for clustering of observations based on the fitted model.

In consequence, this arrives at a modelling approach with an enormous versatility. Firstly, as just expressed, observations can be clustered based on maximum a posteriori (MAP) probabilities of class membership (Murphy 2012, chapter 11). Secondly, projecting the original

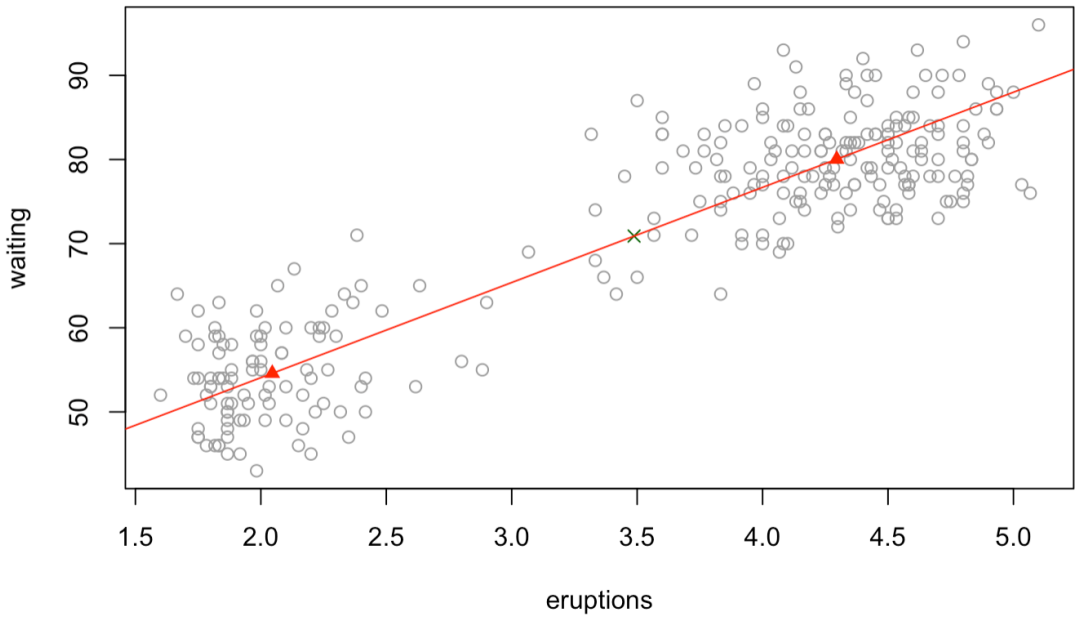

Figure 2.1: Graph showing the estimated one-dimensional space with cluster centres in red and origin of this latent one-dimensional space in green.

data points onto the estimated lower-dimensional space, the dimension of the original multivariate data is reduced (to 1, in the simple framework as discussed in this work), and the compressed data can be used as summary statistic (such as an overall price index across several goods) or for further inferential purposes. Thirdly, the relative order of the posterior random effects (observations 'projected' onto the latent linear subspace) can be used for ranking observations in multivariate data sets. Finally, we will show that it is not difficult to include additional covariates into model (2.1) so that one has de facto a novel tool for multivariate response situations, yielding reduced parameter standard errors as compared to the separate univariate response models. We will give each of these important applications some prominence later in this chapters.

To enable some intuition for how the one-level model operates, let us use the faithful data introduced in Section 1.2.1. The straight line in Figure 2.1 is the one-dimensional latent space $\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} z$ that is parameterized by the latent variable. The red triangles positioned along the straight line are the estimated mixture (cluster) centres. To give some metaphor, one could consider the mixture centres as 'washing pegs' spanning a 'washing rope' holding the clusters.

2.2 Likelihood

The marginal probability density function $f(x_i|\alpha,\beta)$ for observations generated from model (2.1) can be written as

$$f(x_i|\alpha,\beta) = \int f(x_i, z_i|\alpha,\beta)dz_i = \int f(x_i|z_i,\alpha,\beta)\phi(z_i)dz_i, \tag{2.2}$$

where $f(x_i, z_i|\alpha,\beta)$ is the joint probability distribution of observed data x_i and unobserved random effects z_i, and $\phi(\cdot)$ is the density function of the random effect distribution Z. This model is not fully specified since it lacks specific parametrizations of the (unknown) $\Sigma_i = \Sigma(z_i) = \mathrm{Var}(x_i|z_i,\alpha,\beta)$ and ϕ, but let us consider any (additional) parameters involved into these initially as nuisance parameters, and construct appropriate parametrizations for these as we go along.

The initial goal is to find maximum-likelihood estimates for the parameters α and β in model (2.1). Building on the marginal density (2.2), the likelihood of model (2.1) is the following,

$$L(\alpha,\beta) = \prod_{i=1}^{n} \int f(x_i|z_i,\alpha,\beta)\phi(z_i)dz_i$$

with corresponding log-likelihood,

$$l(\alpha,\beta) = \sum_{i=1}^{n} \log\left\{\int f(x_i|z_i,\alpha,\beta)\phi(z_i)dz_i\right\}. \tag{2.3}$$

At this stage a decision needs to be made on how to deal with the integral figuring in Equation (2.3). In principle, one could do this based on a Gaussianity assumption on $\phi(\cdot)$, as common in the mixed model context, in this case leading us back to a factor analysis framework. However, for reasons expressed in Section 2.1, we have decided here differently, and employ instead Aitkin's Nonparametric Maximum Likelihood approach (Aitkin, 1996b). Here, the random effect distribution Z is approximated by a discrete mixture distribution, say $\tilde{Z}$, which is supported on a finite number of mass points $z_1,\ldots,z_K$ with masses $P(\tilde{Z} = z_k) = \pi_k$, $k = 1,\ldots,K$. This discrete mixture facilitates a simple approximation of the marginal likelihood

which just involves sums rather integrals, i.e.

$$l(\alpha, \beta) \approx \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} f(x_i|z_k, \alpha, \beta)\pi_k \right\}. \tag{2.4}$$

Laird (1978) showed that the marginal likelihood (2.3) can be approximated arbitrarily well by (2.4) with a finite set of mass points. We see that this has now become a mixture-type problem, with each mixture component k representing a latent 'class' within the domain of Z (we will use the terms class and component interchangeably henceforth). The EM algorithm (Dempster et al., 1977) is one of the most widely used algorithms for the estimation of parameters in mixture models.

Denote by $f_{ik} = P(x_i|\tilde{Z} = z_k) = f(x_i|z_k, \alpha, \beta)$ the probability density of x_i conditional on class k. Then we know that

$$P(x_i, \tilde{Z} = z_k) = P(x_i|\tilde{Z} = z_k)P(\tilde{Z} = z_k) = f_{ik}\pi_k.$$

Since it is in practice unknown which component each observations belongs to, this is an incomplete data scenario. We describe the missing information on the component membership by an indicator variable

$$G_{ik} = \begin{cases} 1, & \text{if observation } i \text{ belongs to component } k \\ 0, & \text{otherwise.} \end{cases}$$

This defines complete data $(x_i, G_{i1}, ..., G_{iK})$, $i = 1, ..., n$, with probability

$$P(x_i, G_{i1}, ..., G_{iK}) = \prod_{k=1}^{K} (f_{ik}\pi_k)^{G_{ik}}$$

and resulting complete data likelihood $\prod_{i=1}^{n} \prod_{k=1}^{K} (f_{ik}\pi_k)^{G_{ik}}$. Then we can obtain the expected

complete log-likelihood

$$
\begin{aligned}
l_c &= \sum_{i=1}^{n} \mathbb{E}\left[\log\left(\prod_{k=1}^{K}(\pi_k f_{ik})^{G_{ik}}\right)|x_i\right] \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{E}\left[G_{ik}|x_i\right]\log\left(\pi_k f_{ik}\right) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\log\pi_k + \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\log f_{ik}
\end{aligned}
\tag{2.5}
$$

where $w_{ik} = \mathbb{E}\left[G_{ik}|x_i\right] = P(G_{ik} = 1|x_i) = P(\tilde{Z} = z_k|x_i)$, which is the probability of each observation i belonging to component k. For the component-specific densities f_{ik}, we specify, conditional on the mixture centres z_k, a multivariate Gaussian model

$$
f_{ik} = \frac{1}{(2\pi)^{m/2}}\frac{1}{|\Sigma_k|^{1/2}}\exp\left(-\frac{1}{2}(x_i - \alpha - \beta z_k)^T\Sigma_k^{-1}(x_i - \alpha - \beta z_k)\right) \tag{2.6}
$$

where we allow the variance matrices $\Sigma_k = \Sigma(z_k)$ to depend on the cluster k but not on observation i, hence reducing the complexity of the original, fully heteroscedastic, variance specification considerably. The terms $\alpha + \beta z_k$ can be interpreted as the mixture centers in the original data space, spanned along the line $\alpha + \beta z$. Note that the right hand side of (2.4) is then the log-likelihood corresponding to the 'approximative' model

$$
x_i|z_k, \alpha, \beta \sim N(\alpha + \beta z_k, \Sigma_k) \text{ with probability } \pi_k, \tag{2.7}
$$

where we treat the mass points z_k, $k = 1, \ldots, K$, and their associated masses π_k as unknown parameters to be estimated in the EM algorithm alongside with the model parameters α and β. This model can be seen as a Gaussian mixture model with structured mean function and component-specific variances, or as a multivariate response version of the 'nonparametric maximum likelihood' (NPML) approach for the estimation of mixture masses and mass points in random effect models (Aitkin 1996b, Aitkin et al. 2009, chapter 8).

Several reduced, parsimonious, parameterizations of the variance matrices Σ_k are possible in order to describe the shape of the clusters around the mixture centres. The simplest case (i) would be a constant and diagonal matrix $\Sigma_k \equiv \Sigma = \text{diag}(\sigma_j^2)_{\{1\leq j\leq m\}} \in \mathbb{R}^{m\times m}$, which gives the same variance specification to all K components of the mixture. Second (ii), we consider

using different diagonal variance matrices for different components, $\Sigma_k = \text{diag}(\sigma^2_{jk})_{\{1\leq j\leq m\}} \in \mathbb{R}^{m\times m}$, which yields an improvement for estimating data that has clusters of different sizes. Third (iii), we consider using the same full (unrestricted) variance matrix, $\Sigma_k \equiv \Sigma \in \mathbb{R}^{m\times m}$, to capture the correlation of variables. Finally (iv), different full (unrestricted) variance matrices, $\Sigma_k \in \mathbb{R}^{m\times m}$ give better estimations when dealing with clusters that differ by shape and size. In line with (2.6) and (2.7), our notation in what follows will be tailored to this most general case (iv); with the results for the reduced parameterizations naturally deriving from this. Note that after applying the NPML approach, the $\Sigma(z_i)$ becomes $\Sigma(z_k)$, and the assumption of random-effect specific errors feeds into the Σ_k. If we do not actually have Σ's which depend on k, then we also have implicitly $\Sigma(z_i) = \Sigma$. The $\Sigma(z_i)$ are only 'enabled' by variance parametrization (ii) and (iv).

2.3 ECM Algorithm and Computational Considerations

Now we can set up the EM algorithm for estimating model (2.7). It is noted that the developments in this subsection are for a fixed number of components, K. The question of choosing K is considered as a model selection problem and will be addressed through the use of model selection criteria as illustrated in later sections.

E-step

Using the Bayes' theorem, we obtain the posterior probability of observation i belonging to component k (Aitkin et al., 2009),

$$w_{ik} = P(\tilde{Z} = z_k|x_i) = \frac{P(\tilde{Z} = z_k)P(x_i|\tilde{Z} = z_k)}{\sum_l P(\tilde{Z} = z_l)P(x_i|\tilde{Z} = z_l)} = \frac{\pi_k f_{ik}}{\sum_l \pi_l f_{il}}. \tag{2.8}$$

M-step

Using expression (2.6) for the component-wise densities f_{ik}, the expected complete data log-

likelihood becomes

$$\begin{aligned} l_c = \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\log(\pi_k) + \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{1}{2}w_{ik}\log(|\Sigma_k|) + \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{m}{2}\log(2\pi)w_{ik} \\ + \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{1}{2}w_{ik}(x_i-\alpha-\beta z_k)^T\Sigma_k^{-1}(x_i-\alpha-\beta z_k). \end{aligned} \tag{2.9}$$

Taking partial derivatives of l_c with respect to each parameter gives the score equations. We then obtain the following estimators for the parameters α, β, z_k and π_k by setting these score equations to zeros and solving them. (The derivations of the estimators of these parameters can be found in the next section.)

$$\hat{z}_k = \frac{\sum_{i=1}^{n} w_{ik}\hat{\beta}^T\hat{\Sigma}_k^{-1}(x_i-\hat{\alpha})}{\sum_{i=1}^{n} w_{ik}\hat{\beta}^T\hat{\Sigma}_k^{-1}\hat{\beta}}. \tag{2.10}$$

$$\hat{\beta} = \left(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{\Sigma}_k^{-1}\hat{z}_k^2\right)^{-1}\left(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{\Sigma}_k^{-1}(x_i-\hat{\alpha})\hat{z}_k\right) \tag{2.11}$$

$$\hat{\alpha} = \left(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{\Sigma}_k^{-1}\right)^{-1}\left(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{\Sigma}_k^{-1}(x_i-\hat{\beta}\hat{z}_k)\right) \tag{2.12}$$

For the mixture probabilities, since $\sum_{k=1}^{K}\pi_k = 1$, we need to apply a Lagrange multiplier by letting $\partial\left(l-\lambda(\sum_{k=1}^{K}\pi_k-1)\right)/\partial\pi_k = 0$, yielding (Aitkin et al., 2009),

$$\hat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n} w_{ik}. \tag{2.13}$$

Estimators for the flexible variance specifications are again obtained by equating the corresponding partial derivatives to zero, giving results as follows:

(i) $\Sigma = \mathrm{diag}(\sigma_j^2)_{\{1\le j\le m\}} \in \mathbb{R}^{m\times m}$, $k = 1, ..., K$,

$$\hat{\sigma}_j^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(x_{ij}-\hat{\alpha}_j-\hat{\beta}_j\hat{z}_k)^2; \tag{2.14}$$

(ii) $\Sigma_k = \text{diag}(\sigma_{jk}^2)_{\{1\leq j\leq m\}} \in \mathbb{R}^{m\times m}$, $k = 1, ..., K$,

$$\hat{\sigma}_{jk}^2 = \frac{\sum_{i=1}^n w_{ik}(x_{ij} - \hat{\alpha}_j - \hat{\beta}_j \hat{z}_k)^2}{\sum_{i=1}^n w_{ik}}; \tag{2.15}$$

(iii) $\Sigma = \Sigma_1 = ... = \Sigma_k \in \mathbb{R}^{m\times m}$, $k = 1, ..., K$,

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n \sum_{k=1}^K w_{ik}(x_i - \hat{\alpha} - \hat{\beta}\hat{z}_k)(x_i - \hat{\alpha} - \hat{\beta}\hat{z}_k)^T; \tag{2.16}$$

(iv) $\Sigma_k \in \mathbb{R}^{m\times m}, k = 1, ..., K$,

$$\hat{\Sigma}_k = \frac{\sum_{i=1}^n w_{ik}(x_i - \hat{\alpha} - \hat{\beta}\hat{z}_k)(x_i - \hat{\alpha} - \hat{\beta}\hat{z}_k)^T}{\sum_{i=1}^n w_{ik}}. \tag{2.17}$$

It is evident that all of these estimators depend on the weights w_{ik}, hence requiring the use of the EM algorithm which iterates between finding the above estimates and updating the weights given the estimates.

It is noted from Equations (2.10), (2.11) and (2.12) that these involve many inversions of the estimated variance matrices $\hat{\Sigma}_k$. This can make the EM algorithm computationally unstable especiallly under the component-specific variance parameterizations (ii) and (iv). Therefore, in our implementation of the above EM algorithm, we disentangle the M-step updates of $\hat{\Sigma}_k$ from those of $\hat{\alpha}$, $\hat{\beta}$ and $\hat{z}_k$. Specifically, the updates (2.10), (2.11) and (2.12) are executed in a simplified form where $\hat{\Sigma}_k \equiv \text{diag}(\sigma^2)$, for some constant σ^2 which does not need to be specified since it cancels out from the resulting simplified update equations, yielding

$$\hat{z}_k = \frac{\hat{\beta}^T \sum_{i=1}^n w_{ik}(x_i - \hat{\alpha})}{\hat{\beta}^T\hat{\beta}\sum_{i=1}^n w_{ik}}. \tag{2.18}$$

$$\hat{\beta} = \frac{\sum_{i=1}^n \sum_{k=1}^K w_{ik}\hat{z}_k x_i - \frac{1}{n}(\sum_{i=1}^n x_i)(\sum_{i=1}^n \sum_{k=1}^K w_{ik}\hat{z}_k)}{\sum_{i=1}^n \sum_{k=1}^K w_{ik}\hat{z}_k^2 - \frac{1}{n}(\sum_{i=1}^n \sum_{k=1}^K w_{ik}\hat{z}_k)^2}; \tag{2.19}$$

$$\hat{\alpha} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i - \hat{\beta}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k\right); \tag{2.20}$$

That is, in our implementation, within each M-step, we cycle a small number times (five will be sufficient) between (2.18), (2.19), and (2.20), note that $\hat{\alpha}$ and $\hat{\beta}$ from the previous iteration of the M-step are used in the first cycle of the iteration within the M-step in Equation (2.18), then we update $\hat{\pi}_k$ via (2.13), followed by the respective update of the variance matrices according to any of (2.14), (2.15), (2.16), or (2.17) depending on the variance parameterization. The resulting updated parameters are then used in the upcoming E-step (2.8). The simulation studies in Section 2.8 will confirm that this approach yields accurate parameter estimates. It is noted that in spirit, this algorithm is more of an ECM algorithm (Meng & Rubin, 1993) than a general EM algorithm. If we compare the estimators from our methodology to the first motivating example in Meng and Rubin (1993), then, conceptually, the entire cycle involving z_k, β, and α corresponds to Equation (2.2) in Meng and Rubin (1993), while the estimator for the variance corresponds to Equation (2.3) in Meng and Rubin (1993). Thus, for the remainder of the thesis, we will refer to this procedure as an ECM algorithm. A detailed procedure for the ECM algorithm is given in Algorithm 1 below. Note that these expressions resemble the equations for weighted linear regression. Similar observations were drawn by Aitkin et al. (2009) for univariate response models under the GLM framework.

2.4 Derivations

In this section, we are going to show the derivations of the estimators for the parameters used in the ECM algorithm described in Section 2.3. Let us write the expected complete data log-likelihood again,

$$\begin{aligned} l_c = \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\log(\pi_k) + \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{1}{2}w_{ik}\log(|\Sigma_k|) + \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{m}{2}\log(2\pi)w_{ik} \\ + \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{1}{2}w_{ik}(x_i - \alpha - \beta z_k)^T\Sigma_k^{-1}(x_i - \alpha - \beta z_k). \end{aligned} \tag{2.21}$$

Algorithm 1 ECM Algorithm

1. Initialization:

(i) Choose the number of mixture components, K, where K is a positive integer.

(ii) Select a variance parameterization; four options are available.

(iii) Choose starting values for the parameters: π_k, α, β, z_k, Σ_k; four options are available.

(iv) Select the number of iterations, s; 20 iterations is suggested.

2. Iterations:

E-step

For each k, compute the posterior probability of observation i belonging to component k, according to Equation (2.8).

M-step

steps $\leftarrow 0$
while *steps* $\leq s$ **do**
  counter $\leftarrow 0$ ▷ Reset counter for each step
  while *counter* ≤ 5 **do**
    Update α, β and z_k, cycle between Equations (2.20), (2.19), and (2.18).
    counter $\leftarrow$ *counter* $+ 1$
  end while
  Update π_k via Equation (2.13)
  Update Σ_k according to Equation (2.14), (2.15), (2.16), or (2.17), depending on the selected variance parameterization.
  steps $\leftarrow$ *steps* $+ 1$
end while

3. Output: Return the estimated parameters.

Derivation for $\hat{\pi}_k$

We are under the constraint $\sum_{k=1}^{K} \pi_k = 1$, and this can be addressed by applying a Lagrange multiplier. Define,

$$L(\pi_k) = l_c - \lambda(\sum_{k=1}^{K} \pi_k - 1),$$

then by taking the partial derivative of $L(\pi_k)$ with respect to π_k and letting it to be zero, we obtain,

$$\sum_{i=1}^{n} w_{ik} \frac{1}{\pi_k} - \lambda = 0,$$

then,

$$\pi_k = \frac{\sum_{i=1}^{n} w_{ik}}{\lambda},$$

take the summation over k on both sides, we obtain,

$$\sum_{k=1}^{K} \pi_k = \frac{\sum_{k=1}^{K} \sum_{i=1}^{n} w_{ik}}{\lambda} = 1,$$

since $\sum_{k=1}^{K} \sum_{i=1}^{n} w_{ik} = n$, so,

$$\lambda = n,$$

then we obtain,

$$\hat{\pi}_k = \frac{\sum_{i=1}^{n} w_{ik}}{n}. \tag{2.22}$$

Derivation for $\hat{\alpha}$

Using the result (derived by Petersen and Pedersen (2012)) of the derivatives of matrices, vectors, and scalars, where W is symmetric,

$$\frac{\partial}{\partial s}(x - s)^T W (x - s) = -2W(x - s).$$

We obtain the following by taking the partial derivative of the log-likelihood with respect to α,

$$\frac{\partial l_c}{\partial \alpha} = \sum_{i=1}^{n} \sum_{k=1}^{K} -\frac{1}{2} w_{ik} (-2)(\Sigma_k)^{-1}(x_i - \alpha - \beta z_k),$$

then letting it to be zero and solving it,

$$\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\Sigma_k^{-1}(x_i - \alpha - \beta z_k) = 0, \tag{2.23}$$

$$\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\Sigma_k^{-1}(x_i - \beta z_k) = \alpha\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\Sigma_k^{-1},$$

we obtain the estimator for α,

$$\hat{\alpha} = \left(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{\Sigma}_k^{-1}\right)^{-1}\left(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{\Sigma}_k^{-1}(x_i - \hat{\beta}\hat{z}_k)\right), \tag{2.24}$$

which corresponds to Equation (2.12) in Section 2.3.

In our implementation of the ECM algorithm, we assume (only temporarily within each M-step, before actually updating the $\hat{\Sigma}_k$) that $\hat{\Sigma}_k \equiv \text{diag}(\sigma^2)$, for some constant σ^2 which does not need to be specified since it cancels out from the resulting simplified update equations, then Equation (2.23) becomes:

$$\Sigma^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(x_i - \alpha - \beta z_k) = 0,$$

and then multiply Σ on both sides, we obtain,

$$\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(x_i - \alpha - \beta z_k) = 0,$$

then,

$$\hat{\alpha} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i - \hat{\beta}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k\right). \tag{2.25}$$

This is the estimator of α used in the M-step in implementing the ECM algorithm, which corresponds to Equation (2.20) in Section 2.3.

Derivation for $\hat{\beta}$

For the derivation of β, we use the result of the following, which is derived by Petersen and Pedersen (2012),

$$\frac{\partial}{\partial A}(x - As)^T W(x - As) = -2W(x - As)s^T.$$

By taking partial derivative of the l_c with respect to β, we obtain,

$$\frac{\partial l_c}{\partial \beta} = \sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K} -\frac{1}{2} w_{ik}(-2)\Sigma_k^{-1}(x_{ij} - \alpha - \beta z_k) z_k^T.$$

Since z_k is a scalar, $z_k = z_k^T$, and by letting the above equation to be zero and solving it,

$$\sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K} w_{ik}\Sigma_k^{-1}(x_{ij} - \alpha) z_k - \sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K} w_{ik}\Sigma_k^{-1}\beta z_k^2 = 0, \tag{2.26}$$

then,

$$\hat{\beta} = \left(\sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K} w_{ik}\hat{\Sigma}_k^{-1}\hat{z}_k^2\right)^{-1}\left(\sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K} w_{ik}\hat{\Sigma}_k^{-1}(x_{ij} - \hat{\alpha})\hat{z}_k\right) \tag{2.27}$$

which corresponds to Equation (2.11) in Section 2.3.

Again, we assume $\hat{\Sigma}_k \equiv \text{diag}(\sigma^2)$ in implementation, then Equation (2.26) can be rewritten as,

$$\Sigma^{-1}\sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K} w_{ik}(x_{ij} - \alpha) z_k - \Sigma^{-1}\sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K} w_{ik}\beta z_k^2 = 0,$$

multiply Σ on both sides, we could obtain,

$$\hat{\beta} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k x_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k)}{\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k^2 - \frac{1}{n}(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k)^2}, \tag{2.28}$$

which corresponds to Equation (2.19) in Section 2.3 and is being used in the R implementation.

Derivation for $\hat{z}_k$

For the derivation of z_k, we use the result of the following, which is derived by Petersen and Pedersen (2012),

$$\frac{\partial}{\partial s}(x - As)^T W(x - As) = -2A^T W(x - As).$$

By taking partial derivative of the l_c with respect to z_k, we obtain,

$$\frac{\partial l_c}{\partial z_k} = \sum_{i=1}^{n} -\frac{1}{2} w_{ik}(-2)\beta^T\Sigma_k^{-1}(x_i - \alpha - \beta z_k),$$

then,

$$\sum_{i=1}^{n} w_{ik}\beta^T\Sigma_k^{-1}(x_i - \alpha - \beta z_k) = 0, \tag{2.29}$$

we obtain,

$$\hat{z}_k = \frac{\sum_{i=1}^{n} w_{ik}\hat{\beta}^T\hat{\Sigma}_k^{-1}(x_i - \hat{\alpha})}{\sum_{i=1}^{n} w_{ik}\hat{\beta}^T\hat{\Sigma}_k^{-1}\hat{\beta}}, \tag{2.30}$$

which corresponds to Equation (2.10) in Section 2.3.

With the assumption of $\hat{\Sigma}_k \equiv \text{diag}(\sigma^2)$ in implementation, the Equation (2.29) becomes,

$$\sum_{i=1}^{n} w_{ik}\beta^T(\sigma^2 I_m)(x_i - \alpha - \beta z_k) = 0,$$

then,

$$\sigma^2 \sum_{i=1}^{n} w_{ik}\beta^T I_m(x_i - \alpha - \beta z_k) = 0,$$

where the σ^2 can be canceled out, and we will have,

$$\sum_{i=1}^{n} w_{ik}\beta^T(x_i - \alpha - \beta z_k) = 0,$$

The estimator of z_k used in the implementation is the following,

$$\hat{z}_k = \frac{\hat{\beta}^T \sum_{i=1}^{n} w_{ik}(x_i - \hat{\alpha})}{\hat{\beta}^T\hat{\beta} \sum_{i=1}^{n} w_{ik}}, \tag{2.31}$$

which corresponds to Equation (2.18) in Section 2.3, and it is being used in the R implementation.

Derivation for $\hat{\Sigma}_k$

For the derivation of Σ, we use the results of the derivatives of vectors and matrices (Petersen & Pedersen, 2012),

$$\frac{\partial}{\partial W}(x - s)^T W(x - s) = (x - s)(x - s)^T, \tag{2.32}$$

and

$$\frac{\partial}{\partial W}\log(|W|) = (W^{-1})^T. \tag{2.33}$$

Rewrite (2.21) to be

$$\begin{aligned}\tilde{l_c} = \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\log(\pi_k) + \sum_{i=1}^{n}\sum_{k=1}^{K}\frac{1}{2}w_{ik}\log(|\Sigma_k|^{-1}) + \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{m}{2}\log(2\pi)w_{ik} \\ + \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{1}{2}w_{ik}(x_i-\alpha-\beta z_k)^T\Sigma_k^{-1}(x_i-\alpha-\beta z_k).\end{aligned} \tag{2.34}$$

By taking partial derivative of the $\tilde{l_c}$ with respect to Σ_k^{-1} and by letting it to be zero, we obtain,

$$\sum_{i=1}^{n}\frac{1}{2}w_{ik}\left((\Sigma_k^{-1})^{-1}\right)^T + \sum_{i=1}^{n} -\frac{1}{2}w_{ik}(x_i-\alpha-\beta z_k)(x_i-\alpha-\beta z_k)^T = 0,$$

since Σ_k is symmetric, then $\Sigma_k^T = \Sigma_k$,

$$\sum_{i=1}^{n} w_{ik}\Sigma_k = \sum_{i=1}^{n} w_{ik}(x_i-\alpha-\beta z_k)(x_i-\alpha-\beta z_k)^T,$$

we obtain (variance parameterization(iv)),

$$\hat{\Sigma}_k = \frac{\sum_{i=1}^{n} w_{ik}(x_i-\hat{\alpha}-\hat{\beta}\hat{z}_k)(x_i-\hat{\alpha}-\hat{\beta}\hat{z}_k)^T}{\sum_{i=1}^{n} w_{ik}}, \tag{2.35}$$

which corresponds to Equation (2.17) in Section 2.3.

Derivation for $\hat{\sigma}_{jk}^2$

When $\Sigma_k \in R^m$ is diagonal, that is $\Sigma_k = \text{diag}(\sigma_{jk}^2)_{\{1\le j\le m\}}$, where $k = 1,\ldots,K$,

$$\Sigma_k = \begin{pmatrix} \sigma_{1k}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{2k}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{jk}^2 \end{pmatrix}, \tag{2.36}$$

and $|\Sigma_k| = \prod_{j=1}^{m} \sigma_{jk}^2$, since $|\Sigma_k|^{-1} = |\Sigma_k^{-1}|$,

$$|\Sigma_k^{-1}| = \begin{vmatrix} \frac{1}{\sigma_{1k}^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_{2k}^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_{jk}^2} \end{vmatrix} = \prod_{j=1}^{m} \frac{1}{\sigma_{jk}^2}, \tag{2.37}$$

The log-likelihood function from the previous section is the following,

$$\tilde{l_c} = \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\log(\pi_k) + \sum_{i=1}^{n}\sum_{k=1}^{K} \frac{1}{2} w_{ik}\log(|\Sigma_k|^{-1}) + \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{m}{2}\log(2\pi) w_{ik} + \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{1}{2} w_{ik}(x_i - \alpha - \beta z_k)^T \Sigma_k^{-1}(x_i - \alpha - \beta z_k),$$

and the $\log(|\Sigma_k|^{-1})$ will become,

$$\log(|\Sigma_k|^{-1}) = \log(|\Sigma_k^{-1}|) = \log(\frac{1}{\sigma_{1k}^2}.\frac{1}{\sigma_{2k}^2}\cdots\frac{1}{\sigma_{mk}^2}) = -2\sum_{j=1}^{m}\log\sigma_{jk},$$

then the log-likelihood function $\tilde{l_c}$ will become,

$$\tilde{l}_{new} = constant + \sum_{i=1}^{n} -\frac{1}{2} w_{ik}(-2)\sum_{j=1}^{m}\log\sigma_{jk} + \sum_{i=1}^{n}\sum_{j=1}^{m} -\frac{1}{2} w_{ik}\frac{(x_{ij} - \alpha_j - \beta_j z_k)^2}{\sigma_{jk}^2},$$

by taking partial derivative of the $\tilde{l}_{new}$ with respect to σ_{jk} and by letting it to be zero, we obtain,

$$\sum_{i=1}^{n} -w_{ik}\frac{1}{\sigma_{jk}} + \sum_{i=1}^{n} w_{ik}(x_{ij} - \alpha_j - \beta_j z_k)^2\sigma_{jk}^{-3} = 0,$$

then,

$$\sum_{i=1}^{n} w_{ik}\frac{1}{\sigma_{jk}} = \frac{1}{\sigma_{jk}^3}\sum_{i=1}^{n} w_{ik}(x_{ij} - \alpha_j - \beta_j z_k)^2,$$

we then obtain variance parameterization (ii),

$$\hat{\sigma}_{jk}^2 = \frac{\sum_{i=1}^{n} w_{ik}(x_{ij} - \hat{\alpha}_j - \hat{\beta}_j\hat{z}_k)^2}{\sum_{i=1}^{n} w_{ik}}, \tag{2.38}$$

which corresponds to Equation (2.15) in Section 2.3.

Derivation for $\hat{\Sigma}$

For the derivation of parameter Σ, again, we use Equations 2.32 and 2.33. When $\Sigma_k \equiv \Sigma$, the log-likelihood function (2.21) can be rewrite as,

$$\begin{aligned}\tilde{l_c} = \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\log(\pi_k) + \sum_{i=1}^{n}\sum_{k=1}^{K}\frac{1}{2}w_{ik}\log(|\Sigma|^{-1}) + \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{m}{2}\log(2\pi)w_{ik} \\ + \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{1}{2}w_{ik}(x_i - \alpha - \beta z_k)^T\Sigma^{-1}(x_i - \alpha - \beta z_k).\end{aligned} \tag{2.39}$$

By taking partial derivative of the $\tilde{l_c}$ with respect to Σ^{-1} and by letting it to be zero, we obtain,

$$\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{1}{2}w_{ik}\left((\Sigma^{-1})^{-1}\right)^T + \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{1}{2}w_{ik}(x_i - \alpha - \beta z_k)(x_i - \alpha - \beta z_k)^T = 0,$$

since Σ_k is symmetric, and $\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik} = n$ then we obtain variance parameterization (iii),

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(x_i - \hat{\alpha} - \hat{\beta}\hat{z}_k)(x_i - \hat{\alpha} - \hat{\beta}\hat{z}_k)^T, \tag{2.40}$$

which corresponds to Equation (2.16) in Section 2.3.

Derivation for $\hat{\sigma}_j$

When $\Sigma_{m\times m}$ is diagonal, that is $\Sigma = \text{diag}(\sigma_j^2)_{\{1\le j\le m\}}$,

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_j^2 \end{pmatrix}, \tag{2.41}$$

and $|\Sigma| = \prod_{j=1}^{m} \sigma_j^2$, since $|\Sigma|^{-1} = |\Sigma^{-1}|$,

$$|\Sigma^{-1}| = \begin{vmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_j^2} \end{vmatrix} = \prod_{j=1}^{m} \frac{1}{\sigma_j^2}, \tag{2.42}$$

The log-likelihood function from the previous section is the following,

$$\tilde{l_c} = \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\log(\pi_k) + \sum_{i=1}^{n}\sum_{k=1}^{K} \frac{1}{2} w_{ik}\log(|\Sigma|^{-1}) + \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{m}{2}\log(2\pi) w_{ik} + \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{1}{2} w_{ik}(x_i - \alpha - \beta z_k)^T \Sigma^{-1}(x_i - \alpha - \beta z_k),$$

and $\log(|\Sigma|^{-1})$ in the above log-likelihood function will become,

$$\log(|\Sigma|^{-1}) = \log(|\Sigma^{-1}|) = \log(\frac{1}{\sigma_1^2}.\frac{1}{\sigma_2^2}\ldots\frac{1}{\sigma_m^2}) = -2\sum_{j=1}^{m}\log\sigma_j,$$

then the log-likelihood function $\tilde{l_c}$ will become,

$$\tilde{l}_{new} = constant + \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{1}{2} w_{ik}(-2)\sum_{j=1}^{m}\log\sigma_j + \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{m} -\frac{1}{2} w_{ik}\frac{(x_{ij} - \alpha_j - \beta_j z_k)^2}{\sigma_j^2},$$

by taking partial derivative of the $\tilde{l}_{new}$ with respect to σ_j and by letting it to be zero, we obtain,

$$\sum_{i=1}^{n}\sum_{k=1}^{K} -w_{ik}\frac{1}{\sigma_j} + \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(x_{ij} - \alpha_j - \beta_j z_k)^2\sigma_j^{-3} = 0,$$

since $\sum_{i=1}^{n}\sum_{k=1}^{K} -w_{ik} = n$,

$$\frac{n}{\sigma_j} = \frac{1}{\sigma_j^3}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(x_{ij} - \alpha_j - \beta_j z_k)^2,$$

we then obtain variance parameterization (i),

$$\hat{\sigma_j}^2 = \frac{\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(x_{ij} - \hat{\alpha}_j - \hat{\beta}_j\hat{z}_k)^2}{n}, \tag{2.43}$$

which corresponds to Equation (2.14) in Section 2.3.

2.5 Identifiability

Consider again the model for the multivariate data x_i implied by Equation (2.7), i.e.

$$x_i = \alpha + \beta z_k + \varepsilon_i.$$

The product term βz_k makes the parameters $\beta = (\beta_1, \ldots, \beta_m)^T$ and z_k unidentifiable. The vector α is also unidentifiable as, when moving along the estimated straight line, the same model could be attained by translating all z_k's along the line. Therefore, the model is identifiable only under certain restrictions, and in order to fix the problem, we standardize z_k by letting

$$\mathrm{E}(\tilde{Z}) = \sum_{k=1}^{K} \pi_k z_k = 0 \tag{2.44}$$

and

$$\mathrm{Var}(\tilde{Z}) = \sum_{k=1}^{K} \pi_k z_k^2 - (\pi_k z_k)^2 = 1, \tag{2.45}$$

where $\mathrm{Var}[z_k] = \sum_{k=1}^{K} \pi_k z_k^2 - (\pi_k z_k)^2$ (Marques da Silva Júnior et al., 2018). Equation (2.44) solves the problem for α by fixing the position of z_k's along the estimated lower-dimensional subspace, and Equation (2.45) solves the scale problem for β.

However, there is still an identifiability problem in terms of the signs of $\hat{\beta}$ and $\hat{z}_k$. For example, by applying the methodology on the International Adult Literacy Survey data available in R package **mult.latent.reg** (Zhang & Einbeck, 2024b) (this application is discussed in more detail in later sections), we obtain two sets of $\hat{\beta}$ and $\hat{z}_k$, where one is $\hat{\beta}$ = (-7.696774, -9.007150, -7.647289), $\hat{z}$ = (-0.1165172, 1.1067183, -2.9250900), and another is $\hat{\beta}$ = (7.696774, 9.007150, 7.647289), $\hat{z}$ = (0.1165172, -1.1067183, 2.9250900). With $\hat{\alpha}$ being the origin on the one-dimensional line, these two sets of $\hat{z}_k$'s are symmetric with respect to the origin. Usually, for principal component analysis, the direction of sign of the first principal component scores is not important, but for our methodology, it does affect the interpretation of the ranking in a league table.

Correlation between variables can be positive, negative or even no correlation at all.

For simplicity, we now consider a 2-dimensional data set, $x_i \in \mathbb{R}^2$. For model (2.7),

$$\begin{pmatrix} x_{i1} \\ x_{i2} \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} z_k + \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \end{pmatrix}, \tag{2.46}$$

where $i = 1, ..., n$ and $k = 1, ..., K$, $\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} z_k$ is estimated from the ECM algorithm. In most applications, the components of $\hat{\beta}$ will behave according to the following two scenarios, (i) $\hat{\beta}_1 > 0$ and $\hat{\beta}_2 > 0$, (ii) $\hat{\beta}_1 < 0$ and $\hat{\beta}_2 < 0$. An example is shown in Figure 2.1, in which the direction of the latent variable is the same with the first variable.

For instance, if the response variables are educational attainment metrics, with $\hat{\beta}$ being scenario (i), large z_k's mean large achievement, while with $\hat{\beta}$ being scenario (ii), large z_k's mean small achievement. It is possible that not all $\hat{\beta}$ have the same sign. But the thing that matters is that the methodology can guarantee reproducible and unique solutions. Hence, to identify the direction of the latent variable, we enforce $\beta_1 \geq 0$ (but any other component of β could equally be chosen for this).

(We attempted to tackle the identifiability problem by incorporating the constraints 2.44 and 2.45 into the likelihood function. While this approach yielded some reasonable equations, it also complicates a problem that could otherwise be easily solved, as we mentioned above.)

2.6 Starting Values for the ECM Algorithm

Starting values can heavily influence the ability of the ECM algorithm to locate the maximum of the likelihood (see e.g. Panić et al. (2020)). In the **R** implementation of the ECM algorithm of our methodology, the following are the default starting values for parameters $\pi_k, z_k, \alpha, \beta$, and Σ_k:

$$\pi_k^{(0)} = \frac{1}{K},$$

where K is the number of components. We use random numbers from a standard normal distribution as the starting values for the mass points,

$$z_k^{(0)} \sim N(0, 1),$$

which are then re-scaled according to (2.44) and (2.45). As default starting values for the line parameters we use

$$\alpha^{(0)} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\beta^{(0)} = x_r - \alpha^{(0)},$$

where $x_r \in \mathbb{R}^m$ is a randomly selected observation. For all four variance parameterizations, we use a diagonal matrix $\Sigma^{(0)} \in \mathbb{R}^{m\times m}$, not depending on k, as the 'starting variance matrix'. Let

$$s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2},$$

where $j = 1, 2, ..., m$ and $\bar{x}_j$ is the sample mean of the j-th variable. Then, for each diagonal element $(\sigma_j^{(0)})^2$ of the diagonal matrix $\Sigma^{(0)}$, one has the starting value

$$\sigma_j^{(0)} = \frac{s_j}{K}, \;\; j = 1, \ldots, m.$$

Note that the starting values for the parameters described above are the default settings used for the simulations and applications. Additionally, we developed three other choices of starting values in the R package **mult.latent.reg** (Zhang & Einbeck, 2024b). Details of this R package can be found in Section 5.

2.7 Inclusion of Covariates

As briefly introduced in the introduction section, we explore extending the one-level model (2.1) to incorporate covariates, providing another perspective for analyzing multivariate data. Where multivariate response data appear in statistical applications, the most common inferential approach is to define separate regression models for each of the individual variables constituting the multivariate response vector. For instance, while the linear model function `lm` in the statistical programming language **R** does allow for a multivariate response, the resulting fitted models correspond exactly to the individual one-dimensional response models. This approach, however, is ignoring the correlation of the different response variables, which, when taken into account, could lead to reduced parameter standard errors, and hence increased power.

In the original model (2.1), $x_i \in \mathbb{R}^m$ can be explained by a one-dimensional coordinate system. Under the mixture representation of the model (2.7), certain latent groups along the one-dimensional line are driving the data generating process. However, these models do not yet allow for the presence of covariates in the data-generating process of the x_i. To avoid confounding of the latent variable with such covariates (if they are known), the following is an extended model which includes a vector of p covariates related to the response variables,

$$x_i = \alpha + \beta z_i + \Gamma v_i + \varepsilon_i, \tag{2.47}$$

where $x_i \in R^m$, $i = 1, 2, ..., n$, $\alpha \in \mathbb{R}^m$, $\beta \in \mathbb{R}^m$, $v_i \in \mathbb{R}^p$ is the vector of the covariates, and $\Gamma_{m\times p}$ is a matrix of the coefficients of the covariates. Similarly, the 'approximative' model can be written as

$$x_i|z_k, \alpha, \beta, \Gamma \sim N(\alpha + \beta z_k + \Gamma v_i, \Sigma_k) \text{ with probability } \pi_k. \tag{2.48}$$

When we have only one covariate in model (2.47), $v_i \in \mathbb{R}$ and we denote $\Gamma = \gamma \in \mathbb{R}^m$.

The ECM algorithm takes similar shapes as before, with E-step given by $w_{ik} = \frac{\pi_k f_{ik}}{\sum_l \pi_l f_{il}}$. The log-likelihood for model (2.48) is just a simple adaptation of Equation (2.9) with an inclusion of a Γv_i term. By taking partial derivatives of the log-likelihood with respect to each parameter we obtain the score functions, and by equaling these score function to zero and solving them, and again applying the computational simplification as in Section 2.3, we obtain the following estimators, the derivations of these estimators can be found in Appendix A.

$$\hat{\alpha} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i - \hat{\beta}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k - \sum_{i=1}^{n}\hat{\Gamma}v_i\right),$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k x_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k) - \sum_{i=1}^{n}\hat{\Gamma}v_i \sum_{k=1}^{K} w_{ik}\hat{z}_k}{\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k^2 - \frac{1}{n}(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k)^2} + \frac{\frac{1}{n}(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k)(\sum_{i=1}^{n}\hat{\Gamma}v_i)}{\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k^2 - \frac{1}{n}(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k)^2}$$

$$\hat{z}_k = \frac{\hat{\beta}^T \sum_{i=1}^n w_{ik}(x_i - \hat{\alpha} - \hat{\Gamma} v_i)}{\hat{\beta}^T \hat{\beta} \sum_{i=1}^n w_{ik}},$$

$$\hat{\Gamma} = \frac{\sum_{i=1}^n x_i v_i^T - \hat{\alpha} \sum_{i=1}^n v_i^T - \hat{\beta} \sum_{i=1}^n v_i^T \sum_{k=1}^K w_{ik} \hat{z}_k}{\sum_{i=1}^n v_i v_i^T}.$$

$$\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^n w_{ik}.$$

We let $\psi_i = (\psi_{i1}, \ldots, \psi_{im})^T = \Gamma v_i \in \mathbb{R}^m$. Estimators for the flexible variance parameterizations are given as the following,

(i) $\Sigma = \text{diag}(\sigma_j^2)_{\{1 \le j \le m\}} \in \mathbb{R}^{m \times m}, k = 1, ..., K,$

$$\hat{\sigma}_j^2 = \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K w_{ik}(x_{ij} - \hat{\alpha}_j - \hat{\beta}_j \hat{z}_k - \psi_{ij})^2$$

(ii) $\Sigma_k = \text{diag}(\sigma_{jk}^2)_{\{1 \le j \le m\}} \in \mathbb{R}^{m \times m}, k = 1, ..., K,$

$$\hat{\sigma}_{jk}^2 = \frac{\sum_{i=1}^n w_{ik}(x_{ij} - \hat{\alpha}_j - \hat{\beta}_j \hat{z}_k - \psi_{ij})^2}{\sum_{i=1}^n w_{ik}}$$

(iii) $\Sigma = \Sigma_1 = ... = \Sigma_k \in \mathbb{R}^{m \times m}$, $k = 1, ..., K,$

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K w_{ik}(x_i - \hat{\alpha} - \hat{\beta} \hat{z}_k - \hat{\Gamma} v_i)(x_i - \hat{\alpha} - \hat{\beta} \hat{z}_k - \hat{\Gamma} v_i)^T$$

(iv) $\Sigma_k \in \mathbb{R}^{m\times m}, k = 1, ..., K$,

$$\hat{\Sigma}_k = \frac{\sum_{i=1}^n w_{ik}(x_i - \hat{\alpha} - \hat{\beta}\hat{z}_k - \hat{\Gamma}v_i)(x_i - \hat{\alpha} - \hat{\beta}\hat{z}_k - \hat{\Gamma}v_i)^T}{\sum_{i=1}^n w_{ik}}$$

Notably, under model (2.47) with $x_i \in \mathbb{R}^m$, the 'models' for each of the m response variables would be linked through the random effect z_i, hence inducing correlation between units similar as for a multilevel model. An example for the use of this modelling technique is provided in Section 2.10.4.

2.8 Simulations

In this section, we will conduct simulations to assess the accuracy of parameter estimation for a one-level model without covariates, evaluate model selection accuracy, test parameter estimation accuracy for a one-level model with covariates, and use a simple bootstrapped way to obtain standard errors.

2.8.1 Parameter estimation accuracy

The ECM algorithm derived in the previous section, with all four variance parameterizations, is implemented in **R**. Some simulations are set up to test the accuracy of the **R** implementation under different settings.

Under the variance parameterization (i), i.e. the same diagonal matrix for all components, we use 2-dimensional data with three individual sample sizes $n = 100$, $n = 300$, and $n = 500$, and generate 1000 data sets from model (2.7) for each sample size. The true parameter values used for the simulations can be read from the first column of Table 2.1.

The methodology from subsection 2.3 is then applied on each generated data set (with random starting values according to Section 2.6 to initialize the ECM algorithm), and the 1000 estimates for each model parameter (see Table 2.4) are collected. Comparing the average of the estimated values to the true values of the parameters used to generate these data, some key results are shown in Table 2.1, Figure 2.2, Figure 2.3 and Figure 2.4. In Table 2.1, the

Table 2.1: Simulation results under variance parameterization (i). Note that the true values listed for z_k are standardized. The true values before standardization were $z_1 = -0.1$, $z_2 = 13$ and $z_3 = 25$.

		Average estimates		
	True	$n = 100$	$n = 300$	$n = 500$
π_1	0.0500	0.0463	0.0507	0.0498
π_2	0.2500	0.2518	0.2504	0.2512
π_3	0.7000	0.7019	0.6988	0.6990
z_1	-0.6171	-0.6186	-0.6193	-0.6191
z_2	1.1675	1.2262	1.1693	1.1708
z_3	2.8023	2.8457	2.8130	2.8119
α_1	-1.000	-0.9936	-0.9985	-0.9985
α_2	1.000	1.0235	1.0036	0.9982
β_1	1.000	0.9915	0.9986	0.9966
β_2	3.000	2.9974	2.9982	2.9899
σ_1	0.5000	0.5043	0.4966	0.4985
σ_2	2.0000	1.9866	1.9892	1.9912

averaged estimates of the parameters are close to their true values across all parameters and sample sizes, with the bias in the estimates reducing for larger sample sizes. In Figure 2.2, the medians of the estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ in the three box plots are similar, but with the ranges of the boxes getting smaller when increasing n from 100 via 300 to 500. The effect is clearer visible for the $\hat{\beta}_2$'s than the $\hat{\beta}_1$'s since the larger magnitude of the true value of β_2 also comes with larger variability.

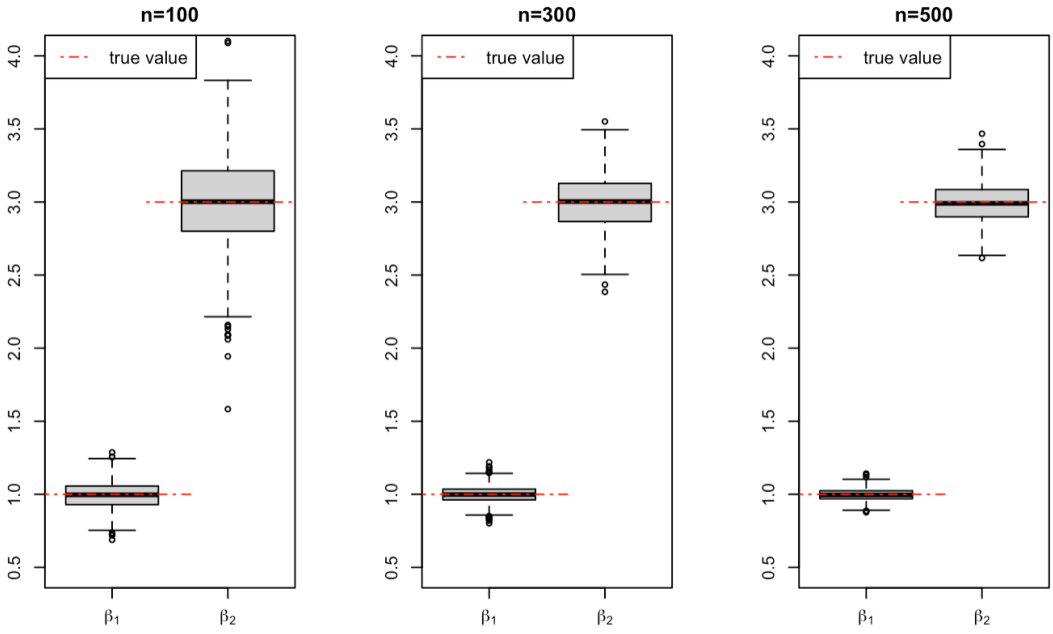

Figure 2.2: Estimations of parameter $\beta = (\beta_1, \beta_2)^T$ with different sample sizes under the variance parameterization (i).

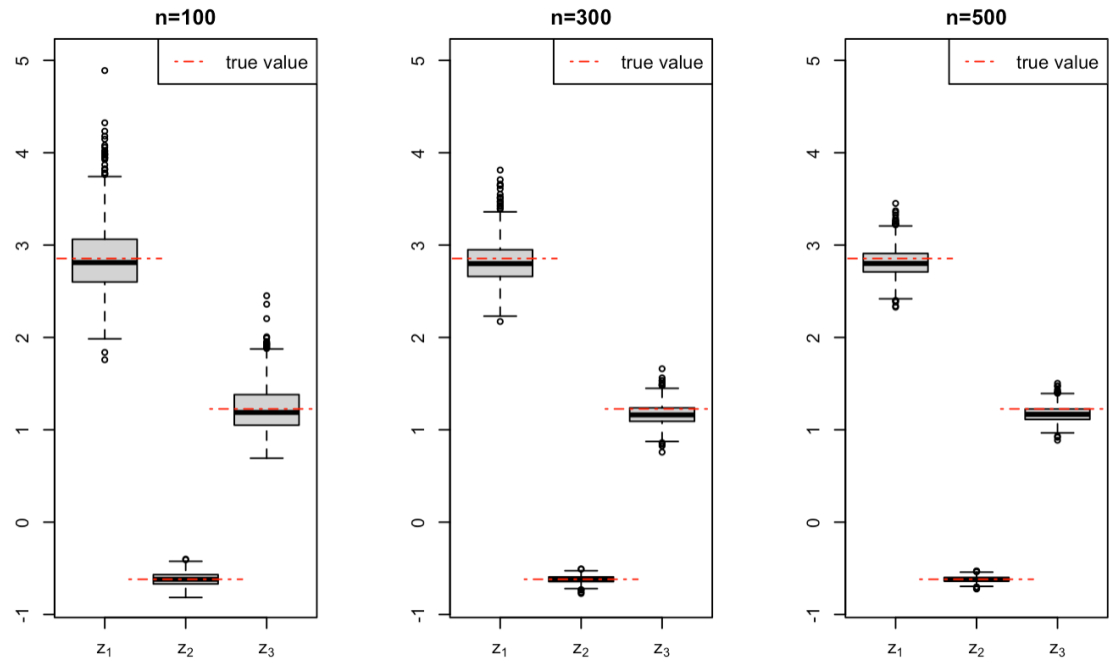

Figure 2.3: Estimations of parameter z_k with different sample sizes under the variance parameterization (i).

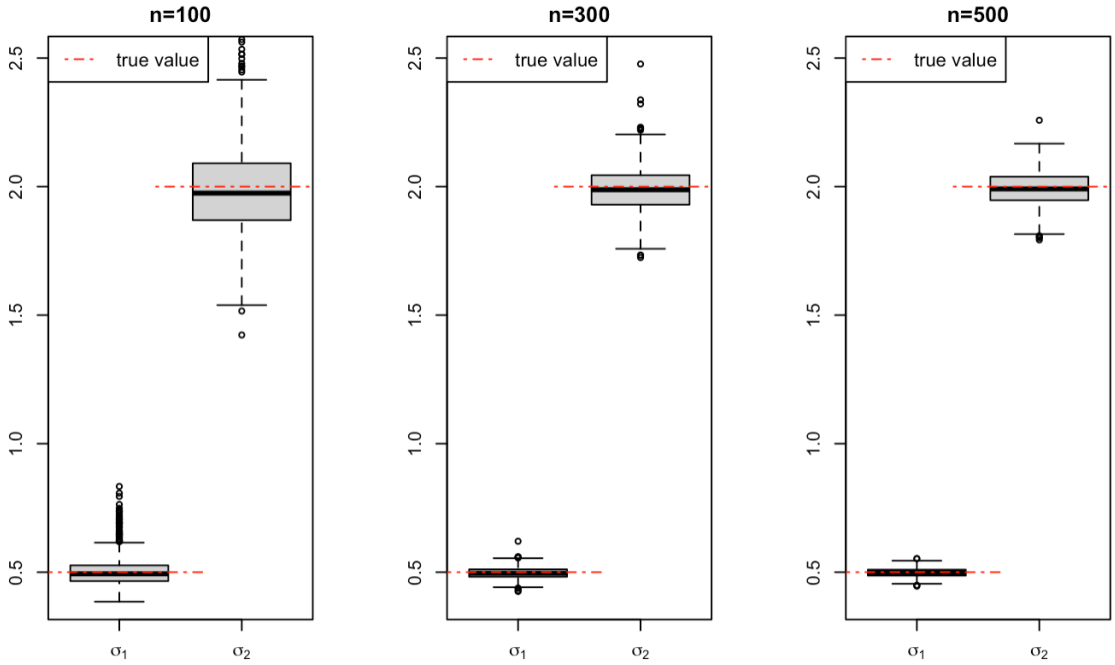

Figure 2.4: Estimations of parameter σ with different sample sizes, where σ_1 and σ_2 are the diagonal components of the variance matrix, under the variance parameterization (i).

Similar simulations were conducted to test the accuracy under variance parameterization (ii), again using 1000 replicates of 2-dimensional data from model (2.7) under each of three sample sizes of $n = 100$, $n = 300$, and $n = 500$. We report the numerical results in Table

Table 2.2: Simulation results under variance parameterization (ii), where σ_{11} and σ_{21} are the diagonal elements of Σ_1, σ_{12} and σ_{22} are the diagonal elements of Σ_2. Note that the true values listed for z_k are standardized. The true values before standardization were $z_1 = 5$ and $z_2 = 25$.

		Average estimates		
	True	$n = 100$	$n = 300$	$n = 500$
π_1	0.2000	0.2004	0.2001	0.2002
π_2	0.8000	0.7996	0.7999	0.7998
z_1	-0.5000	-0.5263	-0.5168	-0.4999
z_2	2.0000	2.0293	2.0182	2.0248
α_1	2.0000	2.0119	2.0024	2.0016
α_2	10.0000	10.0045	9.9998	9.9995
β_1	1.000	0.9929	0.9948	0.9955
β_2	3.000	2.9771	2.9871	2.9926
σ_{11}	0.2000	0.1972	0.1993	0.1998
σ_{21}	0.4000	0.3949	0.3971	0.3991
σ_{12}	1.0000	0.9614	0.9856	0.9948
σ_{22}	2.0000	1.9465	1.9880	1.9862

2.2 and display the estimated variance structures under this model in Figure 2.5. We omit the boxplots for the other parameters as they are similar to those under parameterization (i). For variance parameterization (iii), We generate 2-dimensional data from model (2.7) under three sample sizes of $n = 100$, $n = 300$, and $n = 500$, with 200 replicates. The results and boxplots under parameterization (iii) are shown at the end of this subsection, the main results are summarized in Table 2.3, Figures 2.6, 2.7, and 2.8.

Overall, we can tell from the tables and figures that the estimators give sensible estimates of the parameters, the averaged estimates of the parameters are accurate compared to their true values, there appear to be no systematic biases, and the variability of the estimates reduces with increased sample size. The boxplots illustrate the consistency of estimators, where the boxes are squeezing to the true value (red horizontal line) as the sample size gets larger.

2.8.2 Model selection accuracy (AIC and BIC)

Next, we set up another set of simulations to address the importance of using the correct variance parameterization when fitting a model. For each model with each variance parameterization, we generate 200 replicates, each with sample size of 100, from the model. Then for the data generated from the model with variance parameterization (i), we fit the data to four different models, each with a different variance parameterization. For the remaining data sets generated from the model with variance parameterization (ii), (iii), and (iv), we do the same.

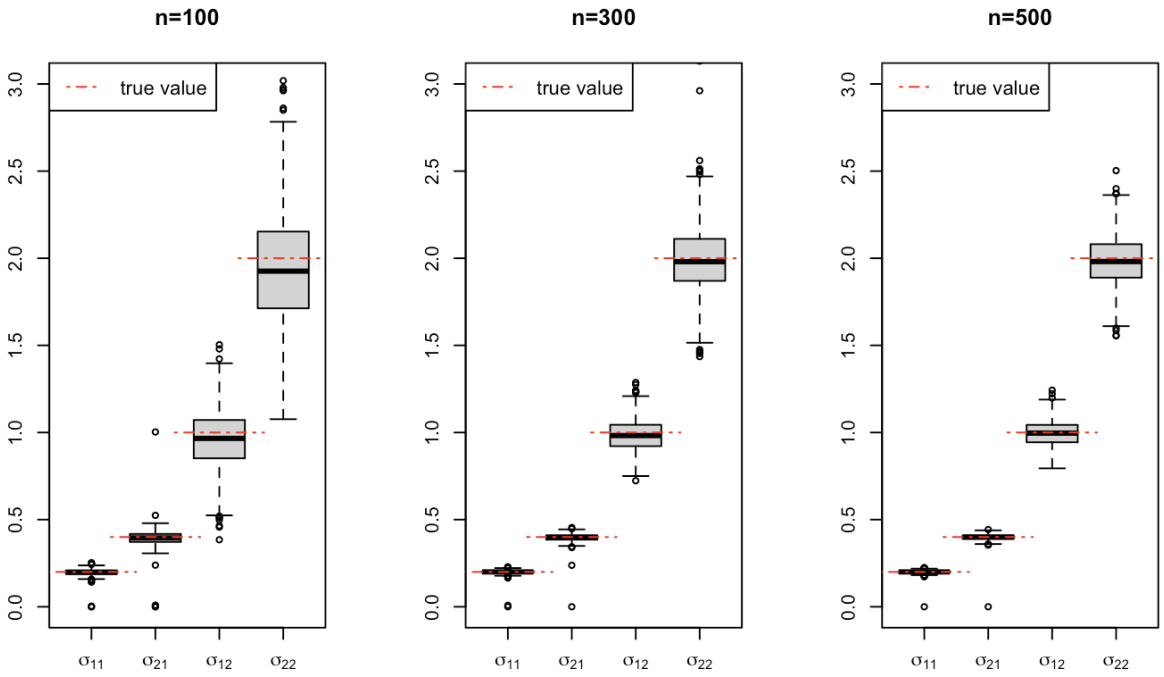

Figure 2.5: Under variance parameterization (ii), estimations of variance parameters with different sample sizes, where σ_{11} and σ_{21} are the diagonal components of the variance matrix for mass point $k = 1$, σ_{12} and σ_{22} are the diagonal components of the variance matrix for mass point $k = 2$.

We consider to use the AIC and BIC (Schwarz, 1978) as the model selection criteria, and we use the approximated likelihood (2.4) as the likelihood in AIC and BIC. For reference, the number of estimated parameters used in the calculation of AIC and BIC are shown in Table 2.4, where m is the dimension of data and K is the number of mass points.

Figure 2.9 shows some key results: for the datasets generated from the model with variance parameterization (i), 73.5% of the fitted models with variance parameterization (i) lead to the smallest AIC values, and 95% of the fitted models with variance parameterization

Table 2.3: Simulation results under variance parameterization (iii). Note that the true values listed for z_k are standardized. The true values before standardization were $z_1 = 5$ and $z_2 = 25$.

		Average estimates		
	True	$n = 100$	$n = 300$	$n = 500$
π_1	0.4000	0.4007	0.3983	0.3971
π_2	0.6000	0.5993	0.6017	0.6028
z_1	-0.8165	-0.8215	-0.8145	-0.8121
z_2	1.2247	1.2296	1.2320	1.2340
α_1	20.0000	20.0001	19.9965	19.9891
α_2	7.0000	7.0174	6.9780	6.9828
β_1	1.000	1.0065	0.9967	0.9980
β_2	3.000	2.9749	2.9864	2.9901

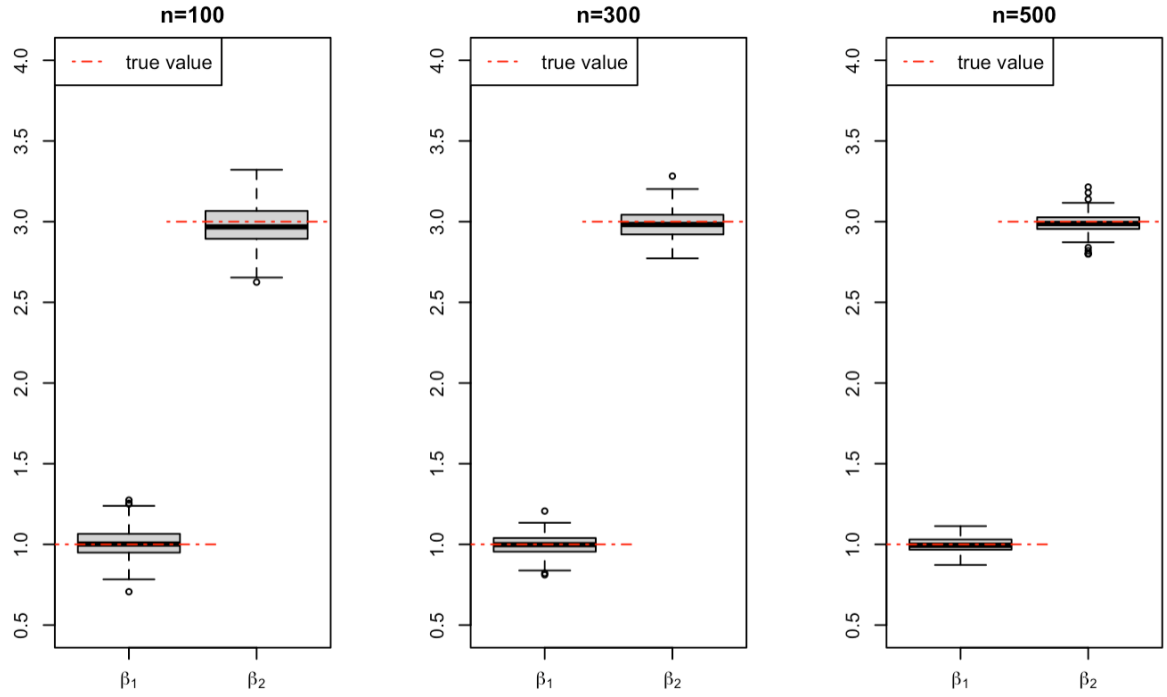

Figure 2.6: Under variance parameterization (iii), estimations of parameter β with different sample sizes.

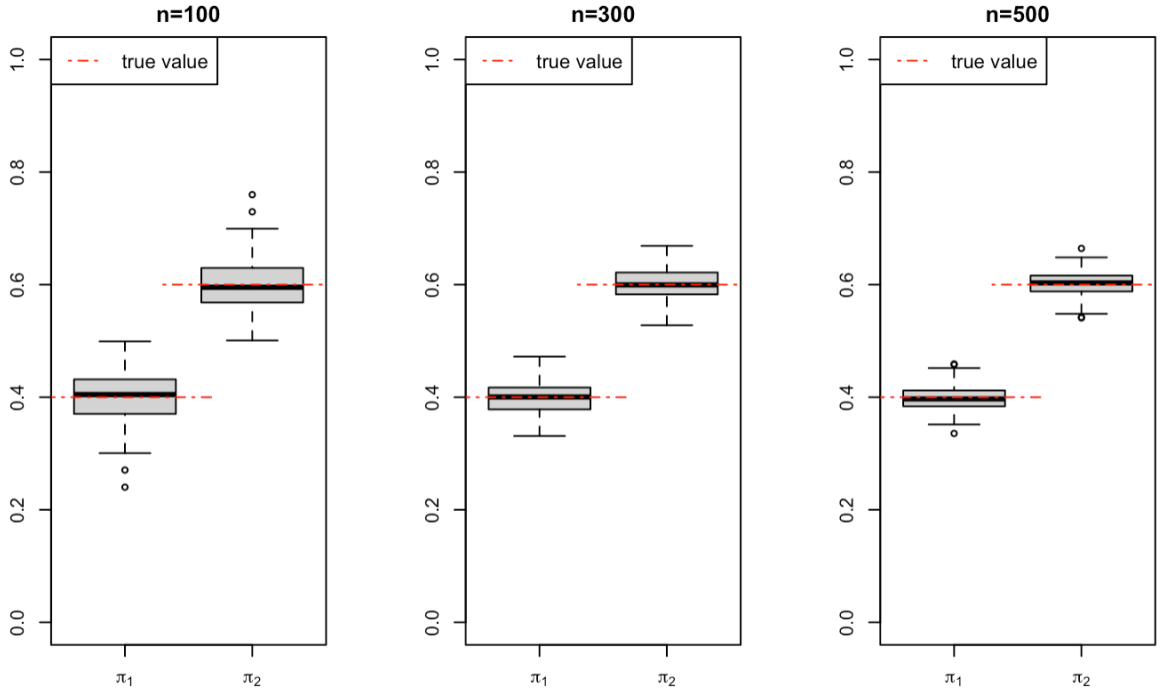

Figure 2.7: Under variance parameterization (iii), estimations of parameter π_k with different sample sizes.

(i) lead to the smallest BIC values. For the datasets generated from the model with variance parameterization (ii), 88% of the fitted models with variance parameterization (ii) lead to the smallest AIC values, and 98% of the fitted models with variance parameterization (ii) lead to the smallest BIC values. For the datasets generated from the model with variance parameterization

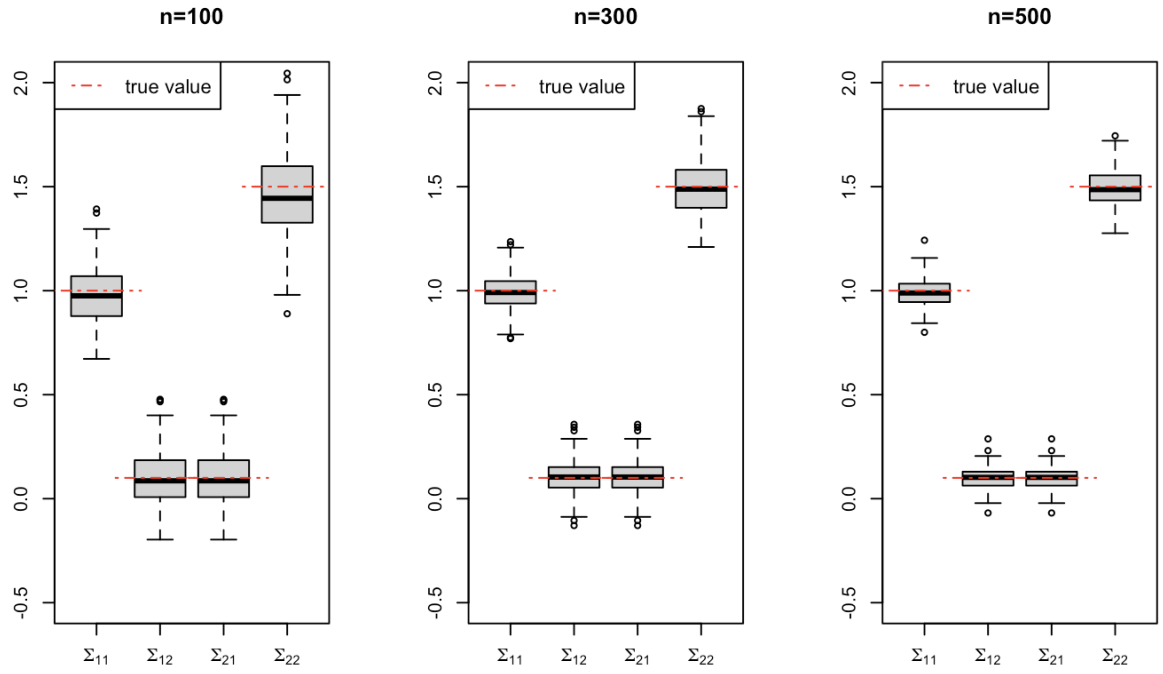

Figure 2.8: Under variance parameterization (iii), estimations of parameter Σ with different sample sizes, where Σ_{11}, Σ_{22} are the diagonal elements and Σ_{12}, Σ_{21} are the off diagonal elements of the variance matrix. The true values are: $\Sigma_{11} = 1.0$, $\Sigma_{22} = 1.5$, $\Sigma_{12} = \Sigma_{21} = 0.1$.

Table 2.4: The number of estimated parameters used for AIC and BIC.

Parameters	Variance (i)	Variance (ii)	Variance (iii)	Variance (iv)
π_k	$K-1$	$K-1$	$K-1$	$K-1$
z_k	K	K	K	K
α	m	m	m	m
β	m	m	m	m
Σ_k	m	mK	$\frac{m(m+1)}{2}$	$\frac{m(m+1)K}{2}$

(iii), 87% of the minimum AIC values and 99% of the minimum BIC values are obtained from a fitted model with the variance parameterization (iii). For datasets generated from the model with variance parameterization (iv), fitting the model with variance parameterization (iv), we obtain 99% of the minimum AIC values and 91.5% of the minimum BIC values. The results indicate that choosing a correct variance parameterization is significant for fitted model selection.

Aitkin et al. (2009) noted that, due to failing the large-sample normality assumption for the likelihood function, the use of AIC or BIC in the context of mixture models may be considered questionable (despite being commonly used). Almohaimeed and Einbeck (2022) discussed further the use of AIC and BIC for model selection under NPML estimation. Although

the BIC might lead to a different choice than AIC, Leroux (1992) showed that using BIC for selecting the number of mixture components for finite mixture models is consistent. We use AIC and BIC as model selection criteria in our methodology.

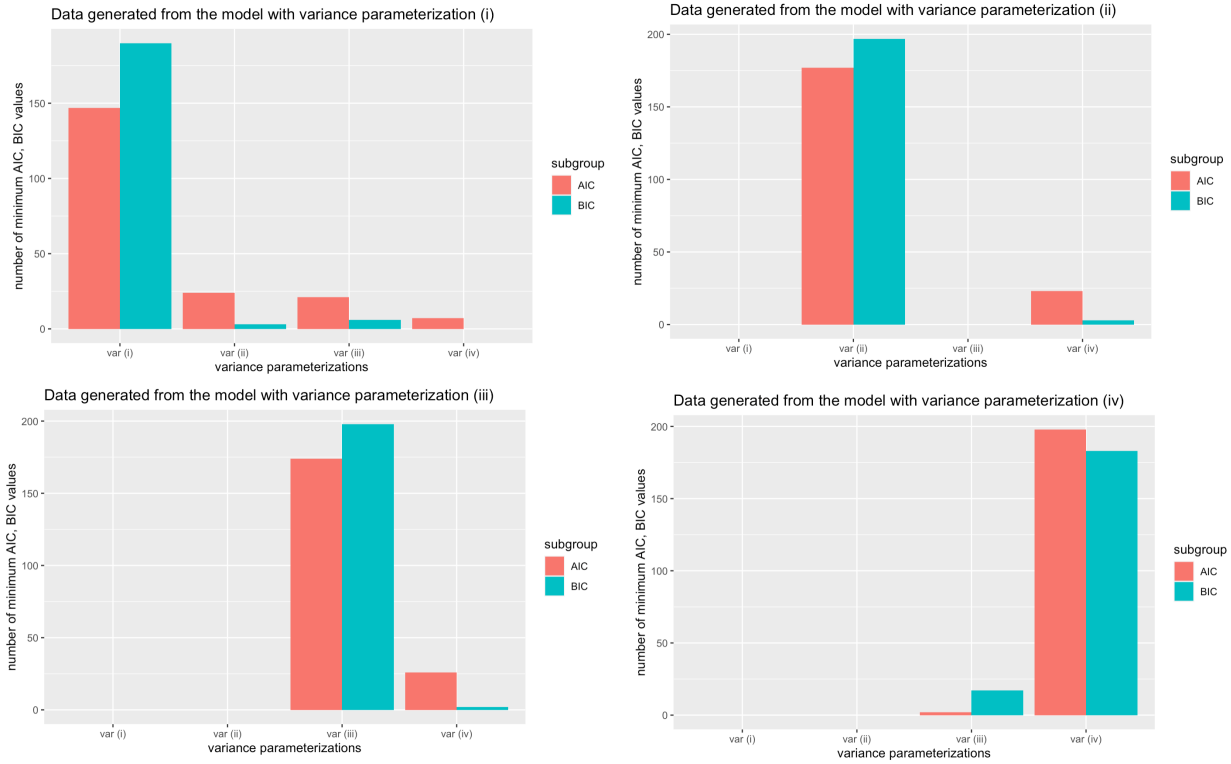

Figure 2.9: Barplots showing the number of minimum AIC, BIC values obtained from fitted models with different variance parameterizations.

2.8.3 Parameter Estimation Accuracy for One-level model with Covariates

For the one-level model with covariates, we conducted a simulation to test the accuracy of parameter estimation using the ECM algorithm implemented in R. Due to the main application of this model being to fit a multivariate response model, the key focus would be on the accuracy of the estimates of parameter Γ, which can be considered as the coefficient matrix. We simulate 4-dimensional data from model (2.48) with two covariates generated from a uniform distribution with a lower limit of 0 and an upper limit of 1, and with $K = 2$ mixture components, under variance parameterization (i). We generate 300 replicates for each of the three sample sizes: $n = 100$, $n = 300$ and $n = 500$, Figure 2.10 shows the structure of data that we are using. Table 2.5 shows the averaged estimates for parameter Γ, the true values used can be found

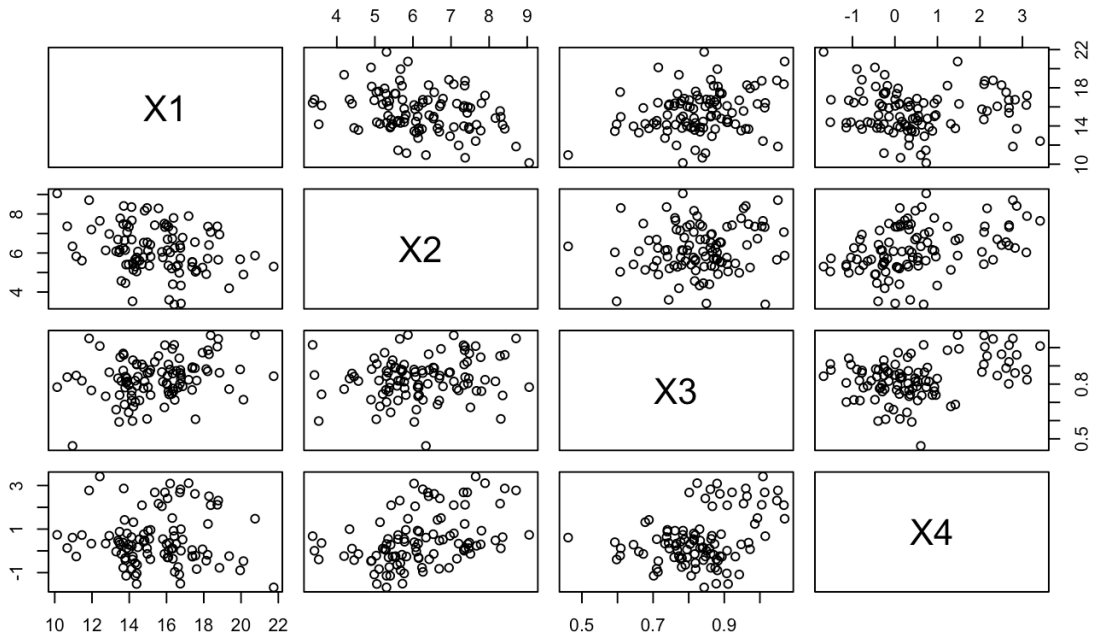

Figure 2.10: A simulated data set with sample size of $n = 100$.

in the first column. Figure 2.11 and 2.12 shows the histograms for each elements in the Γ matrix for all sample sizes each with 300 replicates, we observe that they are all normally distributed and the averaged estimates are very close to their true values across all parameters and sample sizes. However, for parameter β, as seen in Figure 2.13, the estimates are non-symmetrically bimodally distributed. This indicates that the ECM algorithm found a different set of solutions where the parameter is poorly estimated, and this situation corresponds to poor starting points. The reason for this is that the model from which the data points are simulated is a complicated model (4-variate, 2-dimensional, and with 2 covariates). One way to improve the estimation results (specifically for parameters that are not our main concern) is to run each replicate several times (let's say 10 times) and select the results out of the 10 with the smallest AIC values. Due to the computational expense of this simulation, we will illustrate this idea, later in this dissertation, using a small number of replicates. Details can be found in section 5.2 when we introduce the functionalities of the associated R package.

2.8.4 Bootstrapped Standard Error for One-level Model

We conducted a simple simulation following the bootstrapped standard error algorithm which will be introduced in details in Section 2.9.4 with a two-dimensional data set to compare the estimates and the standard errors with fitting separate linear models using `lm()` function. We

Table 2.5: Simulation results for parameter Γ under variance parameterization (i).

		Average estimates		
	True	$n = 100$	$n = 300$	$n = 500$
Γ_{11}	-0.5200	-0.5043	-0.5723	-0.5105
Γ_{12}	0.0500	0.0448	0.0496	0.0587
Γ_{13}	-0.0400	-0.0383	-0.0402	-0.0397
Γ_{14}	0.7200	0.7184	0.7163	0.7213
Γ_{21}	-4.7900	-4.8399	-4.7966	-4.7844
Γ_{22}	2.1900	2.1846	2.1770	2.2029
Γ_{23}	-0.1400	-0.1451	-0.1417	-0.1394
Γ_{24}	0.8700	0.8385	0.8664	0.8727

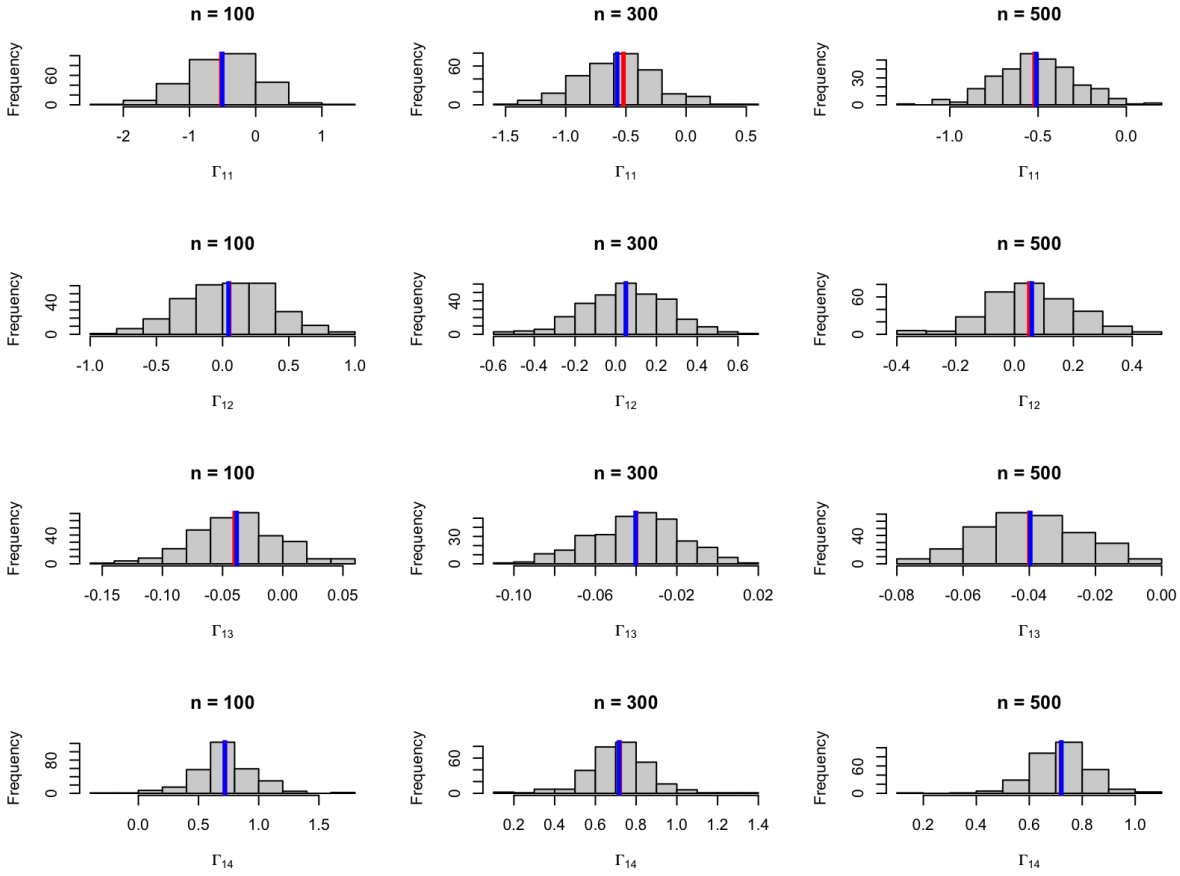

Figure 2.11: A histogram illustrating the overall estimates of 300 replicates for all three sample sizes is shown for the first column of the Γ matrix. The vertical red line is the true value and the blue line is the mean.

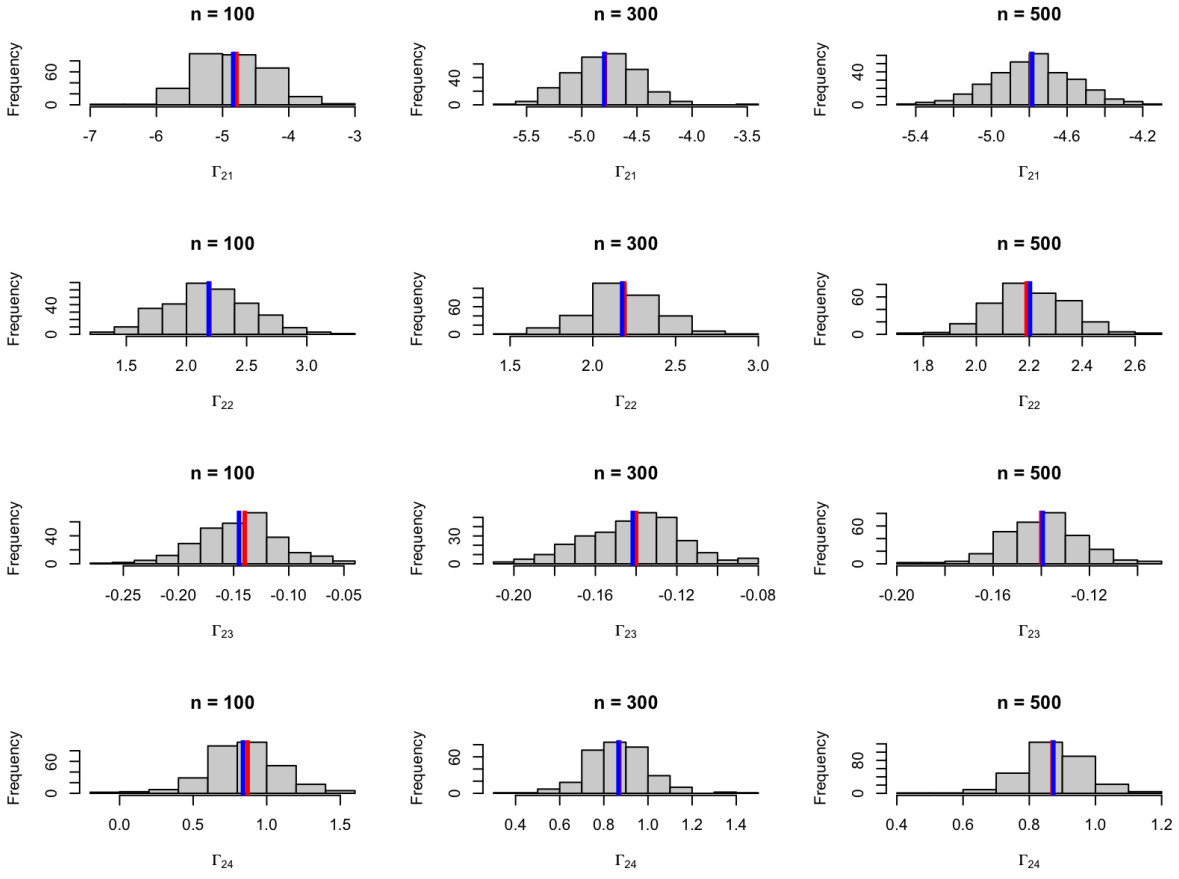

Figure 2.12: A histogram illustrating the overall estimates of 300 replicates for all three sample sizes is shown for the second column of the Γ matrix. The vertical red line is the true value and the blue line is the mean.

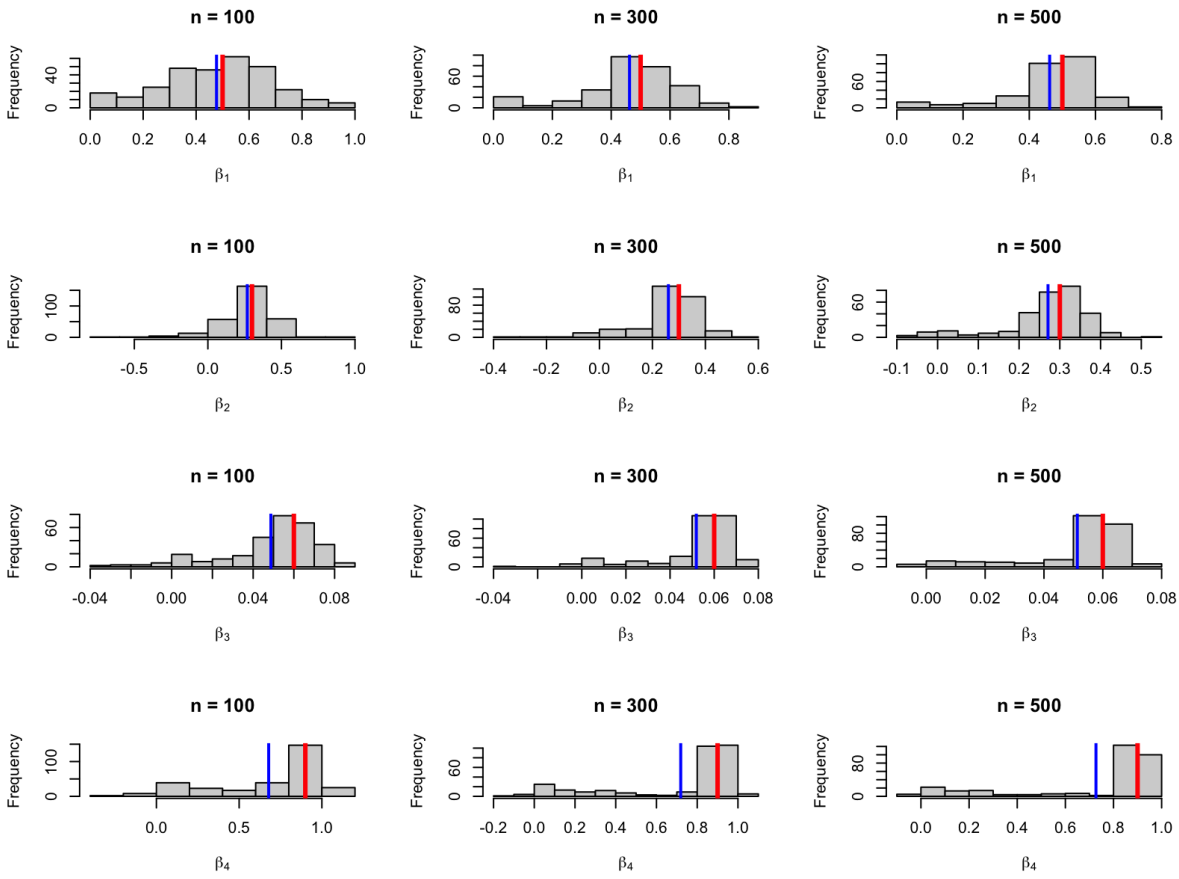

Figure 2.13: A histogram illustrating the overall estimates of 300 replicates for all three sample sizes is shown for β. The vertical red line is the true value and the blue line is the mean.

generated a two-dimensional data set $x_i \in \mathbb{R}^2$ with $\pi = (0.3, 0.7)$, $z = (1.5, -0.6)$, $\alpha = (10, 2)$, $\beta = (1, 3)$, $\gamma = (0.5, 3)$. Then we compared the estimates and standard errors obtained from the use of R function `lm` (when used as a multivariate response model), with those obtained from the procedures outlined in the previous and current subsection with $B = 300$. The results are shown in Table 2.6; overall, our model leads to considerably smaller standard errors for the estimated coefficient parameter γ.

Table 2.6: Estimations (Standard errors) of γ obtained using different methods.

	$\hat{\gamma}_1$ (SE)	$\hat{\gamma}_2$ (SE)
`lm(.)`	0.5025 (0.2139)	2.8763 (0.4871)
model (2.47) with bootstrap	0.4649 (0.1709)	2.7710 (0.3201)

2.9 Additional Inferential Aspects

In the previous sections, for the proposed one-level models (both with and without covariates), the focus was on estimating the parameters from multivariate data $x_i \in \mathbb{R}^m$, and demonstrating that these estimators are (in an empirical sense) consistent, and the variance parameterizations are identifiable. In practice, these steps will rarely form an end in itself, but will be building blocks on the way to a more concrete statistical question. We now refer back to the four application pillars already mentioned in the introduction (Section 2.1), and explain these one by one. Additionally, we will address the important question of how bootstrapped standard errors of covariate parameter estimates are obtained, and how these fare in comparison to univariate response models.

2.9.1 Clustering via MAP estimation

We have already observed in Section 2.3 that the weights w_{ik} correspond to the posterior probability of observation i belonging to component k. The term 'posterior' is here be to be understood as the updated probability of class membership, having knowledge on the value of the observation x_i, as opposed to the 'prior' probability π_k, which does not make use of this information.

Given the availability of w_{ik} from the last iteration of the ECM algorithm, observation x_i is then classified to the cluster $\hat{k}(x_i)$ to which it belongs with highest posterior probability,

$$\hat{k}(x_i) = \arg\max_k w_{ik}.$$

This cluster allocation rule is commonly known as Maximum a posterior (MAP) rule. It is noted in this context that, after convergence of the ECM algorithm, typically most w_{ik} are close to 0 or 1 (with obviously only one of them being close to 1), so that this allocation is in most cases very clear-cut. We will see examples for the application of the MAP rule in Sections 2.10.1 and 2.10.3.

2.9.2 Dimension reduction through predicted latent scores

One application of our methodology is the compression of m-dimensional data to one-dimensional, model-based scores, which can be considered as the summary information of the original data. This is achieved through the use of the 'projection'

$$z_i^* = \sum_{k=1}^{K} w_{ik}\hat{z}_k, \tag{2.49}$$

where $z_i^* \in \mathbb{R}$ (Aitkin, 1996a). Given the fitted model (2.1), z_i^* would be the best prediction of the position for the latent variable z_i that generates the original data x_i. Then the following equation maps the one-dimensional scores back onto the higher dimensional original data space,

$$x_i^* = \alpha + \beta z_i^*,$$

where x_i^* are the compressed counterparts to the original data. It is clear that, unlike in e.g. principal component analysis, the projections $x_i - x_i^*$ are not orthogonal to the linear subspace. However, they still can be meaningful: Under the given approach, all differences between observations to their cluster centres are treated as actual *errors*. The result of this is an increased robustness to such errors, as only clear deviations from a cluster lead to a projection beyond its centre. An example illustrating this behavior is provided in Figure 2.16 in Section 2.10.1.

The one-dimensional scores, z_i^*, can then be used for subsequent inferential procedures, such as a predictor variable in a regression problem involving an external response variable y_i. This approach is illustrated by way of example in Section 2.10.2.

2.9.3 Ranking

The projected z_i^* provide a 'summary score' of all involved variables in the direction spanned by the latent line. Along this line, the positioning of the z_i^* is informative for the degree of which the variables jointly point into the direction of the latent variable. That is, high values of z_i^* would indicate overall high values of the contributing variables, and good agreement of what constitutes 'high'. For instance, if each of three variables constitute price indexes for certain goods, then the higher these constituent indices are, the higher the overall price index will be. Hence, the order statstic of the z_i^*, denoted by $z_{[i]}^*$, can be used to rank the cases i, namely by $[i]$, $i = 1, \ldots, n$. Many of these order statistics will be undistinguishable as the projections will be on (or close) to the same cluster centre. This makes sense from a clustering point of view: If observations cannot be distinguished statistically (i.e., if they are just distinguished by noise), their rank cannot be distinguished. De facto, in many cases, the $z_{[i]}$ will take as many *distinguishable* values as there are mass points. This concept will be explained in more detail by means of an example in Section 2.10.3.

2.9.4 Bootstrapped standard errors and p-values

In statistical practice, not only the estimation of Γ but also an assessment of its accuracy (or in other words, a quantification of its uncertainty) are of interest. Since the direct calculation of standard errors is generally difficult in the context of ECM estimation, we propose a bootstrap procedure for their computation.

The bootstrap process is carried out with the following steps:

(i) We are given a data set $x_i \in \mathbb{R}^m$ and a covariate vector $v_i \in \mathbb{R}^p$, $i = 1, \ldots, n$.

(ii) Fit the data x_i, v_i to model (2.48) to obtain the estimates of the parameters.

(iii) Sampling B data sets from model (2.48) with the estimated parameters obtained from (ii). (Note that in model 2.48, the random effect z_i is replaced by K mass points z_k,

which are randomly chosen with probability π_k.)

(iv) Fit these B data sets to our model and we would obtain B sets of $\hat{\gamma}$. Then calculate the standard deviations across all B replicates of each of the $m \times p$ components of $\hat{\Gamma}$.

The bootstrap process to obtain the p-values is carried out with the following steps: The null hypothesis and the alternative hypothesis for $\hat{\Gamma}$ are the following:

$$H_o : \hat{\Gamma} = \mathbf{0}_{m \times p} \ vs \ H_1 : \hat{\Gamma} \neq \mathbf{0}_{m \times p}$$

(i) We are given a data set $x_i \in \mathbb{R}^m$ and a covariate vector $v_i \in \mathbb{R}^p$, $i = 1, \ldots, n$.

(ii) Fit the data x_i, v_i to model (2.48) to obtain the estimates of the parameters.

(iii) Sampling B data sets from model (2.48) with the estimated parameters obtained from (ii) and with $\Gamma = \mathbf{0}_{m \times p}$.

(iv) Fit these B data sets to model (2.48) and we would obtain B sets of $\hat{\Gamma}$. Then the p-value for each $\gamma \in \Gamma$ would be the proportion of $\hat{\gamma}$ smaller and larger than $\pm\hat{\gamma}_{est.}$ obtained in step (ii).

A simulation study to obtain the bootstrapped standard errors and compare them to the ones obtained from individual linear regression models (using the `lm()` function) can be found in Section 2.8.4. Examples of bootstrapped standard errors and bootstrapped p-values can be found in Section 2.10.4.

2.10 Applications

2.10.1 Faithful Data: Model Selection and Projection

In Section (2.2), we introduced four different variance parameterizations; here we use again the faithful data set to illustrate the effect of using these different variance specifications on model fitting. Figure 2.14 shows the density contour plots for fitting the model with flexible variance parameterizations (i) to (iv). As can be seen from Table 2.7, the AIC and BIC values decrease

Table 2.7: AIC, BIC values for the faithful data under different variance parameterizations.

	Variance (i)	Variance (ii)	Variance (iii)	Variance (iv)
AIC	2333.36	2317.61	2300.37	2286.53
BIC	2365.81	2357.28	2336.43	2333.40

when increasing the complexity of the variance parameterization, even though of course this does not need to be the case generally.

The following are the parameter estimates from a fitted model with the selected parameterization (iv), i.e. different full variance-covariance matrices for each component: $\hat{\pi} = (0.36, 0.64)$, $\hat{\alpha} = (3.49, 70.90)$, $\hat{\beta} = (1.08, 12.20)$, $\hat{z} = (-1.35, 0.74)$, and

$$\hat{\Sigma}_1 = \begin{bmatrix} 0.07 & 0.44 \\ 0.44 & 33.70 \end{bmatrix}, \quad \hat{\Sigma}_2 = \begin{bmatrix} 0.17 & 0.94 \\ 0.94 & 36.05 \end{bmatrix}.$$

Figure 2.15 shows the clustering resulting from these estimates, according to the cluster allocation process that is described in Section 2.9.1.

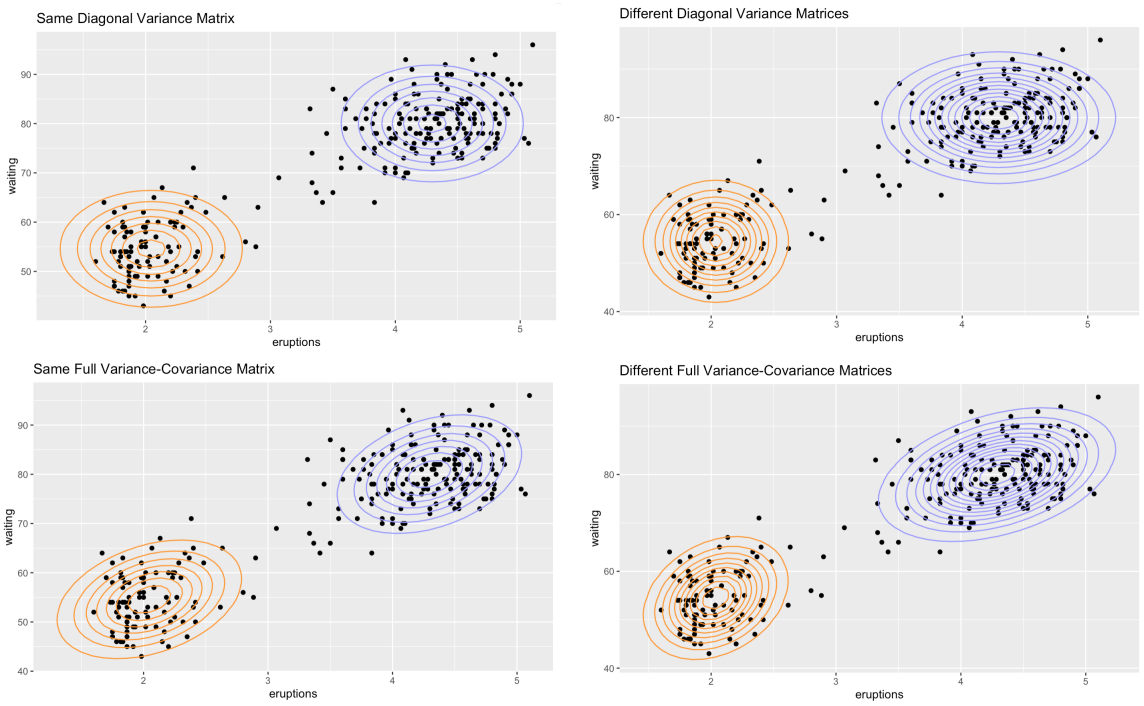

Figure 2.14: Density contour plots with different variance parameterizations; top left (i); top right (ii); bottom left (iii); bottom right (iv).

We can obtain the scores (coordinates of the projected data along the one-dimensional subspace spanned by the latent variable) through the use of Equation (2.49). We use the fol-

lowing images to illustrate the process of projecting the original data points onto the estimated low-dimensional space. In Figure 2.1, the straight line is the one-dimensional latent space, and the red triangles positioned along the straight line are the estimated mixture centres $\hat{\alpha} + \hat{\beta}\hat{z}_k$. Figure 2.15 illustrates how the original data is assigned to different clusters following the MAP rule. The green points in Figure 2.16 on the straight line are the compressed data, x_i^*, after projection onto that line. The most distinctive character between our methodology and the principal component analysis is that the projections are not orthogonal, which can be seen in Figure 2.17.

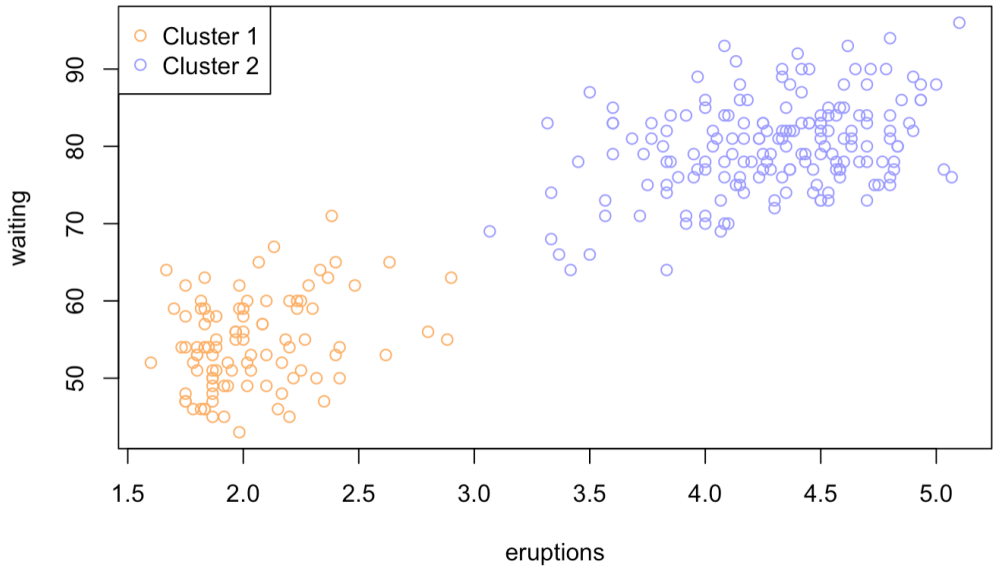

Figure 2.15: For the faithful data, graph showing the original data points being assigned to different clusters according to the Maximum a posterior (MAP) rule.

2.10.2 Soils data: Dimension reduction

In this example, we consider using the model based scores as the explanatory variable to fit a regression model with an additional new variable as the response variable. The data set we used for this analysis is the Soils data set in R package **carData** (Fox et al., 2020), which we have introduced in Section 1.2.2. Some recap of the variables from this dataset that we will use in this application: we construct a data frame with $n = 48$ and six variables: Nitrogen, Phosphorous, Calcium, Magnesium, Potassium and Sodium (which are highly correlated, but do not all use the same units), and use an additional variable 'Density' (bulk density in gm/cm^3) as the response.

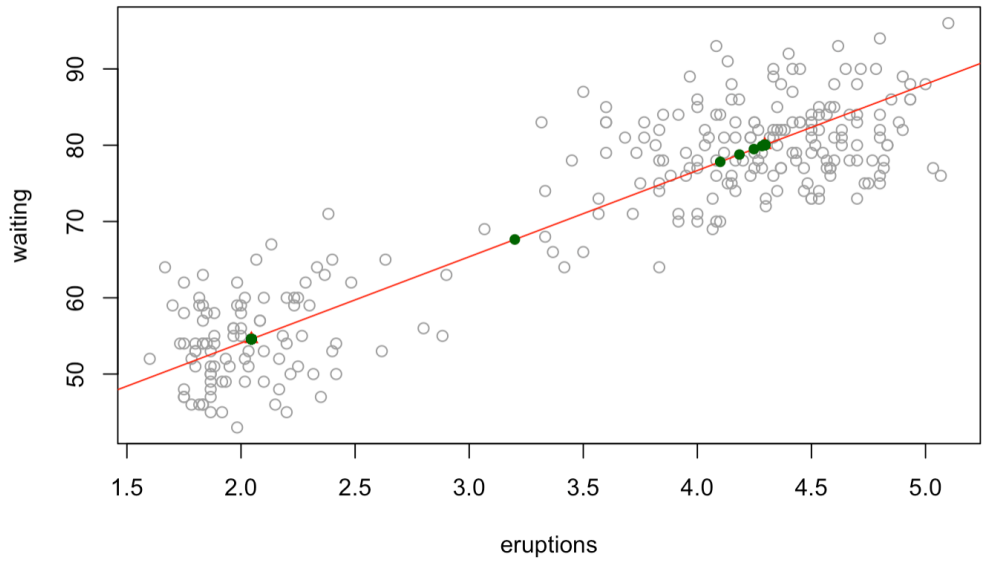

Figure 2.16: For the faithful data, graph showing the projected data points x_i^* in green.

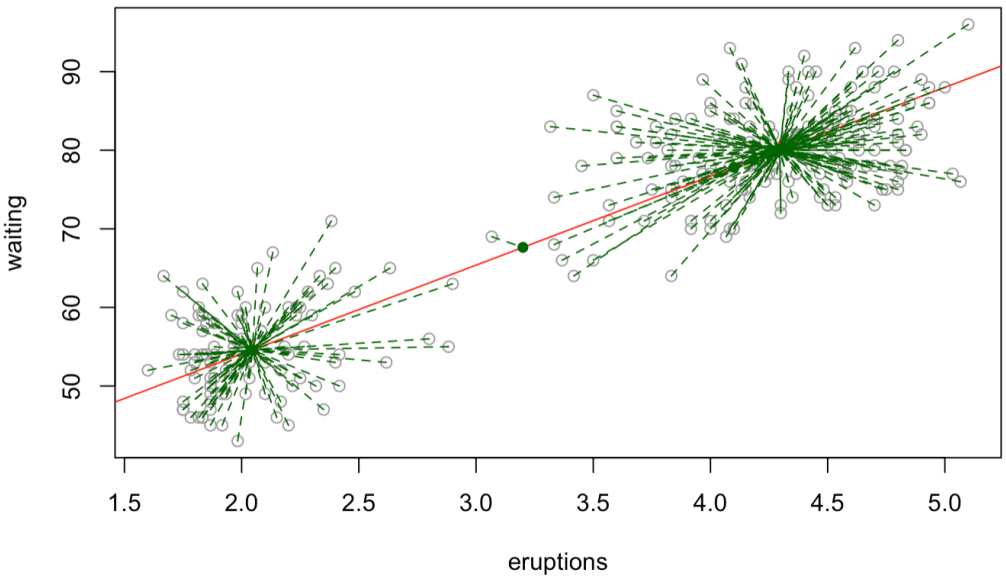

Figure 2.17: For the faithful data, graph showing the projections of the original data points onto the estimated straight latent line.

We apply the methodology laid out in Section 2.3 on the six-dimensional space of variables and use AIC and BIC to inform the choice of parameterizations and number of mass points. Details of the obtained AIC and BIC values using different number of mass points and variance parameterizations can be seen in Table 2.8 and Table 2.9. The AIC and BIC values given in these tables are the minimum values obtained over 20-50 runs with starting values chosen according to Section 2.6; the problem of finding the best solution gets harder when increasing K or the complexity of the error structure. We find that AIC and BIC suggest to use variance parameterization (ii) with 4 mass points or 3 mass points, respectively, to fit the

Table 2.8: AIC values for the Soils data under different variance parameterizations and different number of mass points.

Variance parameterization	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$
(i)	941.07	877.40	881.40	885.35	889.36
(ii)	888.38	827.99	**818.13**	823.82	849.45
(iii)	898.33	879.41	896.68	922.73	903.31
(iv)	842.40	940.30	876.08	826.69	NA

Table 2.9: BIC values for the Soils data under different variance parameterizations and different number of mass points.

Variance parameterization	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$
(i)	980.37	934.84	928.18	935.87	943.62
(ii)	938.91	**893.49**	898.59	919.25	959.85
(iii)	965.70	950.51	971.53	1001.32	985.64
(iv)	949.06	1090.00	1068.81	1062.47	NA

model.

Next we fit a regression model with the scores z_i^* being the predictor and the variable Density as response. We compare our approach to principal component regression which is a commonly used technique for computing regressions when the explanatory variables are highly correlated. For a fair comparison, we construct the first principal component scores by projecting all data points onto the 1-dimensional space and use these scores as the predictor. The fitted lines resulting from using two regression models are shown in Figure 2.18. We see that the data are represented quite differently for our methodology. Table 2.10 shows the statistical measures that evaluate the performance of principal component regression in comparison to our approach (where we have considered both the AIC and BIC solution). We find that our latent variable approach has a better performance for the non-scaled data. It is not unduly affected by scales or units and is robust concerning scaling.

2.10.3 Literacy survey data: Clustering and ranking

League tables are produced for the comparison of different institutions. Aitkin et al. (1981) compared student performance under different teaching techniques using variance component models. Aitkin and Longford (1986) investigated several modelling approaches for the comparison of school effectiveness studies. Sofroniou et al. (2008) used the International Adult

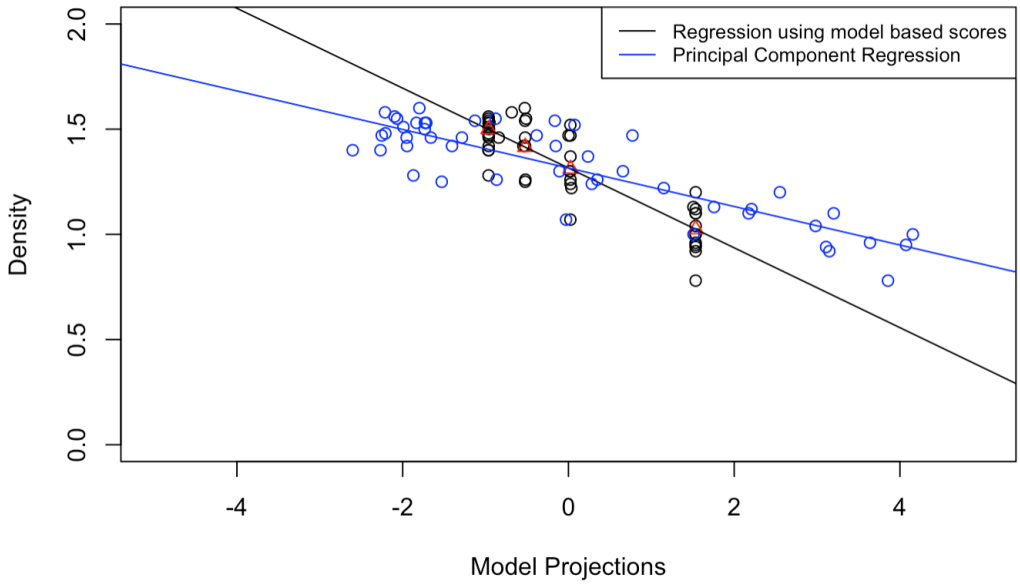

Figure 2.18: Graph showing fitted lines using two regression models. In our methodology, the regression used model based scores (model with $K = 4$) as the explanatory variable and another variable in the Soils data set called 'Density' as the response variable. The points in black correspond to Density values at the model based scores, and points in red are the Density values predicted at the mass points. For the principal component regression, the fitted regression line in blue used the first principal component (the blue points) as the explanatory variable and the variable Density as the response variable.

Table 2.10: Statistical measures of fit for the two regression models.

Regression Model	Non-scaled Data	Scaled Data
Latent variable model ($K = 3$)	$R^2 : 0.7430$ $RMSE : 0.1105$	$R^2 : 0.7231$ $RMSE : 0.1137$
Latent variable model ($K = 4$)	$R^2 : 0.7534$ $RMSE : 0.1084$	$R^2 : 0.7457$ $RMSE : 0.1088$
Principal Component Regression	$R^2 : 0.6226$ $RMSE : 0.1375$	$R^2 : 0.7435$ $RMSE : 0.1097$

Literacy Survey (IALS) data to construct league tables under the NPML estimation approach. In this section we reconsider this data set for analysis. As introduced in Section 1.2.3, the International Adult Literacy Survey (IALS) was collected in 13 countries (or country-type entities) on Prose, Document, and Quantitative scales between 1994 and 1995. The data are reported as the percentage of individuals who could not reach a basic level of literacy (being the worst) in each country, the data can be found in Table 2.12.

As in Sofroniou et al. (2008), we only use the prose scale for the analysis. However, we take the separation of the reported prose results into male and female attainment differently

into account than in that publication. We consider, for each of the 13 countries, male and female prose attainment as a bivariate response, allowing us to employ model (2.1) to describe the data, and so model (2.7) for parameter estimation. Since the gender variable is now being taken into account naturally in the response, no covariates at all are required in the model. Furthermore, since under this modelling approach both female and male prose attainment for a given country are associated with the same random effect, it also eliminates the need to fit a two-level model as in Sofroniou et al. (2008) which is otherwise needed to correlate the female and male observations within each country. So, effectively, by using a gender-defined bivariate response we are 'taking one level out' of the problem.

We fit the model with $K = 3$ mass points and with variance parametrization (ii) which leads to a minimum AIC value of 158.3963 and the smallest BIC value of 166.8705. The scores z_i^* are obtained as the posterior intercept and can be considered as the summary information of the original data. The task is here to rank the observations using the summary information. With the posterior probability matrix $W = (w_{ik})$ obtained at the convergence of the ECM algorithm, upper-level units (countries) can then be classified into different clusters according to their largest posterior probabilities.

Table 2.11 shows the joint ranking of the countries, with the countries being classified into different clusters. In the table, the 3 mass points are ordered from left to right, from the cluster in which the country has the smallest percentage of adults being illiterate to the cluster in which the country has the largest percentage of adults being illiterate. The table shows that Sweden is assigned to mass point 1 which has the smallest number of people being illiterate. Poland is the only country that is assigned to the high illiteracy mass point 3. Netherlands and Germany have posterior probabilities that spread across two mass points but are assigned to mass points 1 and 2 according to their highest posterior probabilities. We also fit the model with $K = 5$ in order to compare to the results obtained by Sofroniou et al. (2008), the results and analysis can be found in Table 2.13 and Table 2.14.

2.10.4 Foetal Movement Data: Covariates and Standard Errors

The foetal movements data we use here has been introduced in Section 1.2.4. For our analysis, the five specific movements recorded during the 4D ultrasound scans: upper face movements,

Table 2.11: Posterior intercepts and 'weight matrix' of posterior probabilities for the IALS data, with implied ranking ('league table'), for $K = 3$. Omitted entries correspond to 0.000.

		Mass points		
		0.154	0.769	0.077
Country	posterior intercept	-1.325	-0.043	3.078
Sweden	-1.325	1.000		
Netherlands	-1.323	0.999	0.001	
Germany	-0.044	0.001	0.999	
Canada	-0.043		1.000	
Australia	-0.043		1.000	
Switzerland (French)	-0.043		1.000	
New Zealand	-0.043		1.000	
Belgium (Flanders)	-0.043		1.000	
Ireland	-0.043		1.000	
United States	-0.043		1.000	
Switzerland (German)	-0.043		1.000	
United Kingdom	-0.043		1.000	
Poland	3.078			1.000

Table 2.12: Proportion of adults not achieving prose level 2 in the IALS data set.

Country	Male	Female
Ireland	24.21	20.93
United States	23.00	18.76
Switzerland (French)	17.46	19.44
Switzerland (German)	18.30	20.66
Canada	18.76	14.44
Belgium (Flanders)	15.55	21.61
Germany	14.31	13.31
United Kingdom	21.38	21.60
Netherlands	10.39	10.49
Poland	43.72	41.74
Sweden	7.31	7.18
Australia	18.33	15.69
New Zealand	19.94	16.52

Table 2.13: Posterior intercepts, weight matrix and implied ranking for the IALS data using model (2.47) with $K = 5$. Omitted entries correspond to 0.00.

		Mass points				
		0.15	0.08	0.32	0.37	0.08
Country	posterior intercept	-1.29	-0.66	-0.17	0.19	2.99
Sweden	-1.29	1.00				
Netherlands	-1.29	1.00				
Germany	-0.66		1.00			
Canada	-0.17			1.00		
Australia	-0.17			1.00		
New Zealand	-0.15			0.95	0.05	
Switzerland (French)	-0.03			0.61	0.39	
Switzerland (German)	0.04			0.42	0.58	
Belgium (Flanders)	0.15			0.12	0.88	
United Kingdom	0.19				1.00	
United States	0.19				1.00	
Ireland	0.19				1.00	
Poland	2.99					1.00

Table 2.14: Classification and ranking for the IALS data in the paper by Sofroniou et al. (2008).

		Mass points				
		0.077	0.093	0.434	0.319	0.077
Country	posterior intercept	-2.602	-2.156	-1.599	-1.379	-0.330
Sweden	-2.60	1.00				
Netherlands	-2.16		1.00			
Germany	-1.72		0.21	0.79		
Australia	-1.60			1.00		
Canada	-1.59			0.97	0.03	
New Zealand	-1.58			0.92	0.08	
Belgium (Flanders)	-1.58			0.89	0.11	
Switzerland (French)	-1.54			0.72	0.28	
Switzerland (German)	-1.45			0.34	0.66	
United States	-1.38			0.01	0.99	
Ireland	-1.38				1.00	
United Kingdom	-1.38				1.00	
Poland	-0.33					1.00

head movements, mouth movements, touch movements and eye blink will be considered as a five-variate response, $x_i \in \mathbb{R}^5$ whereas status ('pre-Covid' or 'during Covid') is the predictor, $v_i \in \mathbb{R}$.

We fit the data to model (2.47) with $K = 3$ and variance parametrization (ii) which leads to the smallest AIC (554.3622) value and BIC (613.473) value among all parametrizations and mass points. In principle, one could fit five separate linear regression models, each taking one of the movements score as the response and status as the predictor. We compare the

estimates of the parameters and the parameter standard errors using this 'naïve' method to our proposed approach, using model (2.47), where the five equations are linked through a common random effect, the results are shown in Table 2.15 and Table 2.16. Our methodology, involving a multivariate response model with random effect, gives parameter estimates which are consistent with the ones obtained from separate linear models, however enjoying reduced standard errors of the coefficients. The bottom row of Table 2.15 and Table 2.16 also gives the p-values of the estimated $\hat{\gamma}$'s. We observe that the p-values also tend to be reduced, leading to a potentially different decision on the significance of a predictor variable if a decision threshold is crossed.

Table 2.15: For the Covid data, estimations of γ obtained using individual linear models for upper face movements, head movements, mouth movements, touch movements and eye blink.

	indiv. linear models				
	upper face	head movements	mouth movements	touch movements	eye blink
estimate ($\hat{\gamma}$)	0.472	0.217	2.600	0.317	0.367
standard error	0.251	0.274	1.135	0.357	0.435
p-value	0.068	0.432	0.028	0.380	0.405

Table 2.16: For the Covid data, estimations of γ obtained using the proposed multivariate response model with random effect. Standard errors and p-values are obtained via the bootstrap.

	multivariate model				
	upper face	head movements	mouth movements	touch movements	eye blink
estimate ($\hat{\gamma}$)	0.460	0.203	2.549	0.297	0.346
standard error	0.193	0.208	0.878	0.250	0.361
p-value	0.051	0.381	0.048	0.224	0.323

2.11 Relationship with existing methodologies

2.11.1 Factor analysis

Some methodologically related techniques have been previously suggested in the literature, partly very long ago. In the homoscedastic case, the one-level model (2.1) can be seen as a one-dimensional factor analysis model (see Murphy 2012, chapter 12), with the difference that we will apply a discrete mixture approximation of the latent variable z_i.

The marginal probability density function $f(x_i|\alpha,\beta)$ for observations generated from model (2.1) can be written as

$$f(x_i|\alpha,\beta) = \int f(x_i, z_i|\alpha,\beta)dz_i = \int f(x_i|z_i,\alpha,\beta)\phi(z_i)dz_i, \tag{2.50}$$

where $f(x_i, z_i|\alpha,\beta)$ is the joint probability distribution of observed data x_i and unobserved random effects z_i, and $\phi(\cdot)$ is the density function of the random effect distribution Z. Previously, we do not make any explicit assumption regarding the distribution of the random effect and applied Aitkin's nonparametric maximum likelihood approach to deal with the integral in (2.50). But in the framework of factor analysis, one could do this based on a Gaussianity assumption on $\phi(\cdot)$. Then the marginal distribution can be written as

$$\begin{aligned} f(x_i|\alpha,\beta) &= \int \mathcal{N}(x_i|\beta z_i + \alpha, \Sigma_{x_i})\mathcal{N}(z_i|\mu_{z_i}, \Sigma_{z_i})dz_i \\ &= \mathcal{N}(x_i|\beta\mu_{z_i} + \alpha, \Sigma_{x_i} + \beta\beta^T), \end{aligned} \tag{2.51}$$

where Σ_{x_i} is a diagonal matrix. Note that the error variance in our model (2.1) is allowed to depend on the random effect, while in factor analysis, it is explicitly stated that the error variance and the random effect are independent of each other.

Then following the derivation by Ghahramani and Hinton (1996), we could obtain,

$$E(z|x) = \eta x,$$

where $\eta = (\Sigma_{x_i} + \beta\beta^T)^{-1} = \Sigma_{x_i}^{-1} - \Sigma_{x_i}^{-1}\beta(I + \beta^T\Sigma_{x_i}^{-1}\beta)^{-1}\beta^T\Sigma_{x_i}^{-1}$, and,

$$E(zz^T|x) = Var(z|x) + E(z|x)E(z|x)^T = I - \eta\beta + \eta x x^T \eta^T.$$

In the **E**-step, $E(z|x)$ and $E(zz^T|x)$ will be computed.

In the **M**-step, we compute,

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i E(z|x_i)^T\right)\left(\sum_{l=1}^{n} E(zz^T|x_l)\right)^{-1},$$

and,

$$\hat{\Sigma}_{x_i} = \frac{1}{n} diag\left\{ \sum_{i=1}^{n} x_i {x_i}^T - \hat{\beta} E(z|x_i) {x_i}^T \right\}.$$

The random effect z_i is estimated through the least square estimates and the derivation can be found in (Krzanowski 2000, chapter 16),

$$\hat{z}_i = (\hat{\beta}^T \hat{\Sigma}_{x_i}^{-1} \hat{\beta})^{-1} \hat{\beta}^T \hat{\Sigma}_{x_i} (x_i - \bar{x}).$$

By comparison, the $\hat{z}_i$ and $\hat{\beta}$ derived under the framework of factor analysis is similar in the structure to the $\hat{z}_k$ and $\hat{\beta}$ with the computations based on an ECM algorithm in the spirit of the Nonparametric Maximum Likelihood approach for the estimation of mixture models in our one-level model. However, the intercept in our model α is omitted in factor analysis, and the implementation of the factor analysis is not designed for small dimensional data, e.g. the function `factanal()` in R package **stats** requires at least three variables. And the estimation of the parameters under the framework of factor analysis still requires EM algorithm, and the computation is not getting easier compared to our approach.

To perform clustering as done in Section 2.10 (e.g. Figure 2.15) based on factor analysis, a two-stage process would be required: first, compute the one-dimensional factor scores, and then perform K-means on these factor scores. In contrast, our methodology allows this to be done in a single step. Additionally, we propose four types of parameterizations of the variance matrices (see Figure 2.14), while the factor analysis model assumes the error variance matrix to be diagonal and unique, which is related to the independence assumption between the error variance and the random effect. With the factor scores, we could perform regression using the factor scores as predictors and an additional variable as the response, as we did in Section 2.10.2, where we compared the fitting of principal component regression and our method. In Section 2.10.4, we applied the proposed model with the inclusion of covariates in regression, which, by accounting for the correlation of the response variables, leads to reduced standard errors for the coefficient estimates. However, factor analysis in its standard form cannot deal with covariates in a straightforward way. There have been methodologies (e.g. Fan et al. 2016; Li and Jung 2017) which use factor analysis in conjunction with covariates but these methods attempt to directly model the factors as a function of covariates, rather than adjusting the

overall model for the effect of the covariates as what has been done in our methodology.

Finally, we consider to use a real data set to show the similarities of our one-level model and the factor analysis method when only one factor is used. The data used here is again the Soils data (used in Section 2.10.2) available in the R package **carData**. We consider the six chemical elements variables: nitrogen, phosphorous, calcium, magnesium, potassium, and sodium. The results are shown in Table 2.17. The similar results indicate that β serves the same purpose in our one-level model as the loadings Γ in factor analysis. However, the conceptual mathematics behind these two methods and the methods of estimating the parameters are different as we have discussed above.

	nitrogen	phosphorous	calcium	magnesium	potassium	sodium
$\hat{\beta}$	0.90	0.80	0.87	-0.54	0.74	-0.87
Loadings	0.95	0.85	0.88	-0.55	0.73	-0.88

Table 2.17: The estimated $\hat{\beta}$ and the loadings Γ from factor analysis using one factor.

2.11.2 GTM

There is also some overlap with the Generative Topographic Mapping (GTM, Bishop et al. 1998), which allows for non-linear manifolds. In GTM, the function $y(t; W)$, where t is a latent variable and W is the parameter matrix, is defined as a continuous and differentiable mapping, rather than just a latent straight line. Pena et al. (2008) revisited the GTM and provided the following equations to explain the mathematical details of this approach. The marginal probability density function for the GTM can be written as

$$f(x|W, \sigma) = \int f(x|t, W, \sigma)\phi(t)dt, \tag{2.52}$$

where x is the original data point, σ^2 is the variance, and $\phi(t)$ is the probability distribution of the latent variable t. In our proposed model, we encounter a situation similar to the one where a decision must be made regarding how to handle the integral over the latent variable. In the GTM the latent variables are parameterized by a fixed and equidistant grid,

$$\phi(t) = \frac{1}{K}\sum_{k=1}^{K} f(x|t_k, W, \sigma), \tag{2.53}$$

where $\phi(t)$ is a set of K equally weighted delta function, which makes the integral to become a summation,

$$f(x|W,\sigma) = \frac{1}{K}\sum_{k=1}^{K} f(x|t_k, W, \sigma). \tag{2.54}$$

Comparing Equation (2.53) to our methodology, we observe that we make no assumptions regarding the distribution of the latent variable, and we used estimable masses and mass points to deal with the integral as compared to Equation (2.54). The way of dealing with the integral in GTM renders the approach less suitable for clustering-type applications. Another difference is that rather than using different variance matrices for different clusters, all mixture components in GTM share the same variance. Under the GTM approach, there is no immediate possibility to include covariates and hence they do not serve as a multivariate response model.

Chapter 3

Two-level Model

In Section 2.7, we proposed a one-level random effect model, in which multivariate observations $x_i \in \mathbb{R}^m$, $i = 1, \ldots, n$, are described by a fixed effect covariate term plus a random effect term based on a one-dimensional latent variable parameterizing a straight line cutting through the multivariate space as follows

$$x_i = \Gamma v_i + \alpha + \beta z_i + \varepsilon_i, \tag{3.1}$$

where $v_i \in \mathbb{R}^p$ is a vector of the covariates, $\Gamma_{m\times p}$ is a matrix of the coefficients of the covariates, $\alpha \in \mathbb{R}^m$ and $\beta \in \mathbb{R}^m$ are m- variate parameter vectors, z_i is a one-dimensional random effect, and $\varepsilon_i \in \mathbb{R}^m$ independent Gaussian errors possibly depending on z_i.

Without the covariate term Γv_i, model (3.1) performs dimension reduction by approximating the original higher dimensional observations by $\alpha + \beta z_i$, which is a straight line through m-variate space, parameterized by the one-dimensional latent variable $z_i \in \mathbb{R}$.

With the covariate term Γv_i, this model can be seen as a multiple regression model for multivariate responses, where the random effect term is ensuring that correlations among the different response measurements are taken into account, potentially resulting in reduced standard errors of parameter estimates. When seen as a clustering technique, this model specification ensures that clusters are adjusted for the specificed covariates, avoiding the potential impact of these covariates on the clustering outcome.

In summary, this approach can lead to quite powerful inferences, despite the seemingly restrictive assumption of a one-dimensional latent linear subspace However, as a limitation, model (3.1) could not yet deal with repeated measures, which would require a two-level extension of the model. We will see that the model formulation (3.1) can be expanded so that a single random effect z_i can be used to account for correlations between observations sharing the same level, thereby ensuring that the estimated models for several response dimensions remain linked. This section aims to provide this extension, giving equal importance to the

applications of clustering (of upper-level units) and multivariate regression for two-level data. The development will begin with the latter, with the former arising as a by-product.

Some related model classes have been developed in the wider context of item response theory, most notably latent class models (Goodman, 1974). These models are commonly used method for the clustering of observed multivariate categorical data (such as questionnaire outcomes on Likert scales) into latent classes. An obvious difference to our methodology is that in latent class models the response variables are categorical rather than continuous. A multilevel version of latent class models was developed by Vermunt (2003). A model selection procedure for deciding the number of latent classes at both levels is proposed by Lukočienė et al. (2010). Gnaldi et al. (2016) introduced a multilevel version latent class-item response theory model applied for educational data in which the collected response variables are dependent of each other. The latent class analysis also allows the inclusion of covariates; Di Mari et al. (2023) proposed a two-step estimator for the multilevel latent class model in which two categorical random effects are used to account for both the upper and lower levels, allowing for clustering of the latent classes on both levels. It remains the case that due to the restriction on categorical outcomes, latent class models cannot be applied or compared with the situations dealt with in this work. However, it should not be left unstated that continuous-outcome versions of multilevel latent class models have also been developed, and are available in specialized commercial software such as Latent GOLD (Vermunt, 2008).

3.1 A Two-level Model for Multivariate Response Data

We consider a scenario where multivariate data $x_{ij} \in \mathbb{R}^m$ has a two-level structure, with the upper level indexed by $i = 1, 2, ..., r$ and the lower level by $j = 1, 2, ..., n_i$. The proposed two-level model takes the form

$$x_{ij} = \alpha + \beta z_i + \Gamma v_{ij} + \varepsilon_{ij}, \tag{3.2}$$

where $\alpha, \beta \in \mathbb{R}^m$, $z_i \in \mathbb{R}$, $v_{ij} \in \mathbb{R}^p$ is the vector of covariates (which may include upper-level variates not depending on j), $\Gamma \in \mathbb{R}^{m \times p}$ is a matrix of the covariate coefficients, and $\varepsilon_{ij} \sim$

$N(0, \Sigma(z_i))$ are independent Gaussian errors. Under such a model, equivalently represented as

$$x_{ij}|z_i, \alpha, \beta, \Gamma \sim N(\alpha + \beta z_i + \Gamma v_{ij}, \Sigma(z_i)) \tag{3.3}$$

the data grouping process is carried out on the upper level, while the lower level units within the same upper level unit share a common random effect term z_i. Thus, the random effect induces a line cutting across the multivariate space of responses, along which the latent values z_i are positioned. Again equivalently, and for later reference, we can write the conditional probability density function of the x_{ij} as

$$\begin{aligned} & f(x_{ij}|z_i, \alpha, \beta, \Gamma) = \\ & = \ (2\pi)^{-m/2}|\Sigma(z_i)|^{-1/2} \exp\left\{-\frac{1}{2}(x_{ij} - \alpha - \beta z_i - \Gamma v_{ij})^T \Sigma^{-1}(z_i)(x_{ij} - \alpha - \beta z_i - \Gamma v_{ij})\right\}. \end{aligned} \tag{3.4}$$

For the distribution of random effects z_i, denoted here by Z, several choices are possible, including a Gaussian distribution. In this work, we consider to use Aitkin's Nonparametric Maximum Likelihood approach (Aitkin, 1999), in which their distribution is approximated by a discrete mixture. However, as will be detailed in the following section, this is not so much a distributional 'assumption', but rather a technical device to approximate the marginal likelihood, allowing for estimation of the model parameters. De facto this approach leads to the estimation of a constrained multivariate mixture model, with mixtures centres spanned a along a straight line through the space of responses. This makes this approach particularly suitable for data which are correlated and clustered at the same time. Bouveyron and Brunet-Saumard (2014) provided a comprehensive review of model-based clustering of high-dimensional data, encompassing constrained and parsimonious models. Celeux and Govaert (1995) explored various clustering situations through the eigenvalue decomposition of the variance matrices of the mixture components. Banfield and Raftery (1993) enabled the variation of all cluster features by a reparameterization of the covariance matrix for Gaussian clustering.

When there is only one covariate $v_{ij} \in \mathbb{R}$, we write $\Gamma = \gamma \in \mathbb{R}^m$. Figure 3.1 illustrates a data scenario corresponding to this concept. The data used here is simulated from model (3.2) in the case that the latent variable obeys a three-point mixture distribution; i.e. the z_i take one of three pre-specified values of z_k with probabilities π_k, $k = 1, \ldots, K$. The grey straight line

represents the one-dimensional latent space $\alpha + \beta z$, and the black triangles positioned along the straight line the mixture centres of each component. The coloured thinner lines are for illustration only and show the trend of lower-level units within each each upper level (which is to some extent a result of the random error and to some part driven by the covariate). The orange triangles are the fitted values: $x_{ij}^* = \hat{\alpha} + \hat{\beta} z_i^* + \hat{\gamma} v_{ij}$, where $z_i^* = \sum_{k=1}^{K} w_{ik} \hat{z}_k \in \mathbb{R}$ are obtained as the posterior random effects using posterior probabilities of component membership w_{ik}, calculated according to Bayes' theorem (Aitkin, 1996a). There is an angle between the fitted values and the one-dimensional latent space (represented by the gray straight line). The fitted values are tilted from the straight line by the effect of the covariate coefficient matrix Γ.

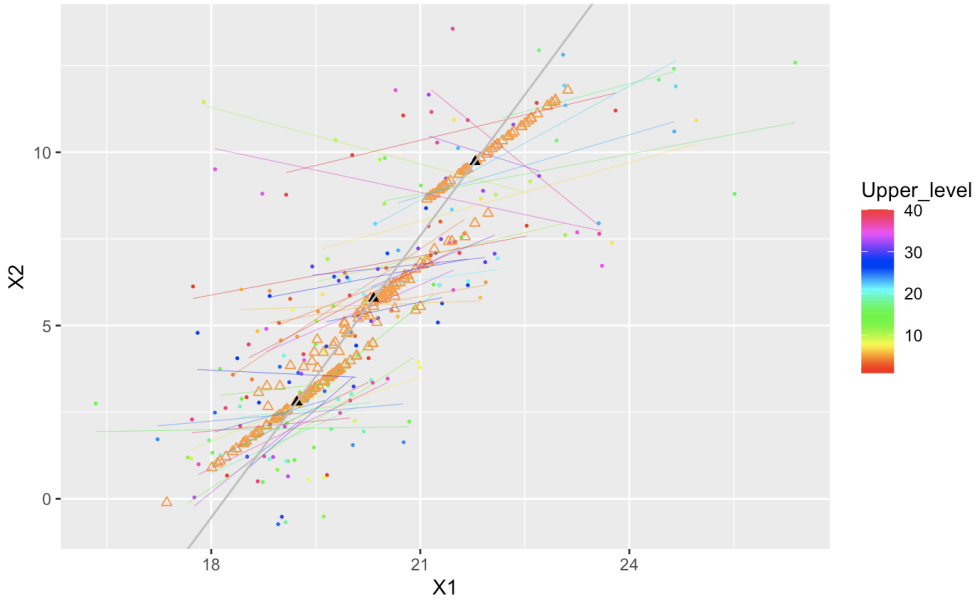

Figure 3.1: Simulated data with 40 upper level units, each with 5 lower level units, with $\alpha^T = (20, 10)$, $\beta^T = (1, 3)$, $\pi_k^T = (0.2, 0.3, 0.5)$, $z_k^T = (1.73, 0.29, -0.87)$, $\gamma = (0.5, 1)$. Observations are generated with component-specific diagonal variance matrices $\Sigma_k = \Sigma(z_k)$.

3.2 Likelihood and Estimators

Let $x_i = (x_{i1}, \ldots, x_{in_i})^T \in \mathbb{R}^{n_i \times m}$ denote the collection of the m-variate lower-level observations relating to the i-th upper level unit. Since these lower-level units are conditionally independent given z_i, we have

$$f(x_i | z_i, \alpha, \beta, \Gamma) = \prod_{j=1}^{n_i} f(x_{ij} | z_i, \alpha, \beta, \gamma).$$

According to model (3.2), the marginal distribution of x_i, which is required for the construction of the likelihood function, can be obtained by integrating over the distribution of z_i, as follows

$$f(x_i|\alpha,\beta,\Gamma) = \int \left[\prod_{j=1}^{n_i} f(x_{ij}|z_i,\alpha,\beta,\Gamma)\right] g(z_i)dz_i, \tag{3.5}$$

where $g(z_i)$ is the density function for the unobserved random effects z_i. Given that we don't make any explicit assumptions regarding the distribution Z of the z_i, $g(z_i)$ can be non-Gaussian, making it infeasible to compute the integration using an analytical approach. Under the Nonparametric Maximum Likelihood approach (Aitkin, 1999), we replace the integral over z_i by a finite sum over K mass points $z_1,\ldots,z_k$ with associated masses $\pi_1,\ldots,\pi_k$, for $k = 1,\ldots,K$. Here we treat the mass points and masses as unknown parameters to be estimated. The value of K will be treated as known in the parameter estimation process and the best choice of K in a fitted model will be selected through the use of model selection criteria, specifically based on the AIC criterion.

The marginal distribution can then be approximated as

$$f(x_i|\alpha,\beta,\Gamma) \approx \sum_{k=1}^{K} \left[\prod_{j=1}^{n_i} f(x_{ij}|z_k,\alpha,\beta,\Gamma)\right] \pi_k, \tag{3.6}$$

in which, by virtue of (3.3),

$$x_{ij}|z_k,\alpha,\beta,\Gamma \sim N(\alpha+\beta z_k+\Gamma v_{ij},\Sigma(z_k)), \tag{3.7}$$

with the component-specific densities $f(x_{ij}|z_k,\alpha,\beta,\Gamma)$ as in Equation (3.4), but with z_i replaced by z_k.

Now, the $\alpha+\beta z_k$ can be interpreted as the locations, in m–dimensional space, of the mixture centers spanned along the one-dimensional latent space, with cluster-wise variances $\Sigma_k \equiv \Sigma(z_k)$ replacing the previous observation-specific variances $\Sigma(z_i)$. The number of parameters to be estimated is effectively reduced by constraining to K distinct variance matrices.

Building on Equation (3.6), the approximated marginal log-likelihood can be obtained as

$$l(\alpha,\beta,\Gamma,z_1,\ldots,z_K|x_1,\ldots x_r) \approx \sum_{i=1}^{r} \log\left\{\sum_{k=1}^{K}\left[\prod_{j=1}^{n_i} f(x_{ij}|z_k,\alpha,\beta,\Gamma)\right]\pi_k\right\}. \tag{3.8}$$

In preparation of the EM algorithm (Dempster et al., 1977) to be used for the parameter estimation (similar to the one-level model, what we are effectively using is an ECM algorithm rather than a general EM algorithm), we define by G_{ik} an indicator variable taking the value 1 if the upper-level unit i belongs to component k, and 0 otherwise (which is, of course, unknown – this is the 'missing information' for the EM machinery). We also denote by $G_i = (G_{i1}, \ldots, G_{iK})^T$ the set of indicators for that unit. This yields 'complete data' $\{x_i, G_i\}$, with probability

$$P(x_i, G_i) = \prod_{k=1}^{K} (f_{ik}\pi_k)^{G_{ik}},$$

where for simplicity of notation we here used $f_{ik} \equiv \prod_{j=1}^{n_i} f(x_{ij}|z_k, \alpha, \beta, \Gamma)$. The complete likelihood can now be written as follows

$$L_c = \prod_{i=1}^{r} \prod_{k=1}^{K} (\pi_k f_{ik})^{G_{ik}}. \tag{3.9}$$

Hence we obtain the complete log-likelihood,

$$l_c = \log L_c = \sum_{i=1}^{r} \sum_{k=1}^{K} G_{ik} \log(\pi_k f_{ik}). \tag{3.10}$$

The expectation $w_{ik} = E[G_{ik}|x_i] = P(G_{ik} = 1|x_i) = \pi_k f_{ik} / \sum_\ell \pi_\ell f_{i\ell}$ is just the 'posterior' probability of each upper-level unit i belonging to component k. Therefore, the expected complete log-likelihood is written as

$$\begin{aligned} l_c^* &= \sum_{i=1}^{r} \sum_{k=1}^{K} \mathbb{E}\left[G_{ik}|x_i\right] \log\left(\pi_k f_{ik}\right) \\ &= \sum_{i=1}^{r} \sum_{k=1}^{K} w_{ik} \log \pi_k + \sum_{i=1}^{r} \sum_{j=1}^{n_i} \sum_{k=1}^{K} w_{ik} \log f(x_{ij}|z_k, \alpha, \beta, \Gamma). \end{aligned} \tag{3.11}$$

Plugging the expression for $f(x_{ij}|z_k, \alpha, \beta, \Gamma)$ into Equation (3.11), we obtain the expected complete log-likelihood as follows

$$\begin{aligned} {l_c}^* = &\sum_{i=1}^{r} \sum_{k=1}^{K} w_{ik} \log(\pi_k) - \frac{1}{2} \sum_{i=1}^{r} \sum_{j=1}^{n_i} \sum_{k=1}^{K} w_{ik} \log(|\Sigma_k|) - \frac{m}{2} \log(2\pi) \sum_{i=1}^{r} n_i \\ &-\frac{1}{2} \sum_{i=1}^{r} \sum_{j=1}^{n_i} \sum_{k=1}^{K} w_{ik} (x_{ij} - \alpha - \beta z_k - \Gamma v_{ij})^T \Sigma_k^{-1} (x_{ij} - \alpha - \beta z_k - \Gamma v_{ij}). \end{aligned} \tag{3.12}$$

By taking partial derivatives of l_c^* with respect to each parameter and letting the score equations to be 0 and solving them, we find that

$$\hat{z}_k = \frac{\sum_{i=1}^{r} w_{ik} \sum_{j=1}^{n_i} \hat{\beta}^T \hat{\Sigma}_k^{-1} (x_{ij} - \hat{\alpha} - \hat{\Gamma} v_{ij})}{\sum_{i=1}^{r} n_i w_{ik} \hat{\beta}^T \hat{\Sigma}_k^{-1} \hat{\beta}}, \; k = 1, \ldots, K. \tag{3.13}$$

$$\hat{\beta} = \left(\sum_{i=1}^{r} \sum_{j=1}^{n_i} \sum_{k=1}^{K} w_{ik} \hat{\Sigma}_k^{-1} \hat{z}_k^2 \right)^{-1} \left(\sum_{i=1}^{r} \sum_{j=1}^{n_i} \sum_{k=1}^{K} w_{ik} \hat{\Sigma}_k^{-1} (x_i - \hat{\alpha} - \hat{\Gamma} v_{ij}) \hat{z}_k \right), \tag{3.14}$$

$$\hat{\alpha} = \left(\sum_{i=1}^{r} \sum_{j=1}^{n_i} \sum_{k=1}^{K} w_{ik} \hat{\Sigma}_k^{-1} \right)^{-1} \left(\sum_{i=1}^{r} \sum_{j=1}^{n_i} \sum_{k=1}^{K} w_{ik} \hat{\Sigma}_k^{-1} (x_i - \hat{\beta} \hat{z}_k - \hat{\Gamma} v_{ij}) \right), \tag{3.15}$$

The solution for $\hat{\Gamma}$ can only be given implicitly in the form of estimating equation

$$\sum_{i=1}^{r} \sum_{j=1}^{n_i} \sum_{k=1}^{K} w_{ik} \hat{\Sigma}_k^{-1} (x_{ij} - \hat{\alpha} - \hat{\beta} z_k) {v_{ij}}^T = \sum_{i=1}^{r} \sum_{j=1}^{n_i} \sum_{k=1}^{K} w_{ik} \hat{\Sigma}_k^{-1} \hat{\Gamma} v_{ij} {v_{ij}}^T. \tag{3.16}$$

We furthermore find the general solution for Σ_k as

$$\hat{\Sigma}_k = \frac{\sum_{i=1}^{r} \sum_{j=1}^{n_i} w_{ik} (x_{ij} - \hat{\alpha} - \hat{\beta} \hat{z}_k - \hat{\Gamma} v_{ij})(x_{ij} - \hat{\alpha} - \hat{\beta} \hat{z}_k - \hat{\Gamma} v_{ij})^T}{\sum_{i=1}^{r} n_i w_{ik}}, \tag{3.17}$$

for $k = 1, \ldots, K$. Finally, since for the mixture probabilities $\sum_{k=1}^{K} \pi_k = 1$, we apply a Lagrange multiplier by letting $\partial \left(l_c^* - \lambda (\sum_{k=1}^{K} \pi_k - 1) \right) / \partial \pi_k = 0$. Hence, we find (Aitkin et al., 2009),

$$\hat{\pi}_k = \frac{\sum_{i=1}^{r} w_{ik}}{r}. \tag{3.18}$$

We note that this set of Equations (3.13) to (3.18) is rather impractical to use directly, because the equations depend on each other in a complex manner, they involve multiple inversions of the estimated matrix $\hat{\Sigma}_k$, and the solution for Γ does not have an explicit form. However, it is also not necessary to apply these equations in full generality. An immediate simplification is suggested by considering the matrices Σ_k. While these variance matrices, under a full unconstrained parameterization, could deal with clusters that differ by shape and size, we found little evidence that such complex variance parameterizations are helpful or necessary in the context of two-level models. This is in line with similar results for one-level models by Zhang and Einbeck (2024d) and what is written in subsection 2.1. Hence, we will restrict this

to diagonal variance matrices

$$\Sigma_k = \text{diag}(\sigma_{lk}^2)_{\{1\leq l\leq m\}}, k = 1, ..., K.$$

To avoid potential identifiability issues, certain restrictions are imposed on the model. First we enforce $\beta_1 \geq 0$ to identify the direction of the latent variable. Then we standardize z_k by $\sum_{k=1}^K \pi_k z_k = 0$, and $\sum_{k=1}^K \pi_k z_k^2 - (\pi_k z_k)^2 = 1$, where $\text{Var}[z_k] = \sum_{k=1}^K \pi_k z_k^2 - (\pi_k z_k)^2$ (Marques da Silva Júnior et al., 2018).

The resulting ECM algorithm, which makes some further simplifications which are however of computational rather than model-related character, is presented in the next subsection.

3.3 ECM Algorithm

We have the following Expectation (E) and Maximization (M) steps resulting from the previous considerations.

E-step

The E-step is obtained from straightforaward application of Bayes' theorem as illustrated in the previous subsection (Aitkin et al., 2009),

$$w_{ik} = \frac{\pi_k f_{ik}}{\sum_l \pi_l f_{il}}. \tag{3.19}$$

M-step

In order to implement the M-step computationally, we adopt the strategy employed in Zhang and Einbeck (2024d) (also written in subsection 2.3). For this, we detach the updates of $\hat{\alpha}$, $\hat{\beta}$, $\hat{z}_k$ and $\hat{\Gamma}$ from those of $\hat{\Sigma}_k$, by invoking, only for the use within expressions (3.13) to (3.16), a further simplification where the variance matrices are assumed to be constant and diagonal, i.e. $\sigma_{lk}^2 \equiv \sigma^2$ for all l and k. This leads to simpler equations for (3.13) to (3.16) as follows

$$\hat{z}_k = \frac{\sum_{i=1}^r w_{ik} \sum_{j=1}^{n_i} \hat{\beta}^T (x_{ij} - \hat{\alpha} - \hat{\Gamma} v_{ij})}{\sum_{i=1}^r n_i w_{ik} \hat{\beta}^T \hat{\beta}}, \tag{3.20}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{K} w_{ik}\hat{z}_k x_{ij} - \frac{1}{n}(\sum_{i=1}^{r}\sum_{j=1}^{n_i} x_{ij})(\sum_{i=1}^{r} n_i \sum_{k=1}^{K} w_{ik}\hat{z}_k)}{\sum_{i=1}^{r} n_i \sum_{k=1}^{K} w_{ik}\hat{z}_k^2 - \frac{1}{n}(\sum_{i=1}^{r} n_i \sum_{k=1}^{K} w_{ik}\hat{z}_k)^2}$$
$$- \frac{\hat{\Gamma}\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{K} w_{ik}\hat{z}_k v_{ij} - \frac{1}{n}(\sum_{i=1}^{r} n_i \sum_{k=1}^{K} w_{ik}\hat{z}_k)(\hat{\Gamma}\sum_{i=1}^{r}\sum_{j=1}^{n_i} v_{ij})}{\sum_{i=1}^{r} n_i \sum_{k=1}^{K} w_{ik}\hat{z}_k^2 - \frac{1}{n}(\sum_{i=1}^{r} n_i \sum_{k=1}^{K} w_{ik}\hat{z}_k)^2}, \quad (3.21)$$

$$\hat{\alpha} = \frac{1}{n}\left(\sum_{i=1}^{r}\sum_{j=1}^{n_i} x_{ij} - \hat{\beta}\sum_{i=1}^{r} n_i \sum_{k=1}^{K} w_{ik}\hat{z}_k - \hat{\Gamma}\sum_{i=1}^{r}\sum_{j=1}^{n_i} v_{ij}\right), \quad (3.22)$$

with the estimator for Γ now being available in explicit form,

$$\hat{\Gamma} = \left(\sum_{i=1}^{r}\sum_{j=1}^{n_i} v_{ij}v_{ij}^T\right)^{-1}\left(\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{K} w_{ik}(x_{ij} - \hat{\alpha} - \hat{\beta}\hat{z}_k)v_{ij}^T\right). \quad (3.23)$$

These four equations are then iterated for a small number of times between each other, where the $\hat{z}_k$, $k = 1, \ldots, K$, are immediately re-standardized to mean 0 and variance 1 after the execution of (3.20) . This routine is then followed by the estimation of the π_k via (3.18), and the update of the variance matrices via $\hat{\Sigma}_k = \text{diag}(\hat{\sigma}_{lk}^2)_{\{1\leq l\leq m\}}, k = 1, ..., K$. Write $\phi_{ij} = \Gamma v_{ij} \in \mathbb{R}^m$ and let $\phi_{ij\ell}$ be its ℓ-th component, $\ell = 1, \ldots, m$. Then,

$$\hat{\sigma}_{lk}^2 = \frac{\sum_{i=1}^{r}\sum_{j=1}^{n_i} w_{ik}(x_{ijl} - \hat{\alpha}_l - \hat{\beta}_l\hat{z}_k - \phi_{ijl})^2}{\sum_{i=1}^{r} n_i w_{ik}}. \quad (3.24)$$

This completes the M-step, and the procedure continues with the E-step (3.19). The derivations of the parameter estimators used in the ECM algorithm can be found in Appendix B. A detailed procedure for the ECM algorithm is given in Algorithm 2 below.

3.4 Intraclass Correlation

In order to understand how the correlation of the lower-level units within each upper-level unit is accounted for, we use the intraclass correlation. We developed a formula to calculate the ICC for the proposed two-level model as follows. We do not assume a specific distribution of the random effect but we can generally write that $z_i \sim (\mu_z, {\sigma_z}^2)$, where μ_z is the mean of z_i

Algorithm 2 ECM Algorithm

1. Initialization:

(i) Choose the number of mixture components, K, where K is a positive integer.

(ii) Choose starting values for the parameters: π_k, α, β, z_k, Γ, Σ_k; four options are available.

(iii) Select the number of iterations, s; 20 iterations is suggested.

2. Iterations:

E-step

For each k, compute the posterior probability of observation i belonging to component k, according to Equation (3.19).

M-step

steps $\leftarrow 0$
while *steps* $\leq s$ **do**
  counter $\leftarrow 0$ ▷ Reset counter for each step
  while *counter* ≤ 5 **do**
    Update z_k, β, α and Γ, cycle between Equations (3.20), (3.21), (3.22), and (3.23).
    counter $\leftarrow$ *counter* $+ 1$
  end while
  Update π_k via Equation (3.18)
  Update Σ_k according to Equation (3.24)
  steps $\leftarrow$ *steps* $+ 1$
end while

3. Output: Return the estimated parameters.

and ${\sigma_z}^2$ is the variance of z_i. We denote Σ to be the diagonal variance matrix for the Gaussian noise ε_{ij}. The expectation for the multilevel multivariate data x_{ij} is

$$\mathbb{E}\left[x_{ij}\right] = \alpha + \beta\mu_z + \Gamma v_{ij},$$

the variance of x_{ij} can be obtained as,

$$\mathrm{Var}(x_{ij}) = \beta\beta^T{\sigma_z}^2 + \Sigma = {\sigma_z}^2 \begin{pmatrix} {\beta_1}^2 & \beta_1\beta_2 & \dots & \beta_1\beta_m \\ \beta_2\beta_1 & {\beta_2}^2 & \dots & \beta_2\beta_m \\ \vdots & \vdots & \ddots & \vdots \\ \beta_m\beta_1 & \beta_m\beta_2 & \dots & {\beta_m}^2 \end{pmatrix} + \begin{pmatrix} {\sigma_1}^2 & 0 & \dots & 0 \\ 0 & {\sigma_2}^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & {\sigma_m}^2 \end{pmatrix}$$

The covariance of two lower-level units x_{ij} and x_{ih} $(j \neq h)$ can be obtained as,

$$\begin{aligned} \mathrm{Cov}(x_{ij}, x_{ih}) &= \mathbb{E}\left[(x_{ij} - \mathbb{E}(x_{ij}))(x_{ih} - \mathbb{E}(x_{ih}))^T\right] \\ &= \mathbb{E}\left[(\beta(z_i - \mu_z) + \varepsilon_{ij})(\beta(z_i - \mu_z) + \varepsilon_{ih})^T\right] \\ &= \beta\beta^T\mathbb{E}\left[(z_i - \mu_z)^2\right] + \beta\mathbb{E}\left[(z_i - \mu_z)\right]\mathbb{E}({\varepsilon_{ih}}^T) + \mathbb{E}(\varepsilon_{ij})\mathbb{E}\left[(z_i - \mu_z)\right]\beta^T + \mathbb{E}(\varepsilon_{ij})\mathbb{E}({\varepsilon_{ih}}^T) \\ &= \beta\beta^T\mathbb{E}\left[(z_i - \mu_z)^2\right], \end{aligned}$$

note that $\mathrm{Var}[(z_i - \mu_z)] = \mathbb{E}\left[(z_i - \mu_z)^2\right] - (\mathbb{E}\left[(z_i - \mu_z)\right])^2 = \mathbb{E}\left[(z_i - \mu_z)^2\right]$. By identifiability (details can be found at the end of Section 3.2), we have $z_i \sim (0, 1)$, that is $\mu_z = 0$ and $\sigma_z = 1$.

So the covariance for two different lower-level units within the same upper-level unit and from the same variable column can be written as,

$$\mathrm{Cov}(x_{ijl}, x_{ihl}) = {\sigma_z}^2{\beta_l}^2 = {\beta_l}^2,$$

and the variance can be written as,

$$\mathrm{Var}(x_{ijl}) = {\sigma_z}^2{\beta_l}^2 + {\sigma_l}^2 = {\beta_l}^2 + {\sigma_l}^2,$$

so the ICC can be written as,

$$\text{Corr}(x_{ijl}, x_{ihl}) = \frac{{\beta_l}^2}{{\beta_l}^2 + {\sigma_l}^2}, \tag{3.25}$$

where $l = 1, 2, \ldots, m$. For a simple random effect model $y_{ij} = \mu + z_i + \varepsilon_{ij}$, where $y_{ij} \in \mathbb{R}$, the variance of the random effect z_i is ${\sigma_z}^2$, and the variance of ε_{ij} is ${\sigma_\varepsilon}^2$, the intraclass correlation is written as $\frac{{\sigma_z}^2}{{\sigma_z}^2+{\sigma_\varepsilon}^2}$. Equation 3.25 is an extension. This has the same shape as the usual expression, just considering $\sigma_z = 1$ and including the squared lth diagonal element ${\beta_l}^2$ as the 'coefficient' of σ_z in our setup. It describes the correlation between two observations from the upper-level unit of the same outcome measurement.

The covariance of two different lower-level units within the same upper-level unit and from different variable column can be written as

$$\text{Cov}(x_{ijl}, x_{ihl'}) = {\sigma_z}^2 \beta_l \beta_{l'} = \beta_l \beta_{l'},$$

and the variance can be written as,

$$\text{Var}(x_{ijl}) = {\sigma_z}^2 {\beta_l}^2 + {\sigma_l}^2 = {\beta_l}^2 + {\sigma_l}^2,$$

$$\text{Var}(x_{ihl'}) = {\sigma_z}^2 {\beta_{l'}}^2 + {\sigma_{l'}}^2 = {\beta_{l'}}^2 + {\sigma_{l'}}^2,$$

so the ICC can be written as,

$$\text{Corr}(x_{ijl}, x_{ihl'}) = \frac{\beta_l \beta_{l'}}{\sqrt{({\beta_l}^2 + {\sigma_l}^2)({\beta_{l'}}^2 + {\sigma_{l'}}^2)}}, \tag{3.26}$$

where $l \neq l'$. By looking at the expression of Equation (3.26), we observe that Equation (3.25) is just a special case of the generalized form of ICC for the proposed multivariate two-level model. It describes the correlation of the two observations within the same upper-level units of different outcome measurements. When considering the same outcome measurements, Equation (3.26) can be written as Equation (3.25).

3.5 Simulations

In this section, we conduct simulations to evaluate the accuracy of parameter estimation. Additionally, we investigate whether an increase in the number of upper- or lower-level units will effectively reduce the variance of the parameter estimates. We also test the importance of choosing the number of mixture components, denoted as K. Finally, we run simulations to examine whether the random effect distribution impacts the estimation of parameter Γ.

3.5.1 Evaluate the Accuracy of Parameter Estimation

We first conduct a simulation study to examine the accuracy of our parameter estimation using the ECM algorithm. Another objective of this simulation is to investigate whether an increase in the number of upper- or lower-level units will effectively reduce the variance of the parameter estimates. We here simulate data from bivariate two-level scenarios with a single covariate, where the number of mixture components is $K = 2$. We first consider a scenario with $r = 50$ upper level units and $n_i = 5$ lower level units, for $i = 1, 2, \ldots, r$. This will be the baseline experiment. Then we keep $r = 50$ unchanged and increase the number of lower-level units to be $n_i = 10$, for $i = 1, 2, \ldots, r$. We consider another sample size with lower-level units $n_i = 5$ for $i = 1, 2, \ldots, r$ unchanged but increase the upper-level units to be $r = 100$. We also further increase the upper level units to be $r = 200$ and keep the lower level units $n_i = 5$ for $i = 1, 2, \ldots, r$. We generate 200 replicated data sets (each with two mixture components, $\pi_1 = 0.4$, $\pi_2 = 0.6$ and true values of z_k's as shown in the first column of Table 3.1) from the model (3.7). In all four scenarios a lower level covariate is generated from a normal distribution with mean 0.3 and standard deviation 0.2, and with true $\gamma = (1, 3)^T$. The simulation results, which are presented in Table 3.1, Table 3.2 and Figure 3.2, indicate that the true parameters are well estimated, and when we increase the number of upper-level units, the parameters' RMSE decreases stronger than when increasing the number of lower-level units.

We also compare the γ estimates from our model to those obtained by fitting individual two-level models. Each of these models uses one of the simulated two-dimensional variables as response variable and treats the covariate as predictor. We used the `lmer()` function in R package **lme4** and the `allvc()` function from the **npmlreg** package for this comparison. The

results, displayed in Table 3.3 and Table 3.4, show that our method produces sensible results when compared to those obtained with `allvc()` and even superior estimates when compared to those obtained with `lmer()`.

Table 3.1: Estimates of key parameters γ, z_k and α with different numbers of upper-level and lower-level units.

		Average estimates			
	True	$r = 50, n_i = 5$	$r = 50, n_i = 10$	$r = 100, n_i = 5$	$r = 200, n_i = 5$
γ_1	1.000	1.033	0.981	0.990	0.997
γ_2	3.000	3.031	3.034	2.993	3.004
z_1	-0.816	-0.804	-0.815	-0.818	-0.814
z_2	1.225	1.279	1.256	1.236	1.235
α_1	2.000	1.986	2.041	2.022	1.990
α_2	10.000	9.995	10.021	10.007	10.001

Table 3.2: RMSE for key parameters γ, z_k and α with different numbers of upper-level and lower-level units.

	RMSE			
	$r = 50, n_i = 5$	$r = 50, n_i = 10$	$r = 100, n_i = 5$	$r = 200, n_i = 5$
γ_1	0.278	0.157	0.166	0.111
γ_2	0.441	0.284	0.263	0.201
z_1	0.130	0.125	0.082	0.057
z_2	0.233	0.200	0.129	0.087
α_1	0.455	0.429	0.310	0.213
α_2	0.179	0.157	0.116	0.077

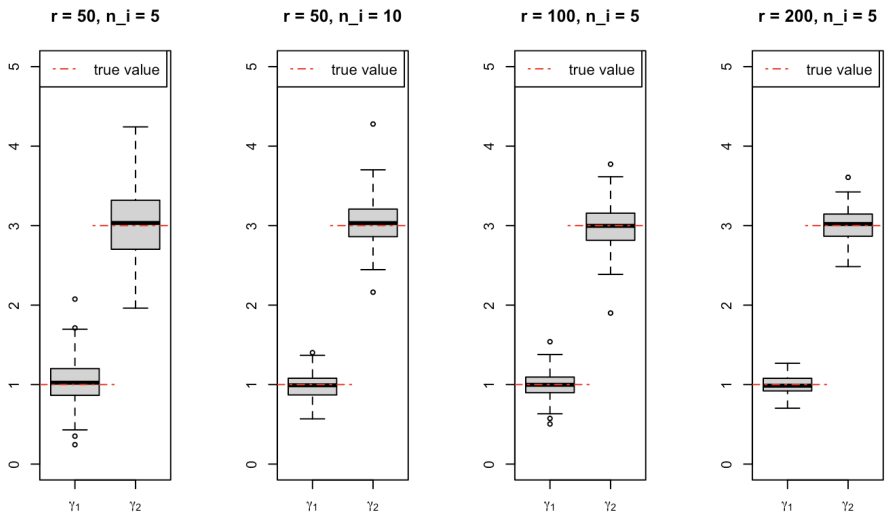

Figure 3.2: Estimates of key parameter γ with different number of upper-level and lower-level units.

Table 3.3: Averaged estimates of γ obtained by fitting individual models.

	Average estimates				
	True	$r = 50, n_i = 5$	$r = 50, n_i = 10$	$r = 100, n_i = 5$	$r = 200, n_i = 5$
			`lmer()`		
γ_1	1.000	0.999	0.987	0.989	0.996
γ_2	3.000	2.972	3.037	3.002	2.999
			`allvc()`		
γ_1	1.000	0.993	0.992	0.989	0.995
γ_2	3.000	2.998	3.037	3.005	2.995

Table 3.4: RMSE for γ obtained by fitting individual models.

	RMSE			
	$r = 50, n_i = 5$	$r = 50, n_i = 10$	$r = 100, n_i = 5$	$r = 200, n_i = 5$
			`lmer()`	
γ_1	0.286	0.182	0.175	0.123
γ_2	0.470	0.325	0.278	0.209
			`allvc()`	
γ_1	0.259	0.167	0.166	0.115
γ_2	0.396	0.284	0.263	0.191

3.5.2 Impact of the Number of Mixture Components

The number of mass points used in a fitted model can be decided through the model selection process. In practice, one can select the value of K that yields the smallest AIC value. However, it is important to note that the choice of K could have an impact on the posterior probabilities and estimations of other parameters. On the other hand, it is important to understand whether the estimation of the γ parameter remains consistent and unaffected by the number of components. To investigate this problem, we set up the following simulation scenario. We generate 200 replicates from model (3.7), employing two mixture components, i.e. $K_{true} = 2$ and allowing one covariate. The covariate was randomly simulated from a normal distribution with mean 0.3 and standard deviation 0.2. Each simulated data set is bivariate with 50 upper-level units and 5 lower-level units within each upper-level unit. Subsequently, we fit the simulated data using three different models: $K = 2$, $K = 3$, and $K = 4$. The results are shown in Table 3.5 and Figure 3.3. It can be seen that the estimation of γ in our method is quite robust to the choice of K.

Table 3.5: Estimates of parameter $\gamma = (\gamma_1, \gamma_2)^T$ with different K.

		Average estimates		
	True	$K = 2$	$K = 3$	$K = 4$
γ_1	0.500	0.519	0.517	0.519
γ_2	2.000	2.082	2.074	2.078

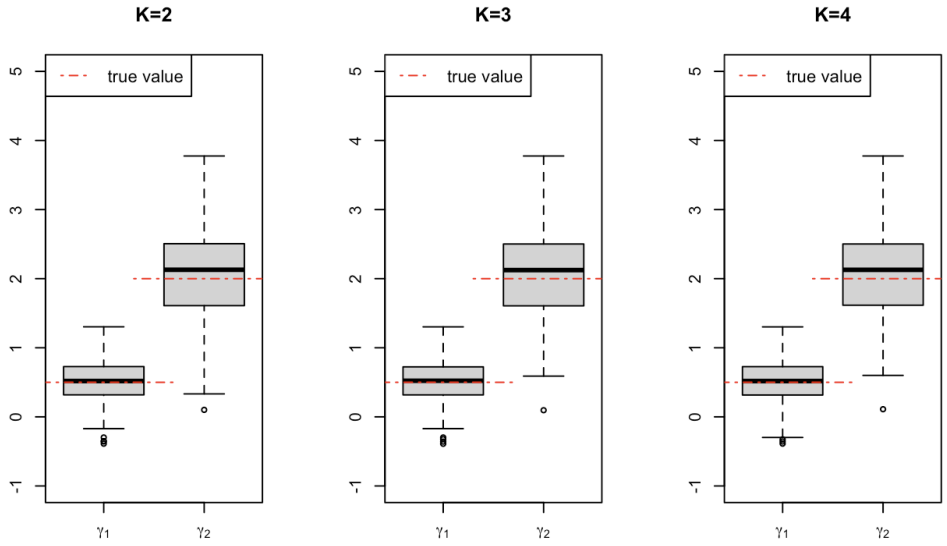

Figure 3.3: Boxplots for the estimated parameter γ with different K.

3.5.3 Impact of the Random Effect Distribution

We conduct another simulation study to investigate whether the random effect distribution impacts the estimation of parameters Γ (here $\Gamma = \gamma$). In our methodology we have explained that there is no explicit assumption on the distribution of Z, since the mixture approximation of Z plays merely the operational role of facilitating the approximation of an integral. However, one may argue that de facto (3.6) then becomes the model, and hence the discrete mixture property of the random effects an implicit assumption. So we wish to check the robustness of different distributional choices for the random effect in comparison with the discrete mixture we used to fit the models. In addition to generating the random effect z_i from a discrete distribution with finite mass points and weights (we use $\pi_1 = 0.4$, $\pi_2 = 0.6$, $z_1 = 1.225$ and $z_2 = -0.816$), we consider several distributions for the random variable Z generating the z_i: Poisson distribution, Gamma distribution, Normal distribution and three versions of a Gaussian mixture distribution. We generate the random variable Z from each of the above distributions, and for each scenario, we generate 200 replicated data sets from model (2.1) with one covariate (the covariate generated from a normal distribution with a mean of 0.3

and a standard deviation of 0.2), with $r = 50$ upper-level units and $n_i = 5$ lower-level units. The random effect generated from these four distributions has lower-level units within each upper-level unit that share the same z_i value.

Table 3.6 and Table 3.7 report the averaged estimates of the covariate coefficient $\hat{\gamma}$. Figure 3.4 and Figure 3.5 show the boxplots of the estimated $\hat{\gamma}$'s with Z from different distributions. One can see that the distribution of the random effect only slightly effects the estimation of the γ's. This is in line with the literature as the estimates of fixed effect parameters are generally robust to the random effect distribution (Drikvandi et al., 2017), (Drikvandi, 2020). Intuitively, an advantage of our methodology with a mixture discrete distribution for random effects is when there is a clustering structure along the latent space in the data, as it allows the random effects to capture such variability, which subsequently helps reduce the standard errors of estimates.

Table 3.6: Average estimated $\bar{\gamma}$ from simulated data with different distributions of Z.

	γ_{true}	
	0.500	2.000
Distribution of Z	$\hat{\gamma}_1$	$\hat{\gamma}_2$
Discrete mixture ($K = 2$)	0.519	2.082
Poisson ($\lambda = 0.2$)	0.440	1.810
Gamma ($\alpha = 2$, $\beta = 1.5$)	0.575	2.265
Normal ($\mu = 0$, $\sigma = 1$)	0.494	2.021

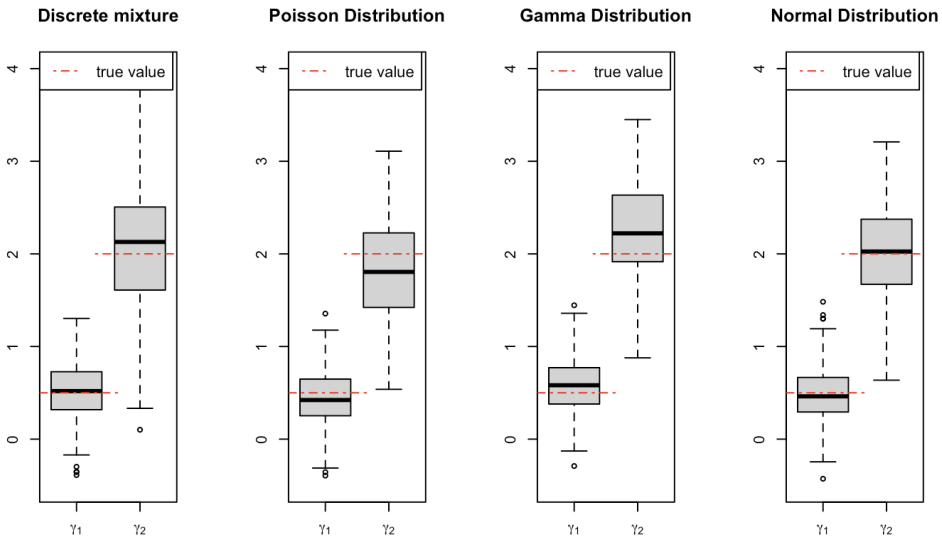

Figure 3.4: Boxplots (corresponding to Table 3.6) for the estimated parameter $\hat{\gamma}$ from simulated data in which Z is from different distributions.

Table 3.7: Average estimated $\bar{\gamma}$ from simulated data with Z generated from mixtures of Gaussians with $\mu = (1.225, -0.816)$ and $\pi = (0.4, 0.6)$, each mixture with a different set of σ's.

	γ_{true}	
	0.500	2.000
Distribution of Z	$\hat{\gamma}_1$	$\hat{\gamma}_2$
Mixture of Gaussians ($\sigma = (1, 1.5)$)	0.459	1.843
Mixture of Gaussians ($\sigma = (0.2, 0.5)$)	0.409	1.700
Mixture of Gaussians ($\sigma = (0.1, 0.2)$)	0.451	1.823
Discrete mixture ($K = 2$)	0.519	2.082

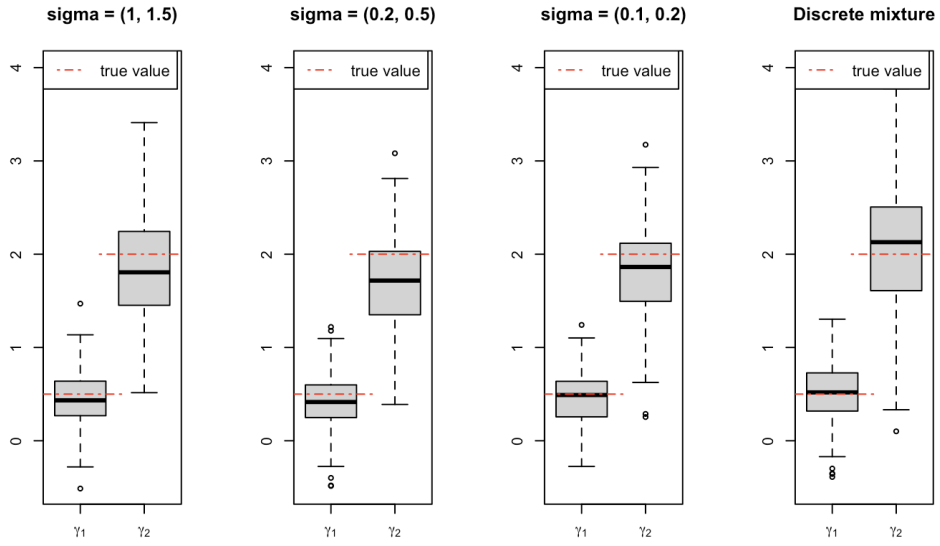

Figure 3.5: Boxplots (corresponding to Table 3.7) for the estimated parameter $\hat{\gamma}$ from simulated data in which Z is from different mixtures of Gaussians.

3.6 Additional Inferential Aspects for the Two-level Model

3.6.1 Clustering

For two-level models, the clustering always operates on the upper-level units. The MAP rule for the two-level model is similar to the MAP rule for the one-level model. Given the availability of w_{ik} from the last iteration of the ECM algorithm, which correspond to the posterior probability of upper-level unit i belonging to component k, upper-level unit $x_i, i = 1, 2, \ldots, r$, is then classified to the cluster $\hat{k}(x_i)$ to which it belongs with highest posterior probability,

$$\hat{k}(x_i) = \arg\max_k w_{ik}.$$

Based on the MAP rule, we have come up with another way of classifying the observa-

tions. We call it the robust clustering rule (at the 95% confidence level), where the significance can take on various values.

$$\hat{k}(i) = \begin{cases} \arg\max_k w_{ik}, & \text{if } \arg\max_k w_{ik} > 0.95 \\ \text{uncertain group}, & \text{otherwise} \end{cases}$$

Examples for clustering using both of these clustering techniques can be found in Section 3.7.2 and Section 3.7.3.

3.6.2 Ranking

As we have described in detail when we were talking about the one-level model, and as we mentioned in the illustration plot (Figure 3.1) of the two-level model, we use model-based scores z_i^* for ranking purposes in constructing the league tables,

$$z_i^* = \sum_{k=1}^{K} w_{ik} \hat{z}_k,$$

where the z_i^* is a common random effect shared by the lower-level units within the same upper-level unit. For example, in the import and export data introduced in Section 1.2.6, z_i^* is a country-specific random effect.

3.6.3 Bootstrapped Standard Error

When using the proposed multivariate response models (for both the one-level model and the two-level model), no analytic calculation of the standard errors is possible. Therefore, we propose the following bootstrapped algorithm to obtain the standard errors for the covariate coefficients. (This bootstrapped standard error algorithm for the two-level model is similar to the one used in the previous section for the one-level model.) The bootstrap process to obtain the standard error is carried out with the following steps:

(i) We are given a data set $x_{ij} \in \mathbb{R}^m$ and a covariate vector $v_{ij} \in \mathbb{R}^p$, $i = 1, \ldots, r$, $j = 1, \ldots, n_i$. (Note that for multilevel data, there are two types of covariate: the lower-level covariate which is the covariate collected on the lower levels, and the upper-level covariate which is the covariate collected on the upper levels.)

(ii) Fit the data x_{ij}, v_{ij} to model (3.7) to obtain the estimates of the parameters.

(iii) Sampling B data sets from model (3.7) with the estimated parameters obtained from (ii). (Note that in model 3.7, all the common random effect z_i, shared by the lower-level units within the same upper-level unit, is replaced by K mass points z_k, which are randomly chosen with probability π_k. The only part related to different levels is the covariate term Γv_{ij}, indicating that the simulated data, in fact, originates from the lower-level.)

(iv) Fit these B data sets to our model and we would obtain B sets of $\hat{\gamma}$. Then calculate the standard deviations across all B replicates of each of the $m \times p$ components of $\hat{\Gamma}$.

Examples of bootstrapped standard errors can be found in Section 3.7.1.

3.7 Applications

In this section, we analyze the real datasets from the case studies introduced in Section 1.2. We focus on regression in the first case study and on classification in the second case study, while in the third case study both regression and classification are of interest.

3.7.1 Twins Data

For our analysis of the twins data set, we consider the two types of touches, self touch and other touch, as a bivariate response variable, and include the three mental health variables as covariates into model (3.2). Under this model, the observations within upper-levels (we consider each mother as a upper-level unit) share a common, mother-specific, random effect z_i, which accounts for correlated touch behaviour of fetuses from the same mother. Notably there is only one such random effect variable, which applies to both response variables.

An examination of the AIC values across different values of K yields that the minimum AIC is attained for $K = 2$, with AIC value 428.6266, and hence we use this choice of K for our analysis. The traditional method of dealing with such a structured data would be fitting separate two-level models, each using one of the touch movements as the response variable and the three mental health measurements as covariates. Table 3.8 shows the estimates of the coefficients and their standard errors obtained through using the `lmer()` function in R

package **lme4** (Bates et al., 2015), and Table 3.9 shows the estimates from our model and the bootstrapped standard errors. (Note that the bootstrap applied here is a straightforward extension of the bootstrap technique developed by Zhang and Einbeck (2024d), adjusted to the context of the two-level models, ensuring that all units on the upper level get associated with the same random effect; the bootstrap process is written in detail in Section 3.6.3.) Our approach gives estimates similar to the linear model but with the main advantage of reduced standard errors of parameter estimates.

Table 3.8: For the twins data, estimations of γ obtained using individual two-level models (`lmer()`) for self touch and other touch as response and depression, perceived stress scale (PSS) and anxiety as predictors, with standard errors given in brackets.

	indiv. two-level models		
	depression	stress	anxiety
self touch	-27.34 (39.18)	11.31 (13.81)	-11.46 (24.89)
other touch	-92.70 (49.86)	55.62 (25.55)	-60.30 (38.76)

Table 3.9: For the twins data, estimations of γ obtained using the proposed multivariate response model with random effect. Standard errors (in the first brackets) are obtained via the bootstrap. Note that the model is fitted once using a bivariate response variable and three covariates simultaneously with a mother-specific random effect.

	multivariate response model		
	depression	stress	anxiety
self touch	-26.82 (37.97)	12.10 (13.30)	-7.12 (23.48)
other touch	-83.43 (49.25)	46.82 (15.70)	-73.72 (27.79)

3.7.2 Import and Export Data

We consider a data set concerning trade in goods and services, or more specifically the transactions in goods and services between residents and non-residents, measured in million USD. The data is extracted from the OECD website (Organisation for Economic Co-operation and Development, 2023b). The variables are given as the percentage of imports and exports in relation to the overall GDP. The dataset comprises data from 44 countries, and for our analysis, we specifically selected the time period between 2018 and 2022, during which a varying number of observations is available for different countries. Specifically, Australia, Japan, Korea, Mexico, New Zealand, Turkey, United States, China, and Colombia have four observations each,

while India, Russia, and Brazil have three observations each. The remaining countries have five observations each. We are interested in clustering the data with respect to their overall export/import activity relative to GDP size.

In our analysis of this data, the logs of imports and exports constitute a bivarate response variable, with $r = 44$ countries defining the upper level, and $n_i \in \{3, 4, 5\}, i = 1, \ldots, r$. This is a two-level scenario with 44 countries on the upper level, and with three to five repeated measurements each.

Fitting a bivariate response model of type (3.2) with a country-specific random effect, but without covariate, with $K = 4$ mass points leads to an AIC value of 117.8696, which is the smallest AIC value one can get across all $K \geq 1$. For each country, we obtain the posterior probabilities w_{ik} according to (3.19), an excerpt of the full matrix $(w_{ik})_{1 \leq i \leq r, 1 \leq k \leq K}$ is given in Table 3.10. The countries are ordered in this table by their posterior intercepts $z_i^* = \sum_{k=1}^{K} w_{ik} \hat{z}_k$, with smaller values corresponding to smaller import/export volume relative to GDP. We can think of this column as representing predicted values of a latent variable which we could describe as 'international trade volume per GDP'. So, according to this sort of linearized view on the problem, Luxembourg shows the largest trade volume per GDP, and the US the smallest.

A sensible way of clustering the observations is to follow the MAP rule, i.e. each upper-level unit (country) i is assigned to the cluster k to which it belongs with the largest probability w_{ik}. We can see from Table 3.10 that, according to this rule, Luxembourg is the only country assigned to its high-volume mass point. The second-largest mass point encompasses a wide range of countries ranging from Germany and Sweden to Ireland, followed by the second smallest mass point featuring countries from Israel to Iceland. The mass point corresponding to the smallest trading volume per GDP comprises of 10 countries, most of which very large countries including Australia and all BRIC countries. Figure 3.6 top left provides a graphical representation of this clustering approach, with observations colored by MAP classifications. Note that three to five observations correspond to each country.

The second mass point has smaller variance compared to the first and third mass points (see Table 3.11), and all countries allocated to this cluster according to the MAP rule share some probability mass with the third cluster (but not all countries in the third cluster share

probability mass with the second). There is no obvious characteristic distinguishing these clusters, even though countries in the third cluster tend to be smaller in size, especially those that have no or little probability mass shared with the second cluster.

We note that neither the ranking (by posterior intercepts) nor the clustering (by the MAP rule) gives a clear evidence on how *well* two countries, or two clusters, can actually be distinguished. However, the available posterior probabilities help us to provide a principled way of doing so. If the largest posterior probability of the observation exceeds a certain level of confidence, say 0.95, it is clustered into that specific cluster with 95% confidence. That is, it can be robustly distinguished from observations (countries) that are allocated to *other* mass points with the same level confidence. So, according to this criterion, we can produce a 'robust' clustering of countries, which is illustrated in Figure 3.6 top right. For example, Luxembourg is classified as part of the highest mass point 4 with a probability of 1, and it can be reliably distinguished from countries in mass point 3, such as Austria, Poland, ..., Ireland. Conversely, all countries for which the largest posterior probability of an observation is less than 0.95 are considered an uncertain observation that does not belong to any specific mass point, colored as gray points in Figure 3.6 top right. This specifically concerns the countries of United Kingdom, Turkey and France with their largest probabilities being below 0.95. At this level of confidence, the second mass point is eradicated entirely. However, changing the confidence level to 90% allows a robust clustering of United Kingdom, Turkey and France to that mass point, as shown in Figure 3.6 bottom. We can see that the robust clustering of observations into 'confidence-adaptive' clusters would depend on the level of confidence. It should be pointed out that, beyond the conclusions immediately drawn from this cluster-based approach, some more conclusions could be drawn with careful reasoning: for instance, even under a 95% level of confidence, all countries between the United Kingdom and Greece can in fact be robustly distinguished from both mass points 1 and 4, as they feature $> 95\%$ probability mass between them. They just cannot be robustly distinguished between mass points 2 and 3, at that level.

As an alternative approach to robustly distinguish cluster-level units based on the posterior information derived from the EM algorithm, Einbeck et al. (2017) suggested a method for measuring the uncertainty of posterior intercepts and probabilities based on an analytical

approach and an NPML-bootstrap process.

Table 3.11: Estimated σ_{lk}, where $l = 1, 2$ and $k = 1, \ldots, 4$ for the fitted model in Section 3.7.3.

	Mass points			
k	1	2	3	4
$\hat{\sigma}_{1k}$	0.204	0.177	0.329	0.039
$\hat{\sigma}_{2k}$	0.285	0.172	0.347	0.036

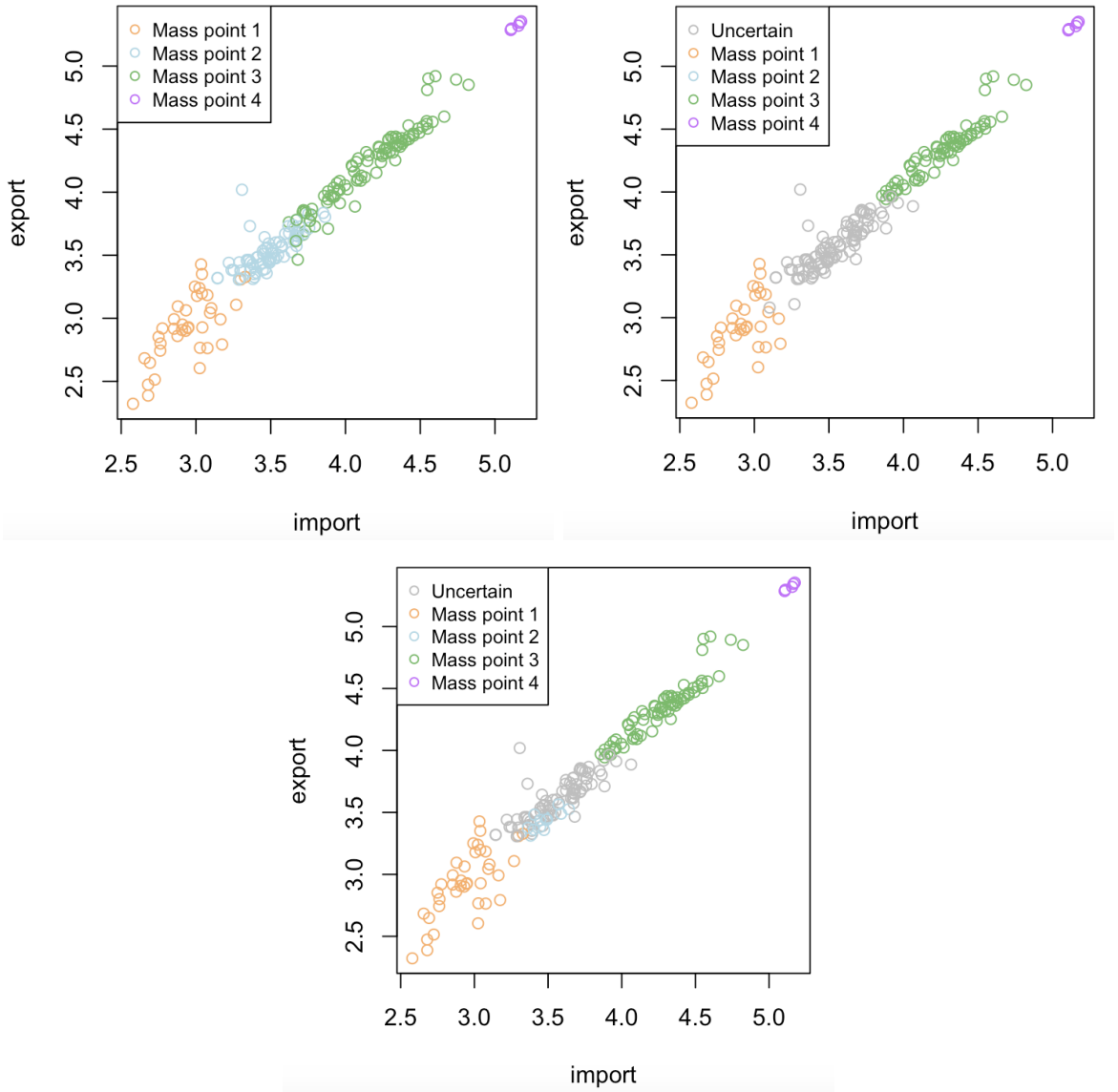

Figure 3.6: Clustering of the imports and exports data; top left using the Maximum a posterior (MAP) rule; top right with 95% confidence; bottom with 90% confidence.

3.7.3 PIAAC Data

The data considered in this section is from the Programme for the International Assessment of Adult Competencies (PIAAC) survey of adult skills, carried out in 2011 and 2012 by the OECD.

We now analyse the PIAAC data set, where Literacy, Numeracy, and Problem solving constitute a three-variate response (all three skill types are provided on a continuous scale ranging from 0 to 500), and gender and employment status serve as two covariates. Again this is a two-level model, with 28 countries and two sub-national regions defining the upper levels. The lower levels are defined through the different combinations of the covariate factor levels within each country; i.e. there are four lower-level 'observations' for each country corresponding to the average score for this combination of covariates.

We fit model (3.2) with $K = 4$ mass points, which leads to a minimum AIC value of 711.702. Posterior intercepts can again be obtained through the use of $z_i^* = \sum_{k=1}^{K} w_{ik}\hat{z}_k$, with posterior probabilities w_{ik} according to (3.19). These posterior intercepts can be seen as the summary information for each country, providing the residual performance after the covariates have been taken into account. The role of the covariates is to 'take out' the effects of such variables in the clustering process. The estimates of the covariate coefficients and the bootstrapped standard errors are shown in Table 3.12. Coverage properties of 95% confidence intervals arising from these standard errors are also reported in Table 3.12. The results show how gender (male = 1, female = 0) and employment status (employee = 1, self-employed = 0) relate to literacy, numeracy, and problem-solving skills. For example, it indicates that employees have expected problem-solving scores that are 6.056 higher than for self-employed. Providing the z_i^* in rank order results in a league table, shown in Table 3.13. The posterior probabilities obtained at the convergence of the ECM algorithm are also given in this table, and can be used for classification of countries according to their skill levels.

We can distinguish two countries in terms of their cluster membership if they fall with 95% confidence into two different mass points. In Table 3.13, Chile, Mexico and Turkey are assigned to the worst performance mass point 1 with probabilities exceeding 0.95. Consequently, they can be robustly distinguished from all countries starting from Greece (allocated to mass point 2) up to New Zealand (in the best 'performance' mass point 3) with 95% confidence.

Slovenia is the only country that is allocated to mass point 3 with 95% confidence, it can be robustly concluded to have performed better than Greece which is allocated to mass point 2 with 95% confidence. The largest two component probabilities of England(UK) and Israel spread across two mass points, so that the membership of these countries in either mass point 2 or mass point 3 cannot be definitively determined. All what we can say is that there is at least 80% confidence that Ireland and England(UK) do belong to mass point 3, thus they could at this (low) confidence level be distinguished from countries in mass point 2. Note that it is not possible to determine a comparative ranking among countries belonging to the same mass point, for example, we cannot say that Mexico has performed better than Chile. We cannot even robustly conclude that the England(UK) has performed better than United States, since with more than 80% confidence the two countries belong to the same mass point.

A somewhat similar analysis (using Stata) was carried out by Grilli et al. (2016) using data from the TIMSS&PIRLS database. Their multivariate approach jointly considers educational achievement in Reading, Mathematics and Science, where the coefficients for each response were estimated separately and combined using multiple imputation formulas. However, they did not consider the ranking problem, and their approach cannot be used for clustering purposes.

3.7.4 IALS Data

We again consider the Literacy Survey Data introduced in the beginning in Section 1.2.3 and used in Section 2.10.3 as an example to illustrate how the proposed one-level model can be applied in dealing with the joint ranking of multiple continuous variables (via the posterior random effect) with a view to constructing league tables. Previously, only the prose measurement was used and split to form a bivariate variable. Now, we analyze the data considering the 3-variate response Prose, Document, and Quantitative, additionally including the lower-level covariate gender in the model; i.e. $m = 3$, $p = 1$ and $\Gamma = \gamma \in \mathbb{R}^3$. The country-specific random effect z_i accounts for the correlation among the observations within upper-level units and the correlation among the three response dimensions of the model. We fit the model with $K = 4$ mass points and component-specific diagonal variances Σ_k, leading to an AIC value of 235.5 which does not drop significantly when increasing K further or with other variance

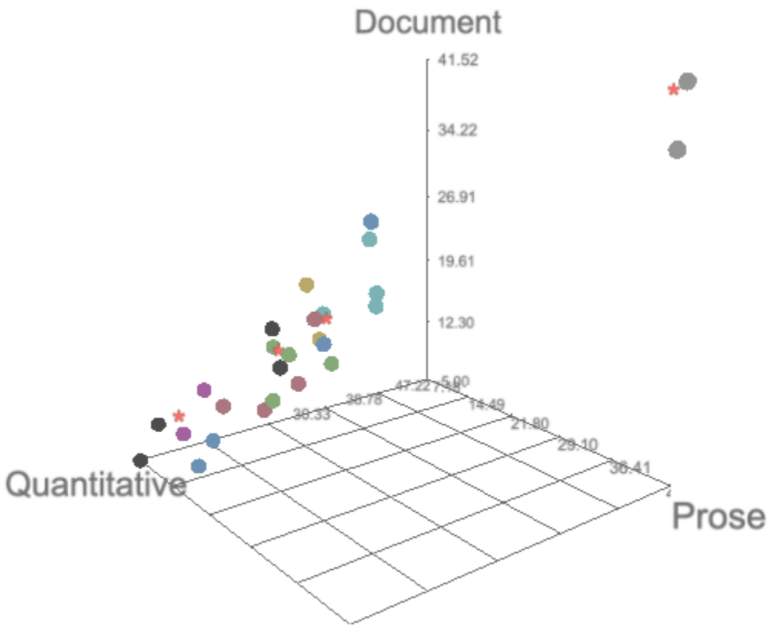

Figure 3.7: IALS Data with 3 variables

parametrizations. Figure 3.7 shows the scatterplot of these 3 variables with the mixture centers labeled with a red star. Table 3.14 presents the joint ranking via the posterior random effect and classification of the countries. The table shows that Sweden, Germany, and the Netherlands are assigned to mass point 1 with the smallest number of people being illiterate. Poland is the only country that is assigned to the high illiteracy mass point 4. The US and Ireland have posterior probabilities that spread across two mass points but are assigned to different components. Using all three measurements as a multivariate response, the component allocation of each country is more decisive compared to the results (Table 2.13) using just Prose.

3.8 Level reduction

The one-level model can be considered as a particular type of multi-level (i.e. here, two-level) model, with the upper level corresponding to observations x_i and the lower level to 'measurements' x_{ij} on the 'repeated responses'. As demonstrated in the example in Section 2.10.3, the shared random effect on the 'upper level' is directly obtained from the inferential framework without resorting to two-level ('variance component') modelling in a traditional sense. Spinning this thought further, the present methodology (both the one-level model and the two-level model) allows for a reduction of the number of levels in a genuine multilevel

scenario. For instance, assume one has repeated measures of some quantity taken on the left and right ear of some individuals over time (a detailed analysis of this data can be found below). Then, rather than fitting a three-level model, the two ears could define the axes of a bivariate response model, reducing the problem to a two-level model. By employing a multivariate response model, one can effectively 'take one level out'.

We have two examples from public health research and one education data to illustrate such situations.

The first example is the Visual impairment data from the Baltimore Eye Survey (Tielsch et al., 1989). The data described by Fahrmeir et al. (1994) and Liang et al. (1992) was collected to research on the effect of race and age on visual impairment. The repeated examinations were conducted on over 5000 people aged above 40. The left eye and the right eye from the same individuals are recorded as a binary variable (whether or not an eye was visually impaired) forming the response variables. The traditional way of dealing with such a data will require a two-level model, with the individuals as the upper level, and perhaps 'left eye' or 'right eye' as a binary covariate. But with our model, we would have a bivariate response for the left and the right eye, and our single random effect will take care of correlating the individuals, that is each 2-dimensional observation will always relate a single predicted random effect on the latent axis. So, we have effectively reduced a two-level to a one-level model. Further covariates can then of course be added.

The example 2 we consider the Hearing data, which is the from an ongoing multidisciplinary observation study (Shock, 1984), the Baltimore Longitudinal Study of Aging (BLSA). The study collects the hearing threshold sound pressure level at 11 different frequencies on both ears from the same individuals. The analysis in Verbeke et al. (1997) only uses two levels as they fit each 'ear' separately, see Figure (3.8). In principle the analysis of this data requires three levels: Individuals, time, and ear (left and right). With our proposed two-level model, we could fit both ears simultaneously with a two-level model, not requiring three levels.

Unfortunately, we were unable to access the above two data sets. Nevertheless, this example demonstrates how our model could reduce the level when fitting a model.

In the third example, we consider again the The International Adult Literacy Survey

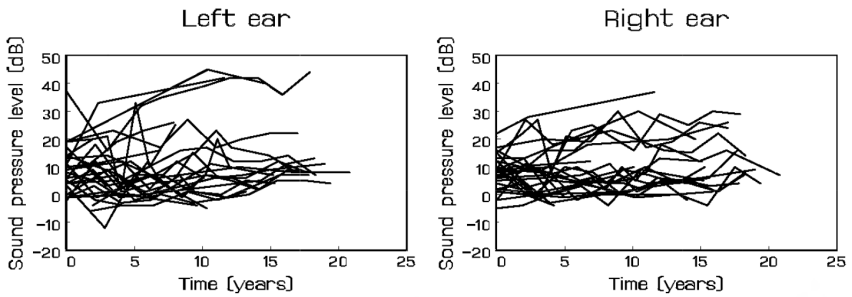

FIGURE 2.4. *Hearing Data. Individual profiles of 30 randomly selected subjects in the hearing data set, for the left and the right ear separately.*

Figure 3.8: Figure from Verbeke et al. (1997)

(IALS) data introduced in the introduction and previous application of clustering and ranking Section (2.10.3). The data was collected on Prose, Document and Quantitative scales, and we only considered the Prose measurement and take the separation of the reported prose results into male and female attainment differently into account than in that publication (Sofroniou et al., 2008), which include the gender variable as a covariate in a two-level model. By doing this, we are essentially taking one level out and fit a one-level model to the data.

Table 3.10: Classification and ranking for the trade and service data with $K = 4$. Posterior probabilities: $0.10 < p < 0.90$, $0.90 \leq p < 0.95$, $0.95 \leq p < 1$.

Country	posterior intercept		Mass points				MAP
		k	1	2	3	4	
		$\hat{\pi}_k$	0.236	0.310	0.431	0.023	
	z_i^*	$\hat{z}_k$	-1.402	-0.321	0.846	2.933	
United States	-1.402		1.000	0.000	0.000	0.000	1
Brazil	-1.401		1.000	0.000	0.000	0.000	1
Japan	-1.401		0.999	0.001	0.000	0.000	1
China	-1.401		0.999	0.001	0.000	0.000	1
India	-1.401		0.999	0.001	0.000	0.000	1
Colombia	-1.399		0.999	0.001	0.000	0.000	1
Indonesia	-1.378		0.979	0.021	0.000	0.000	1
Australia	-1.378		0.979	0.021	0.000	0.000	1
Russia	-1.373		0.974	0.025	0.001	0.000	1
New Zealand	-1.321		0.928	0.070	0.002	0.000	1
Israel	-0.574		0.260	0.715	0.025	0.000	2
South Africa	-0.383		0.099	0.862	0.039	0.000	2
Canada	-0.325		0.056	0.896	0.048	0.000	2
United Kingdom	-0.292		0.035	0.907	0.058	0.000	2
Turkey	-0.272		0.025	0.909	0.066	0.000	2
France	-0.251		0.018	0.906	0.076	0.000	2
Chile	-0.218		0.010	0.893	0.097	0.000	2
Italy	-0.211		0.009	0.889	0.102	0.000	2
Costa Rica	-0.156		0.004	0.850	0.146	0.000	2
Republic of Korea	-0.149		0.004	0.844	0.152	0.000	2
Spain	-0.127		0.003	0.828	0.169	0.000	2
Mexico	-0.080		0.002	0.790	0.208	0.000	2
Norway	-0.040		0.021	0.719	0.260	0.000	2
Finland	0.156		0.000	0.591	0.409	0.000	2
Iceland	0.180		0.000	0.570	0.430	0.000	2
Germany	0.357		0.000	0.418	0.582	0.000	3
Sweden	0.490		0.000	0.304	0.696	0.000	3
Portugal	0.492		0.000	0.303	0.697	0.000	3
Greece	0.614		0.000	0.198	0.802	0.000	3
Austria	0.790		0.000	0.047	0.953	0.000	3
Poland	0.806		0.000	0.034	0.966	0.000	3
Denmark	0.824		0.000	0.018	0.982	0.000	3
Switzerland	0.839		0.000	0.005	0.995	0.000	3
Latvia	0.843		0.000	0.002	0.998	0.000	3
Czech Republic	0.844		0.000	0.001	0.999	0.000	3
Estonia	0.845		0.000	0.000	1.000	0.000	3
Netherlands	0.845		0.000	0.000	1.000	0.000	3
⋮	⋮		⋮	⋮	⋮	⋮	⋮
Slovak Republic	0.846		0.000	0.000	1.000	0.000	3
Ireland	0.846		0.000	0.000	1.000	0.000	3
Luxembourg	2.933		0.000	0.000	0.000	1.000	4

Table 3.12: Estimates of covariate coefficients (matrix Γ) for the PIAAC data. Standard errors (in brackets) are obtained via the bootstrap. Coverage of 95% confidence intervals are reported in the squared brackets.

	Gender	Employment status
Literacy	-0.417 (1.336) [0.947]	2.683 (1.336) [0.949]
Numeracy	8.817 (1.458) [0.943]	-0.517 (1.427) [0.956]
Problem solving	1.833 (1.163) [0.947]	6.056 (1.218) [0.950]

Table 3.13: Classification and ranking for the PIAAC data using model $x_{ij} = \alpha + \beta z_i + \Gamma v_{ij} + \varepsilon_{ij}$ with $K = 4$. Posterior probabilities: $0.05 < p < 0.10$, $0.10 < p < 0.90$, $0.90 \leq p < 0.95$, $0.95 \leq p < 1$.

		Mass points			
		0.100	0.115	0.275	0.510
Country	posterior intercept	-2.650	-0.634	-0.069	0.700
Chile	-2.650	1.000	0.000	0.000	0.000
Mexico	-2.650	1.000	0.000	0.000	0.000
Turkey	-2.650	1.000	0.000	0.000	0.000
Greece	-0.632	0.000	0.997	0.003	0.000
Spain	-0.603	0.000	0.945	0.055	0.000
Republic of Korea	-0.592	0.000	0.926	0.074	0.000
Italy	-0.296	0.000	0.403	0.597	0.000
United States	-0.100	0.000	0.066	0.927	0.007
Poland	-0.092	0.000	0.058	0.930	0.012
Slovenia	-0.065	0.000	0.018	0.963	0.019
Ireland	-0.010	0.000	0.013	0.902	0.085
France	-0.006	0.000	0.007	0.905	0.088
Israel	0.012	0.000	0.005	0.885	0.110
England(UK)	0.023	0.000	0.022	0.843	0.135
Denmark	0.398	0.000	0.000	0.392	0.608
Germany	0.451	0.000	0.000	0.323	0.667
Flanders(Belgium)	0.567	0.000	0.000	0.173	0.827
Norway	0.654	0.000	0.000	0.006	0.940
Czech Republic	0.659	0.000	0.000	0.054	0.946
Hungary	0.667	0.000	0.000	0.043	0.957
Austria	0.676	0.000	0.000	0.031	0.969
Australia	0.683	0.000	0.000	0.021	0.979
Estonia	0.686	0.000	0.000	0.018	0.982
Finland	0.689	0.000	0.000	0.014	0.986
Canada	0.692	0.000	0.000	0.010	0.990
Japan	0.696	0.000	0.000	0.005	0.995
Slovak Republic	0.696	0.000	0.000	0.005	0.995
Netherlands	0.699	0.000	0.000	0.002	0.998
Sweden	0.699	0.000	0.000	0.001	0.999
New Zealand	0.700	0.000	0.000	0.000	1.000

Table 3.14: Posterior probabilities and intercepts for the IALS data. In the column 'mass points', the first two rows give estimated $\hat{\pi}_k$ and $\hat{z}_k$.

Table 3.15: Posterior probabilities for the IALS data

		Mass points			
		0.2308	0.5391	0.1532	0.0769
Country	posterior intercept	-1.1576	-0.0819	0.5904	2.8703
Sweden	-1.15760	1.0000	0.0000	0.0000	0.0000
Germany	-1.15756	1.0000	0.0000	0.0000	0.0000
Netherlands	-1.15754	0.9999	0.0001	0.0000	0.0000
Canada	-0.08188	0.0000	1.0000	0.0000	0.0000
Australia	-0.08188	0.0000	1.0000	0.0000	0.0000
Switzerland(French)	-0.08188	0.0000	1.0000	0.0000	0.0000
New Zealand	-0.08173	0.0000	0.9998	0.0002	0.0000
Belgium(Flanders)	-0.08163	0.0000	0.9996	0.0004	0.0000
Switzerland(German)	-0.08114	0.0000	0.9989	0.0011	0.0000
United States	-0.08036	0.0000	0.9977	0.0023	0.0000
Ireland	0.58386	0.0000	0.0098	0.9902	0.0000
United Kingdom	0.58912	0.0000	0.0019	0.9981	0.0000
Poland	2.87028	0.0000	0.0000	0.0000	1.0000

Chapter 4

One-level Quadratic Model

4.1 Model and Estimations

In Section 2.1 we proposed a linear random effect model. To motivate the developments in this chapter, here again we apply model (2.1) on a real data set, namely the Mussels' muscles data (available in R package **dr**) introduced in Section 1.2.8. We focus on describing the data structure rather than analyzing the effects of predictors on the response variables through some regression models. Therefore, we only consider four variables: Shell height (denoted as H), Shell length (denoted as L), Shell mass (denoted as S) and Shell width (denoted as W) for our application. The projections of the fitted model onto 2-dimensional pairs plot are shown in Figure 4.1. We observe that this model captured the overall trend of the data. For the combinations involving shell height, shell width, and shell length, the relationships among them are linear. However, for the plots on the second row and the bottom right, it is obvious that the relationship between the two variables is non-linear. The model (2.1) proposed in Section 2.1 is limited by its linear nature, particularly in handling multivariate data structures with curvature. Given the existence of a non-linear latent structure in the example data, it becomes apparent that a non-linear model is needed to capture its curvature effectively. In this chapter we further extend our linear model to a non-linear model to allow approximating multivariate data with non-linear latent structures through the use of a smooth curve. (Note that the results of the mussels data fitted with the non-linear model can be found in Section 4.6.)

We describe the non-linear model first in a general framework. This can be written as,

$$x_i = \sum_{j=0}^{p} \alpha_j B_j(z_i) + \varepsilon_i, \tag{4.1}$$

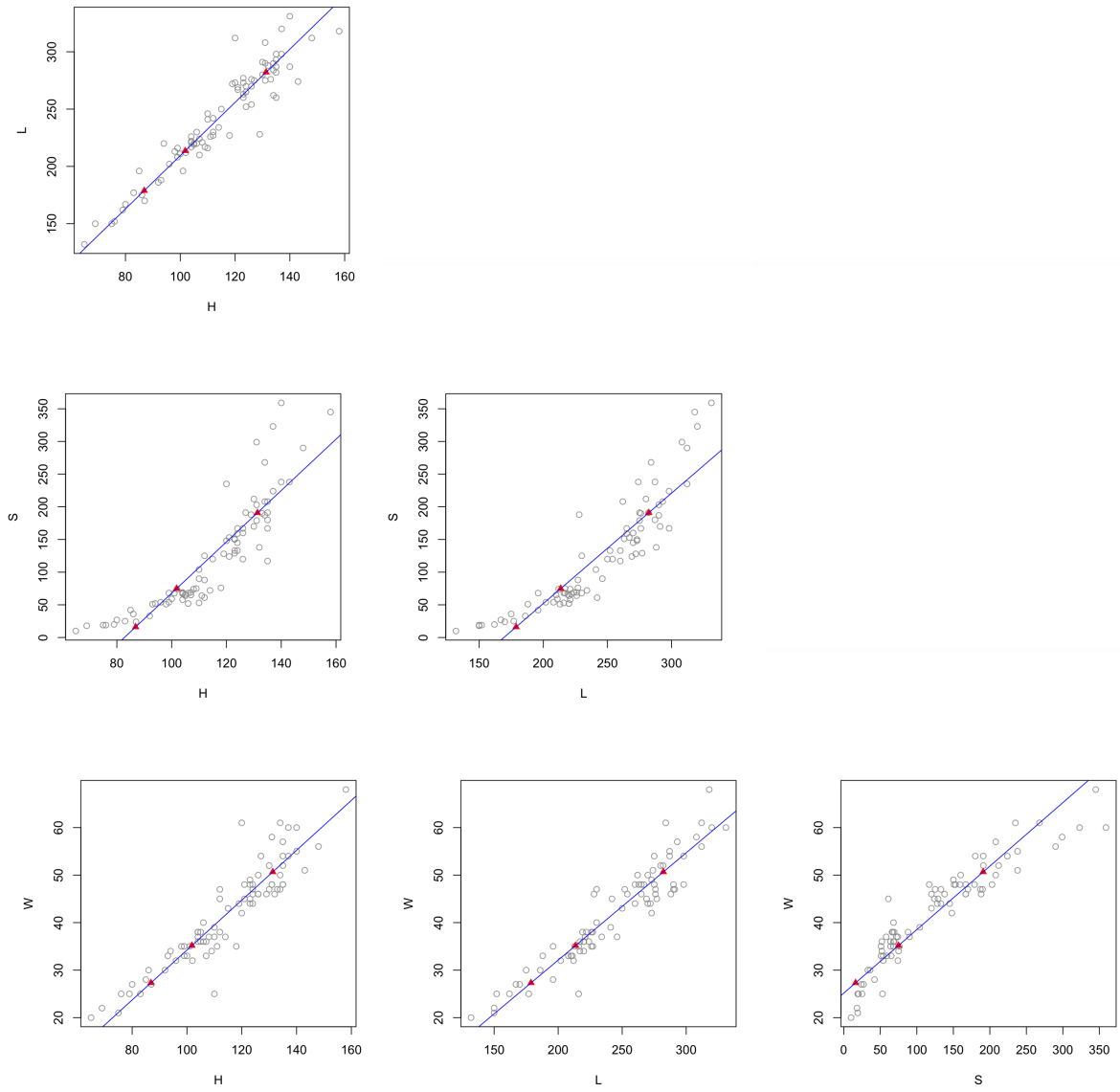

Figure 4.1: Mussels data with fitted model. Note that shell height is denoted as H, shell length is denoted as L, shell mass is denoted as S, and shell width is denoted as W.

where $B_j(z_i)$ can be regarded as any real-valued basis functions, such as B-splines, or polynomials, $x_i \in \mathbb{R}^m$, $z_i \in \mathbb{R}$ is a one-dimensional random effect and α_j is m-variate parameter vectors.

To examine the feasibility of this approach, we consider, in this exploratory chapter, a simple scenario involving a quadratic curve. The model can be written as the following,

$$x_i = \alpha + \beta z_i + \eta {z_i}^2 + \varepsilon_i, \tag{4.2}$$

where $B_j(z_i) = z_i^j$ is a polynomial, $x_i \in \mathbb{R}^m$, α, β and η are m-variate parameter vectors, z_i is a one-dimensional random effect, with $z_i \in \mathbb{R}$, and $\varepsilon_i \sim N(0, \Sigma(z_i))$ are independent Gaussian errors. If we compare model (4.2) to the full curve model (4.1), then it will be: $\alpha = \alpha_1$, $\beta = \alpha_2$, $\eta = \alpha_3$, $B_0(z_i) = 1$, $B_1(z_i) = z_i$, and $B_2(z_i) = z_i^2$. Similar to what we have developed in the one-level model, we could include covariates for multivariate response situations. Furthermore, we could have a two-level extension of the model so that we can deal with repeated measures. But here we will not develop such components further, and will instead only focus on the implementation of model (4.2).

The model (4.2) can be equivalently written as,

$$x_i | z_i, \alpha, \beta, \eta \sim N(\alpha + \beta z_i + \eta {z_i}^2, \Sigma(z_i)),$$

Again equivalently, and for later reference, we can write the conditional probability density function of the x_i as,

$$f(x_i | z_i, \alpha, \beta, \eta) = (2\pi)^{-m/2} |\Sigma(z_i)|^{-1/2} \exp\left\{ -\frac{1}{2} (x_i - \alpha - \beta z_i - \eta {z_i}^2)^T \Sigma^{-1}(z_i)(x_{ij} - \alpha - \beta z_i - \eta {z_i}^2) \right\}.$$

The marginal probability density function $f(x_i | z_i, \alpha, \beta, \eta)$ for observations generated from model (4.2) can be written as,

$$f(x_i | \alpha, \beta, \eta) = \int f(x_i, z_i | \alpha, \beta, \eta) dz_i = \int f(x_i | z_i, \alpha, \beta, \eta) \phi(z_i) dz_i,$$

where $f(x_i, z_i | \alpha, \beta, \eta)$ is the joint probability distribution of observed data x_i and unobserved random effects z_i, $\phi(\cdot)$ is the density function of the random effect distribution Z. Here the

distribution can be Gaussian or non-Gaussian, we don't make any explicit assumptions regarding the distribution Z of the z_i. If it is non-Gaussian, then it is difficult to find an analytical form for the marginal density function. So, again, we consider to use Nonparametric Maximum Likelihood approach, where the integral over z_i is replaced by a finite summation over K mass points $z_1, z_2, \cdots, z_k$ with their responding masses $\pi_1, \pi_2, \cdots, \pi_k$, where $k = 1, 2, \cdots, K$. Then model (4.2) can be rewritten as,

$$x_i|z_k, \alpha, \beta, \eta \sim N(\alpha + \beta z_k + \eta {z_k}^2, \Sigma(z_k)), \tag{4.3}$$

with probability π_k, and the marginal distribution can be approximated as,

$$f(x_i|\alpha, \beta, \eta) = \sum_{k=1}^{K} \pi_k f(x_i|z_k, \alpha, \beta, \eta),$$

To find the maximum likelihood estimates for the parameters, building on the marginal density, the likelihood of model (4.2) is the following,

$$l = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k f(x_i|z_k, \alpha, \beta, \eta),$$

and the expected complete log-likelihood of model (4.2) is the following,

$$\begin{aligned} l_c = &\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik} \log(\pi_k) + \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{1}{2} w_{ik} \log(|\Sigma_k|) + \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{m}{2} \log(2\pi) w_{ik} \\ &+ \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{1}{2} w_{ik} (x_i - \alpha - \beta z_k - \eta {z_k}^2)^T \Sigma_k^{-1} (x_i - \alpha - \beta z_k - \eta {z_k}^2). \end{aligned} \tag{4.4}$$

4.2 ECM Algorithm

We will use the ECM algorithm for parameter estimation, and the number of mixture components, denoted by K, will be selected based on the AIC and BIC criteria.

E-step

We update the posterior probability of observation i belonging to component k obtained using

Bayes' theorem.

$$w_{ik} = \frac{\pi_k f_{ik}}{\sum_l \pi_l f_{il}}. \tag{4.5}$$

M-step

The parameters π_k, α, β, η, and σ_j are updated in the M-step. The derivations for these parameter estimators are provided below, together with the corresponding estimators used in the M-step. Note that the derivation for parameters π_k, α, β, σ_j, and even η is similar to what we have done for the one-level and two-level models with covariates. In the calculations, the term $\eta {z_k}^2$ is just an additive term, like the covariate term, i.e. model (2.48) and (3.7) respectively. However, from a modelling point of view, the key difference here is that $\eta {z_k}^2$ is a random effect term, whereas Γv_i in the one-level model or Γv_{ij} in the two-level model is fixed effect term. The primary difference in the derivation process concerns z_k: in the quadratic model, we do not have an analytical form for the estimator of z_k. Instead, we estimate z_k by solving the cubic equation for z_k within the M-step. Details can be found below. A detailed procedure for the ECM algorithm is given in Algorithm 3 below.

4.3 Derivation for Parameter Estimators

Derivation for $\hat{\pi}_k$

The derivation of

$$\hat{\pi}_k = \frac{\sum_{i=1}^{n} w_{ik}}{n} \tag{4.6}$$

for the quadratic model is the same as that for $\hat{\pi}_k$ of the one-level model. Details can be found in Section 2.4.

Derivation for $\hat{z}_k$

For the derivation of z_k, let us first write the log-likelihood (4.4) in an indexed form,

$$l_{index} = constant + \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{m} -\frac{1}{2} w_{ik} \frac{(x_{ij} - \alpha_j - \beta_j z_k - \eta_j {z_k}^2)^2}{{\sigma_{jk}}^2},$$

then by taking partial derivative of the above log-likelihood l_{index} with respect to the parameter z_k, we obtain the following,

$$\frac{\partial l_{index}}{\partial z_k} = \sum_{i=1}^{n}\sum_{j=1}^{m} -\frac{1}{2} w_{ik}.2.\frac{(x_{ij} - \alpha_j - \beta_j z_k - \eta_j {z_k}^2)}{{\sigma_{jk}}^2}(-\beta_j - 2\eta_j z_k).$$

Again, we assume $\hat{\Sigma}_k \equiv \text{diag}(\sigma^2)$ in implementation, and letting the partial derivative equal to zero, then equation above can be rewritten as,

$$\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ik}(x_{ij} - \alpha_j - \beta_j z_k - \eta_j {z_k}^2)(\beta_j + 2\eta_j z_k) = 0,$$

now let us expend the above equation, and then we will obtain the following,

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{m}(w_{ik}x_{ij}\beta_j + 2\cdot w_{ik}x_{ij}\eta_j z_k - w_{ik}\alpha_j\beta_j - 2\cdot w_{ik}\alpha_j\eta_j z_k - w_{ik}{\beta_j}^2 z_k \\ -2\cdot w_{ik}\beta_j\eta_j {z_k}^2 - w_{ik}\eta_j\beta_j {z_k}^2 - 2\cdot w_{ik}{\eta_j}^2 {z_k}^3) = 0,\end{aligned} \tag{4.7}$$

then we organize it into the general cubic equation form,

$$\begin{aligned}(\sum_{i=1}^{n}\sum_{j=1}^{m} 2\cdot w_{ik}\hat{\eta}_j^2){z_k}^3 + 3\cdot(\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ik}\hat{\beta}_j\hat{\eta}_j){z_k}^2 \\ -(\sum_{i=1}^{n}\sum_{j=1}^{m} 2\cdot w_{ik}x_{ij}\hat{\eta}_j - \sum_{i=1}^{n}\sum_{j=1}^{m} 2\cdot w_{ik}\hat{\alpha}_j\hat{\eta}_j - \sum_{i=1}^{n}\sum_{j=1}^{m} w_{ik}\hat{\beta}_j^2)z_k \\ -(\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ik}x_{ij}\hat{\beta}_j - \sum_{i=1}^{n}\sum_{j=1}^{m} w_{ik}\hat{\alpha}_j\hat{\beta}_j) = 0,\end{aligned} \tag{4.8}$$

Unfortunately, obtaining an analytical form of z_k as we did in previous models is not possible in here. However, (in the M-step) we can still estimate z_k by solving the cubic equation (4.8) since each part of the coefficients is known in each iteration. It's important to note that the roots of a cubic equation fall into two scenarios: (i) one real root and two complex roots, in which case we consider the real root as the estimate of z_k, and (ii) three real roots, where we always choose the smallest absolute root in the implementation. In the R implementation, we use `polyroot()` function (available in **base** R package) to solve this cubic equation.

Derivation for $\hat{\beta}$

For the derivation of β, we use the result of the following, which is which is derived by Petersen and Pedersen (2012)

$$\frac{\partial}{\partial A}(x - As)^T W(x - As) = -2W(x - As)s^T$$

By taking partial derivative of the l_c with respect to β, we obtain,

$$\frac{\partial l_c}{\partial \beta} = \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{1}{2} w_{ik}(-2)\Sigma_k^{-1}(x_i - \alpha - \beta z_k - \eta {z_k}^2) z_k^T,$$

Since z_k is a scalar, $z_k = z_k^T$, and by letting the above equation to be zero and solving it,

$$\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\Sigma_k^{-1}(x_i - \alpha - \eta {z_k}^2) z_k - \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\Sigma_k^{-1}\beta z_k^2 = 0, \tag{4.9}$$

then,

$$\hat{\beta} = \left(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{\Sigma}_k^{-1}\hat{z}_k^2\right)^{-1}\left(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{\Sigma}_k^{-1}(x_i - \hat{\alpha} - \hat{\eta}\hat{z}_k^2)\hat{z}_k\right), \tag{4.10}$$

Again, we assume $\hat{\Sigma}_k \equiv \text{diag}(\sigma^2)$ in implementation, then Equation (4.9) can be rewritten as,

$$\Sigma^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(x_i - \alpha - \eta {z_k}^2) z_k - \Sigma^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\beta z_k^2 = 0,$$

multiply Σ on both sides, we could obtain,

$$\hat{\beta} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(x_i - \hat{\alpha} - \hat{\eta}\hat{z}_k^2)\hat{z}_k}{\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k^2}, \tag{4.11}$$

which is being used in the R code.

Derivation for $\hat{\alpha}$

Using the result (derived by Petersen and Pedersen (2012))of the derivatives of matrices, vectors, and scalars, where W is symmetric,

$$\frac{\partial}{\partial s}(x - s)^T W(x - s) = -2W(x - s)$$

We obtain the following by taking the partial derivative of the log-likelihood with respect to α,

$$\frac{\partial l_c}{\partial \alpha} = \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{1}{2} w_{ik}(-2)(\Sigma_k)^{-1}(x_i - \alpha - \beta z_k - \eta {z_k}^2),$$

then letting it to be zero and solving it,

$$\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\Sigma_k^{-1}(x_i - \alpha - \beta z_k - \eta {z_k}^2) = 0, \tag{4.12}$$

$$\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\Sigma_k^{-1}(x_i - \beta z_k - \eta {z_k}^2) = \alpha \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\Sigma_k^{-1},$$

we obtain the estimator for α,

$$\hat{\alpha} = \left(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{\Sigma}_k^{-1}\right)^{-1} \left(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{\Sigma}_k^{-1}(x_i - \hat{\beta}\hat{z}_k - \hat{\eta}\hat{z}_k^2)\right). \tag{4.13}$$

In our implementation of the ECM algorithm, we assume that $\hat{\Sigma}_k \equiv \text{diag}(\sigma^2)$, for some constant σ^2 which does not need to be specified since it cancels out from the resulting simplified update equations, then Equation (4.12) becomes:

$$\Sigma^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(x_i - \alpha - \beta z_k - \eta {z_k}^2) = 0,$$

and then multiply Σ on both sides, we obtain,

$$\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(x_i - \alpha - \beta z_k - \eta {z_k}^2) = 0,$$

then,

$$\hat{\alpha} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i - \hat{\beta}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z_k} - \eta\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}{z_k}^2\right). \tag{4.14}$$

This is the estimator of α used in the M-step in implementing the ECM algorithm.

Derivation for $\hat{\eta}$

For the derivation of η, we use the result of the following, which is derived by Petersen and Pedersen (2012),

$$\frac{\partial}{\partial A}(x - As)^T W(x - As) = -2W(x - As)s^T,$$

By taking partial derivative of the l_c with respect to η, we obtain,

$$\frac{\partial l_c}{\partial \eta} = \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{1}{2} w_{ik}(-2)\Sigma_k^{-1}(x_i - \alpha - \beta z_k - \eta {z_k}^2)({z_k}^2)^T, \tag{4.15}$$

Letting the above equation to be 0 and solving it,

$$\hat{\eta} = \left(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{\Sigma}_k^{-1}(x_i - \hat{\alpha} - \hat{\beta}\hat{z}_k)({z_k}^2)^T\right)\left(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{\Sigma}_k^{-1}{z_k}^4\right)^{-1}, \tag{4.16}$$

In our implementation, we have the assumption that $\hat{\Sigma}_k \equiv \text{diag}(\sigma^2)$, then the partial derivative (4.15) becomes,

$$\Sigma^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{1}{2} w_{ik}(-2)(x_i - \alpha - \beta z_k - \eta {z_k}^2)({z_k}^2)^T = 0,$$

then we multiply Σ on both sides and we obtain,

$$\hat{\eta} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}x_i\hat{z}_k^2 - \hat{\alpha}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k^2 - \hat{\beta}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k^3}{\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k^4}. \tag{4.17}$$

Derivation for $\hat{\Sigma}_k$

For the derivation of Σ, again, we use Equations 2.32 and 2.33. Rewrite (4.4) to be

$$\begin{aligned}\tilde{l_c} = &\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\log(\pi_k) + \sum_{i=1}^{n}\sum_{k=1}^{K}\frac{1}{2} w_{ik}\log(|\Sigma_k|^{-1}) + \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{m}{2}\log(2\pi)w_{ik} \\ &+ \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{1}{2} w_{ik}(x_i - \alpha - \beta z_k - \eta {z_k}^2)^T\Sigma_k^{-1}(x_i - \alpha - \beta z_k - \eta {z_k}^2).\end{aligned} \tag{4.18}$$

By taking partial derivative of the $\tilde{l_c}$ with respect to Σ_k^{-1} and by letting it to be zero, we obtain,

$$\sum_{i=1}^{n} \frac{1}{2} w_{ik} \left((\Sigma_k^{-1})^{-1} \right)^T + \sum_{i=1}^{n} -\frac{1}{2} w_{ik} (x_i - \alpha - \beta z_k - \eta {z_k}^2)(x_i - \alpha - \beta z_k - \eta {z_k}^2)^T = 0,$$

since Σ_k is symmetric, then $\Sigma_k^T = \Sigma_k$,

$$\sum_{i=1}^{n} w_{ik} \Sigma_k = \sum_{i=1}^{n} w_{ik} (x_i - \alpha - \beta z_k - \eta {z_k}^2)(x_i - \alpha - \beta z_k - \eta {z_k}^2)^T,$$

we obtain (variance parameterization(iv)),

$$\hat{\Sigma}_k = \frac{\sum_{i=1}^{n} w_{ik} (x_i - \hat{\alpha} - \hat{\beta} \hat{z}_k - \hat{\eta} \hat{z}_k^2)(x_i - \hat{\alpha} - \hat{\beta} \hat{z}_k - \hat{\eta} \hat{z}_k^2)^T}{\sum_{i=1}^{n} w_{ik}}. \tag{4.19}$$

Derivation for $\hat{\sigma}_{jk}^2$

When $\Sigma_k \in R^m$ is diagonal, that is $\Sigma_k = \text{diag}(\sigma_{jk}^2)_{\{1 \le j \le m\}}$, where $k = 1, \ldots, K$,

$$\Sigma_k = \begin{pmatrix} \sigma_{1k}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{2k}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{jk}^2 \end{pmatrix}, \tag{4.20}$$

and $|\Sigma_k| = \prod_{j=1}^{m} \sigma_{jk}^2$, since $|\Sigma_k|^{-1} = |\Sigma_k^{-1}|$,

$$|\Sigma_k^{-1}| = \begin{vmatrix} \frac{1}{\sigma_{1k}^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_{2k}^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_{jk}^2} \end{vmatrix} = \prod_{j=1}^{m} \frac{1}{\sigma_{jk}^2}, \tag{4.21}$$

The log-likelihood function from the previous section is the following,

$$\begin{aligned} \tilde{l_c} = & \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \log(\pi_k) + \sum_{i=1}^{n} \sum_{k=1}^{K} \frac{1}{2} w_{ik} \log(|\Sigma_k|^{-1}) + \sum_{i=1}^{n} \sum_{k=1}^{K} -\frac{m}{2} \log(2\pi) w_{ik} \\ & + \sum_{i=1}^{n} \sum_{k=1}^{K} -\frac{1}{2} w_{ik} (x_i - \alpha - \beta z_k - \eta {z_k}^2)^T \Sigma_k^{-1} (x_i - \alpha - \beta z_k - \eta {z_k}^2), \end{aligned}$$

and $\log(|\Sigma_k|^{-1})$ in the log-likelihood function above will become,

$$\log(|\Sigma_k|^{-1}) = \log(|\Sigma_k^{-1}|) = \log(\frac{1}{\sigma_{1k}^2}.\frac{1}{\sigma_{2k}^2}\ldots\frac{1}{\sigma_{mk}^2}) = -2\sum_{j=1}^{m}\log\sigma_{jk},$$

then the log-likelihood function $\tilde{l}_c$ will become,

$$\tilde{l}_{new} = constant + \sum_{i=1}^{n} -\frac{1}{2}w_{ik}(-2)\sum_{j=1}^{m}\log\sigma_{jk} + \sum_{i=1}^{n}\sum_{j=1}^{m} -\frac{1}{2}w_{ik}\frac{(x_{ij}-\alpha_j-\beta_j z_k-\eta {z_k}^2)}{\sigma_{jk}^2},$$

by taking partial derivative of the $\tilde{l}_{new}$ with respect to σ_{jk} and by letting it to be zero, we obtain,

$$\sum_{i=1}^{n} -w_{ik}\frac{1}{\sigma_{jk}} + \sum_{i=1}^{n} w_{ik}(x_{ij}-\alpha_j-\beta_j z_k-\eta {z_k}^2)^2\sigma_{jk}^{-3} = 0,$$

then,

$$\sum_{i=1}^{n} w_{ik}\frac{1}{\sigma_{jk}} = \frac{1}{\sigma_{jk}^3}\sum_{i=1}^{n} w_{ik}(x_{ij}-\alpha_j-\beta_j z_k-\eta {z_k}^2)^2,$$

we then obtain variance parameterization (ii),

$$\hat{\sigma}_{jk}^2 = \frac{\sum_{i=1}^{n} w_{ik}(x_{ij}-\hat{\alpha}_j-\hat{\beta}_j\hat{z}_k-\hat{\eta}\hat{z}_k^2)^2}{\sum_{i=1}^{n} w_{ik}}. \tag{4.22}$$

Derivation for $\hat{\Sigma}$

For the derivation of parameter Σ, again, we use Equations 2.32 and 2.33. When $\Sigma_k \equiv \Sigma$, the log-likelihood function (4.4) can be rewrite as,

$$\begin{aligned}\tilde{l}_c = &\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\log(\pi_k) + \sum_{i=1}^{n}\sum_{k=1}^{K}\frac{1}{2}w_{ik}\log(|\Sigma|^{-1}) + \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{m}{2}\log(2\pi)w_{ik} \\ &+\sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{1}{2}w_{ik}(x_i-\alpha-\beta z_k-\eta {z_k}^2)^T\Sigma^{-1}(x_i-\alpha-\beta z_k-\eta {z_k}^2).\end{aligned} \tag{4.23}$$

By taking partial derivative of the $\tilde{l}_c$ with respect to Σ^{-1} and by letting it to be zero, we obtain,

$$\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{1}{2}w_{ik}\left((\Sigma^{-1})^{-1}\right)^T + \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{1}{2}w_{ik}(x_i-\alpha-\beta z_k-\eta {z_k}^2)(x_i-\alpha-\beta z_k-\eta {z_k}^2)^T = 0,$$

since Σ_k is symmetric, and $\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik} = n$ then we obtain variance parameterization (iii),

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(x_i - \hat{\alpha} - \hat{\beta}\hat{z}_k - \hat{\eta}\hat{z}_k^2)(x_i - \hat{\alpha} - \hat{\beta}\hat{z}_k - \hat{\eta}\hat{z}_k^2)^T. \tag{4.24}$$

Derivation for $\hat{\sigma}_j$

When $\Sigma_{m\times m}$ is diagonal, that is $\Sigma = \text{diag}(\sigma_j^2)_{\{1\le j\le m\}}$,

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_j^2 \end{pmatrix}, \tag{4.25}$$

and $|\Sigma| = \prod_{j=1}^{m}\sigma_j^2$, since $|\Sigma|^{-1} = |\Sigma^{-1}|$,

$$|\Sigma^{-1}| = \begin{vmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_j^2} \end{vmatrix} = \prod_{j=1}^{m}\frac{1}{\sigma_j^2}, \tag{4.26}$$

The log-likelihood function from the previous section is the following,

$$\begin{aligned}\tilde{l_c} = &\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\log(\pi_k) + \sum_{i=1}^{n}\sum_{k=1}^{K}\frac{1}{2}w_{ik}\log(|\Sigma|^{-1}) + \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{m}{2}\log(2\pi)w_{ik} \\ &+\sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{1}{2}w_{ik}(x_i - \alpha - \beta z_k - \eta {z_k}^2)^T\Sigma^{-1}(x_i - \alpha - \beta z_k - \eta {z_k}^2),\end{aligned}$$

and $\log(|\Sigma|^{-1})$ will become,

$$\log(|\Sigma|^{-1}) = \log(|\Sigma^{-1}|) = \log(\frac{1}{\sigma_1^2}.\frac{1}{\sigma_2^2}\ldots\frac{1}{\sigma_m^2}) = -2\sum_{j=1}^{m}\log\sigma_j,$$

then the log-likelihood function $\tilde{l_c}$ will become,

$$\tilde{l}_{new} = constant + \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{1}{2}w_{ik}(-2)\sum_{j=1}^{m}\log\sigma_j + \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{m} -\frac{1}{2}w_{ik}\frac{(x_{ij} - \alpha_j - \beta_j z_k - \eta {z_k}^2)^2}{\sigma_j^2},$$

by taking partial derivative of the $\tilde{l}_{new}$ with respect to σ_j and by letting it to be zero, we obtain,

$$\sum_{i=1}^{n}\sum_{k=1}^{K} -w_{ik}\frac{1}{\sigma_j} + \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(x_{ij} - \alpha_j - \beta_j z_k - \eta {z_k}^2)^2\sigma_j^{-3} = 0,$$

since $\sum_{i=1}^{n}\sum_{k=1}^{K} -w_{ik} = n$,

$$\frac{n}{\sigma_j} = \frac{1}{\sigma_j^3}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(x_{ij} - \alpha_j - \beta_j z_k - \eta {z_k}^2)^2,$$

we then obtain variance parameterization (i),

$$\hat{\sigma}_j^2 = \frac{\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(x_{ij} - \hat{\alpha}_j - \hat{\beta}_j\hat{z}_k - \hat{\eta}\hat{z}_k^2)^2}{n} \tag{4.27}$$

Note that in the **R** implementation and simulation for the quadratic model, we only used variance parameterization (i), i.e. Equation (4.27).

4.4 Identifiability

We implement the ECM algorithm with the same diagonal variance matrices for all components in **R**. Subsequently, we use simulated data to test the accuracy. Initially, we use two-dimensional data with a sample size of 500 and set the true mixture components to be $K = 4$. We then observed very large estimation values and some very small values, indicating potential issues with identifiability. To further confirm whether another set of values exists which will lead to the same curve, we assume the following reparameterization for z_k,

$$z_k = a + d\tilde{z}_k,$$

where a and d are real-valued scalars. By substituting the reparameterization of z_k into model (4.3), we can obtain reparameterizations for all other parameters,

$$\alpha = \tilde{\alpha} - \frac{d}{a}\tilde{\beta} + \frac{d^2}{a^2}\tilde{\eta},$$

Algorithm 3 ECM Algorithm

1. Initialization:

(i) Choose the number of mixture components, K, where K is a positive integer.

(ii) Choose starting values for the parameters: π_k, α, β, z_k, η, σ_j.

(iii) Select the number of iterations, s; 20 iterations is suggested.

2. Iterations:

E-step

For each k, compute the posterior probability of observation i belonging to component k, according to Equation (4.5).

M-step

steps $\leftarrow 0$
while *steps* $\leq s$ **do**
  counter $\leftarrow 0$ ▷ Reset counter for each step
  while *counter* ≤ 5 **do**
    Update z_k, β, α and η, cycle between Equations (4.8), (4.11), (4.14), and (4.17).
    counter $\leftarrow$ *counter* $+ 1$
  end while
  Update π_k via Equation (4.6)
  Update σ_j according to Equation (4.27)
  steps $\leftarrow$ *steps* $+ 1$
end while

3. Output: Return the estimated parameters.

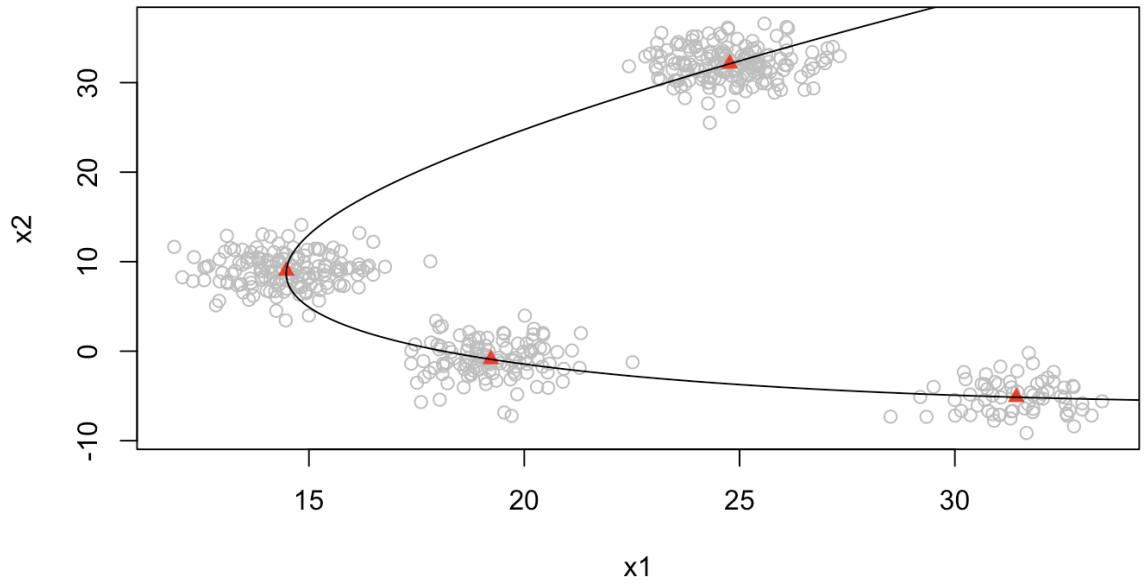

Figure 4.2: Simulated data generated with true values: $\pi_k^T = (0.35, 0.15, 0.30, 0.20)$, $\alpha^T = (15, 5)$, $\beta^T = (-5, 15)$ and $\eta^T = (10, 5)$.

$$\beta = \frac{\tilde{\beta} - \frac{2d}{a}\tilde{\eta}}{a},$$

$$\eta = \frac{\tilde{\eta}}{a^2},$$

where α, β, η and z_k are the estimated parameters, while $\tilde{\alpha}$, $\tilde{\beta}$, $\tilde{\eta}$ and $\tilde{z}_k$ represent the transformed parameters. Now it is clear that there is an identifiability problem, which we can resolve by simply standardizing z_k (subtracting the mean of z_k and then dividing by its standard deviation). Additionally, to identify the direction of the latent variable, we enforce $\beta_1 \geq 0$ (but any other component of β could equally be chosen for this).

After addressing the identifiability problem, we use this simulated data set to illustrate how the model approximates the non-linear data using a curve that is parameterized by a single random effect. In Figure 4.2, the data points represent simulated two-dimensional data with 4 mixture components. The curve in gray that is passing through the mixture centers ($\alpha + \beta z_k + \eta z_k^2$, marked with red triangles) is the fitted quadratic curve. Figure 4.2 also suggests that this methodology could be used as another approach to estimate principal curves and we return to this idea in Section 4.7.

Table 4.1: Simulation results under variance parameterization (i).

		Average estimates		
	True	$n = 100$	$n = 300$	$n = 500$
π_1	0.1500	0.1269	0.1269	0.1355
π_2	0.2000	0.1949	0.1954	0.1946
π_3	0.3000	0.2916	0.2963	0.2971
π_4	0.3500	0.3866	0.3814	0.3727
z_1	1.2696	1.2834	1.2847	1.2769
z_2	0.2402	0.2091	0.2021	0.2179
z_3	-0.4461	-0.4635	-0.4579	-0.4497
z_4	-1.0637	-1.0289	-1.0288	-1.0451
α_1	5.0000	5.4230	5.3237	5.3444
α_2	15.0000	14.9822	15.0769	15.0253
β_1	-5.0000	-6.6284	-6.5173	-6.0477
β_2	15.0000	14.1142	14.3244	14.3750
η_1	5.0000	4.8833	4.8010	4.9135
η_2	10.0000	10.4077	10.3825	10.2027
σ_1	1.0000	1.1239	1.1701	1.1103
σ_2	2.0000	2.2411	2.3176	2.2107

4.5 One-level Quadratic Model Simulation

We conduct a simulation with 2-dimensional data with four mixture component to test the accuracy of the implementation. The structure of the simulated data is shown in Figure 4.2 in Section 4.4. For sample sizes $n = 100$, $n = 300$ and $n = 500$, we generate 300 replicates respectively from model (4.3). One thing to notice here is that for each of the replicates, we ran the ECM algorithm 20 times to select the best estimates with the smallest AIC value; this process is selecting a good starting value for the ECM algorithm. The averaged estimates for each parameter can be found in Table 4.1, and the true values are shown in the first column. We can observe that the averaged estimates of the parameters are generally close to their true values, and overall, the bias is reduced with the increase of the sample size. Boxplots for parameters α, β, z_k, η, π_k, and σ are shown is Figure 4.3, Figure 4.4, Figure 4.5, Figure 4.6, Figure 4.7 and Figure 4.8. They also show that these parameter estimators give sensible and consistent estimates.

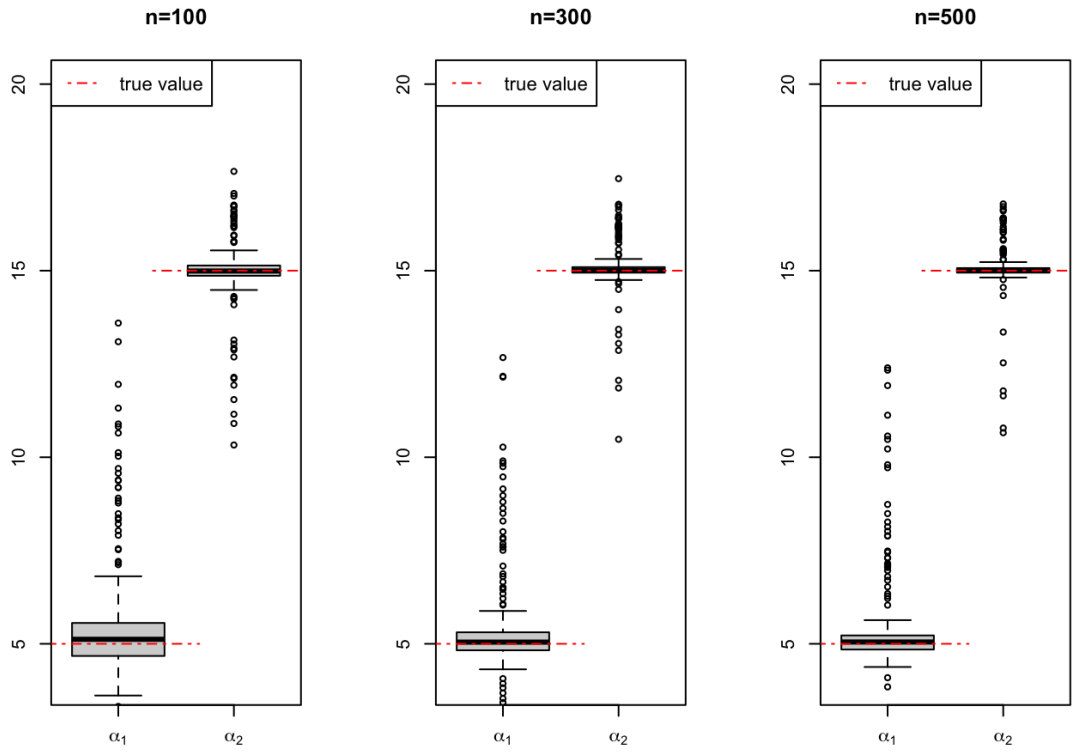

Figure 4.3: Estimations of parameter $\alpha = (\alpha_1, \alpha_2)^T$ with different sample sizes under the variance parameterization (i)

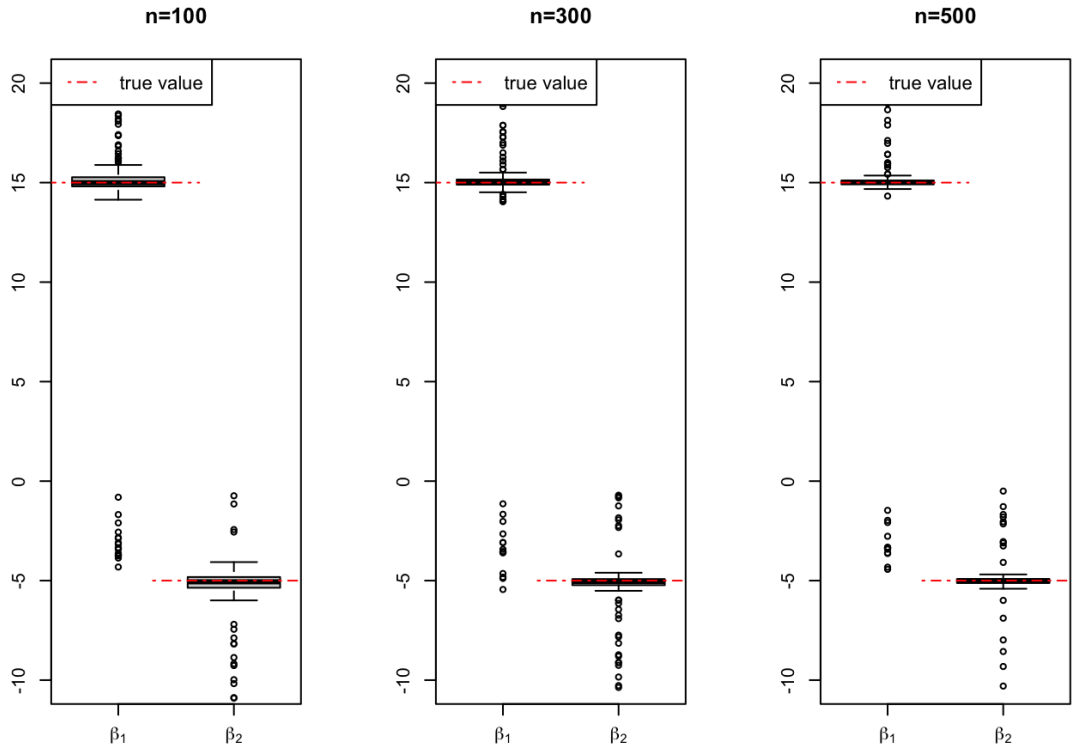

Figure 4.4: Estimations of parameter $\beta = (\beta_1, \beta_2)^T$ with different sample sizes under the variance parameterization (i)

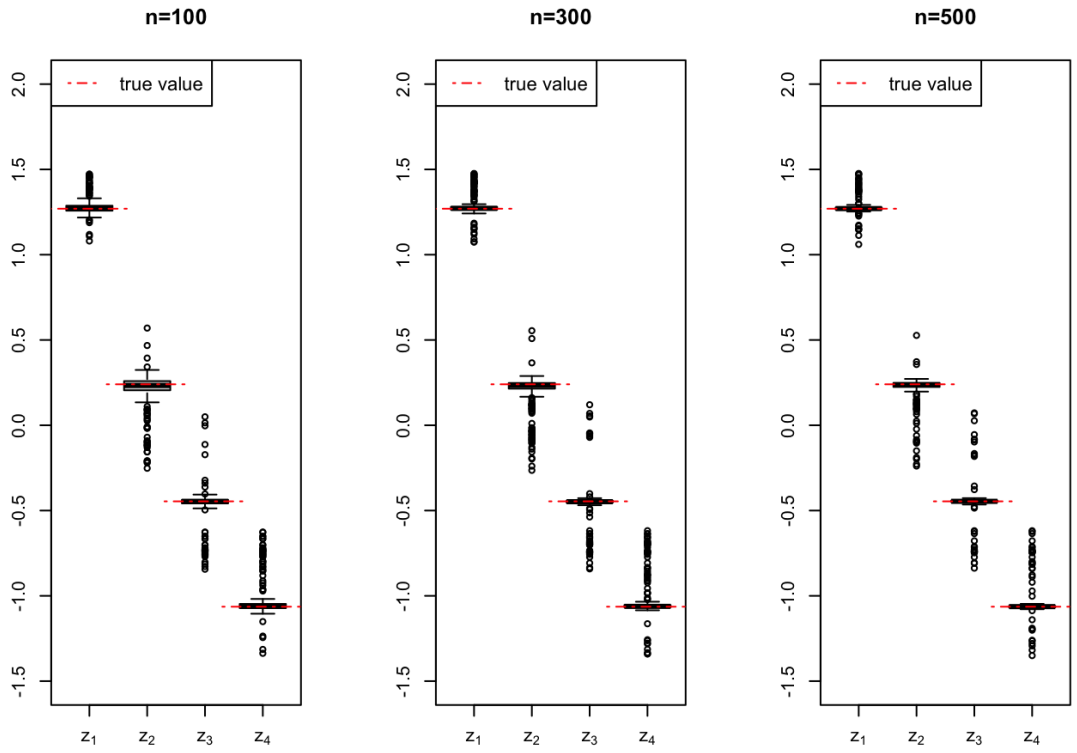

Figure 4.5: Estimations of parameter z_k with different sample sizes under the variance parameterization (i)

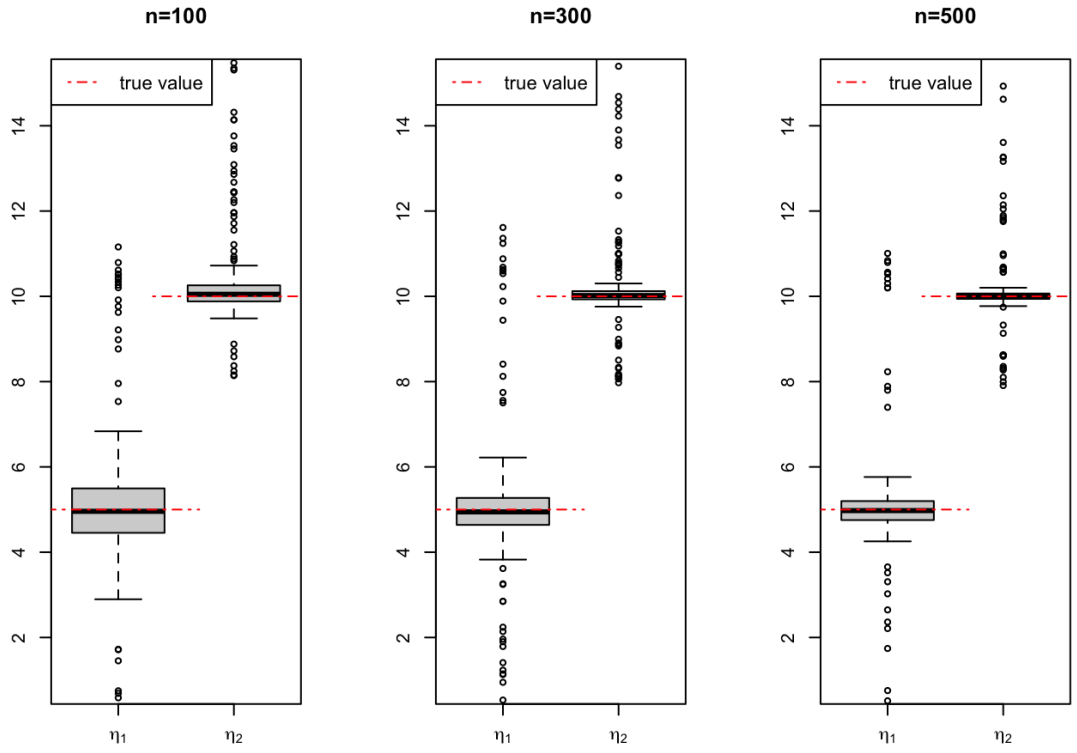

Figure 4.6: Estimations of parameter $\eta = (\eta_1, \eta_2)^T$ with different sample sizes under the variance parameterization (i)

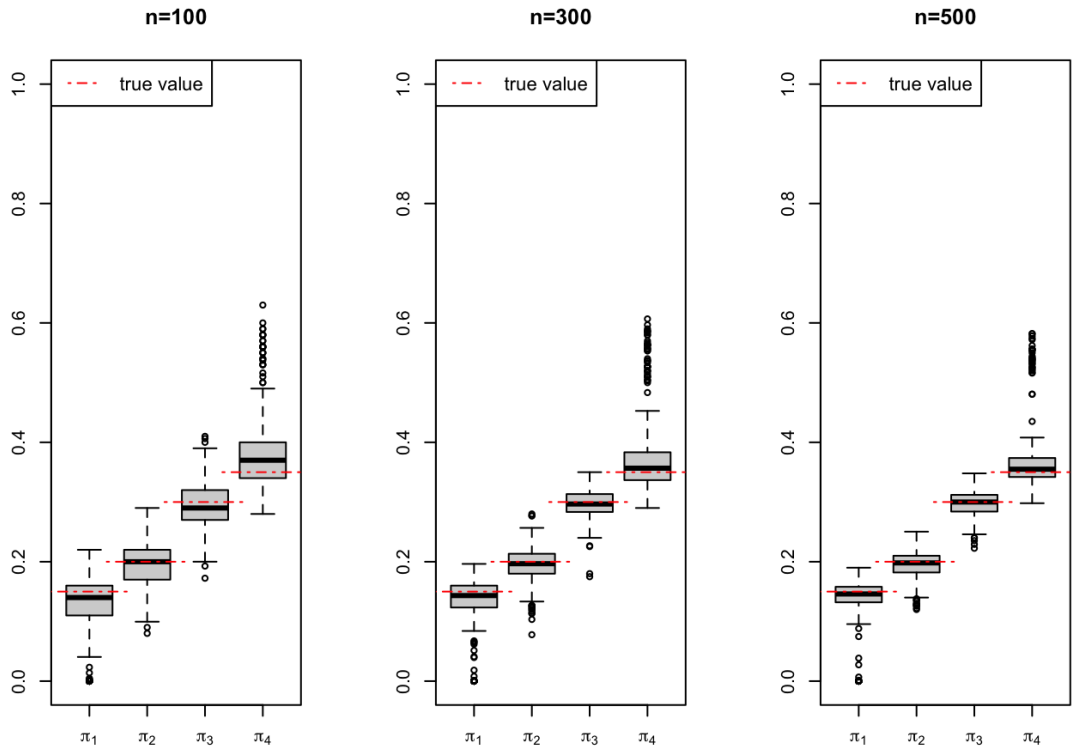

Figure 4.7: Estimations of parameter π_k with different sample sizes under the variance parameterization (i)

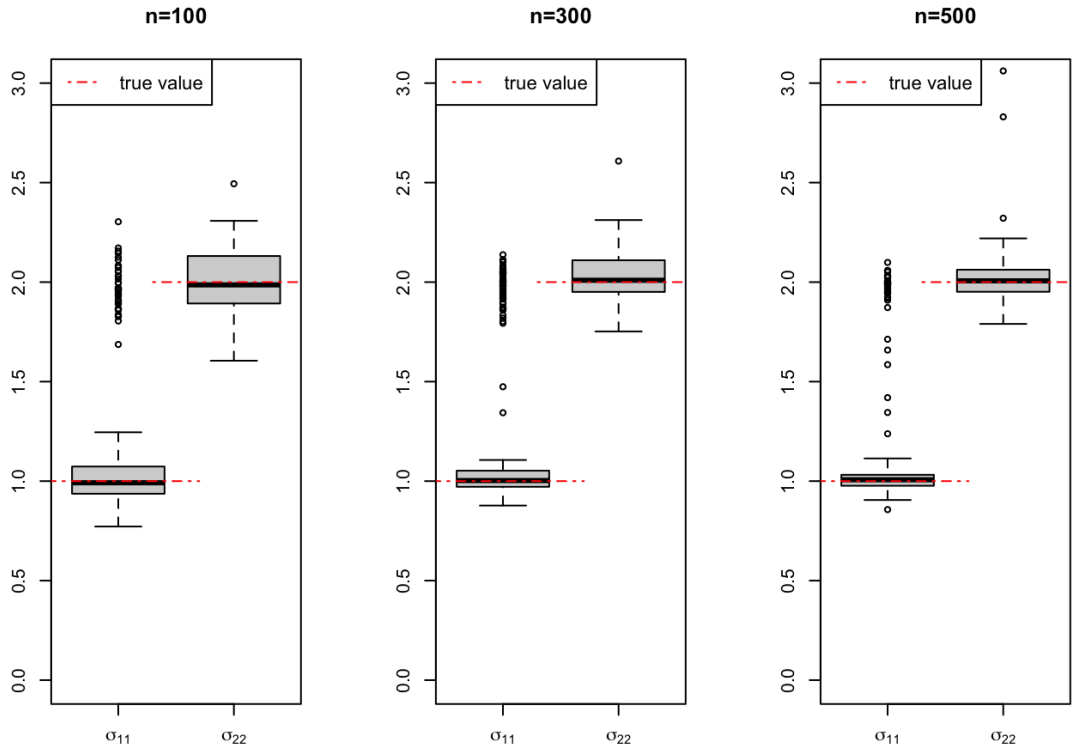

Figure 4.8: Estimations of parameter σ with different sample sizes under the variance parameterization (i)

4.6 Mussels data

As discussed in the beginning of this chapter, a curve may provide a better fit to the mussels data. Now we fit the mussels data to the quadratic model with mixture component $K = 3$, the projections of the fitted curve onto 2-dimension plot are shown in Figure 4.9. We can also compare the AIC and BIC values from the linear and non-linear models, see Table 4.2. It is evident that a quadratic model have a better performance in fitting the mussels data.

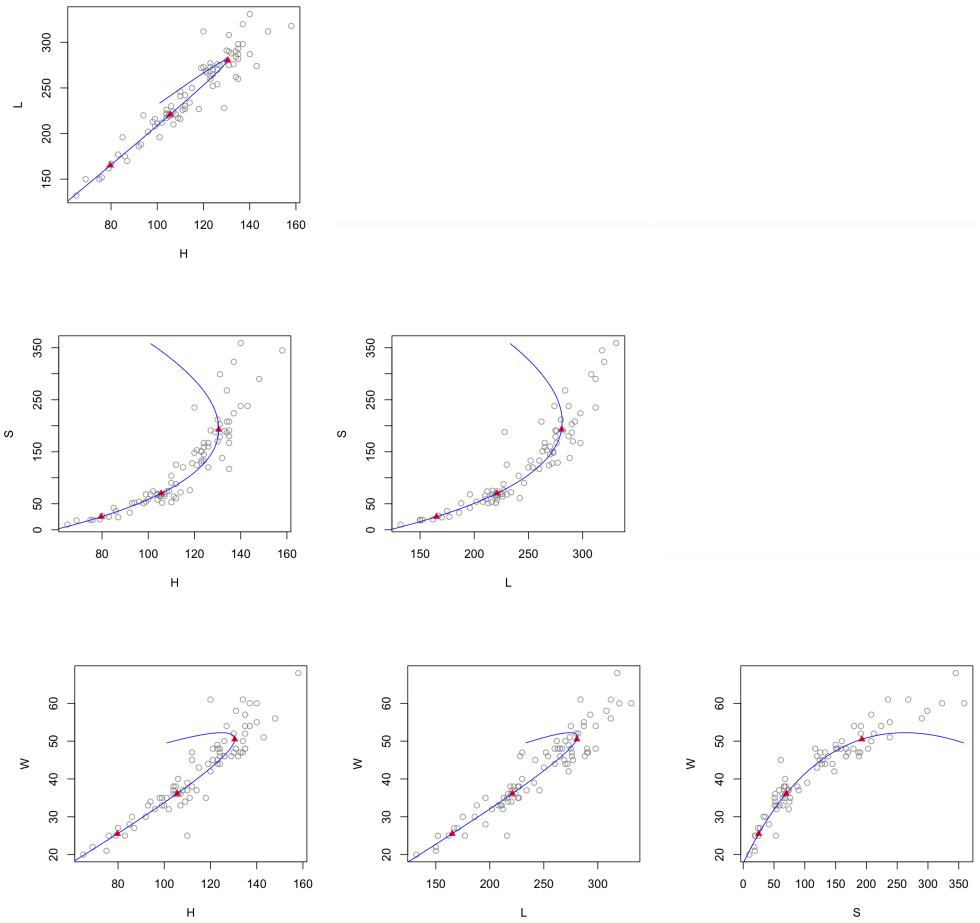

Figure 4.9: Mussels data with 2-dimensional projections of fitted curve.

Table 4.2: The AIC and BIC values for the fitted linear model and the fitted curve.

	Linear model	Quadratic model
AIC	2836.840	2813.742
BIC	2877.754	2864.283

4.7 Comparison with Principal Curves

The quadratic model we have proposed is a way of using a random effect model to estimate principal curves. Both methodologies aim to approximate multivariate datasets with smooth one-dimensional curves. The generalization of the non-linear model can be written as follows,

$$x_i = g(\lambda_i) + \varepsilon_i, \tag{4.28}$$

where Hastie and Stuetzle (1989) attempted to use this model for the principal curve. However, as illustrated by Tibshirani (1992) when using (4.28) as the data generating function, then g is not the principal curve for the data. In our proposed model, $g(\lambda_i) = g(z_i) = \sum_{j=0}^{p} \alpha_j B_j(z_i)$, which is the generative model. HS defined the principal curve to be a curve that satisfies the self-consistent property: the curve g is called self-consistent, if $E(X|\lambda_r(X) = \lambda) = g(\lambda)$ for a.e. λ, where X is a random vector in R^p and λ_r is defined as the projection index.

The most distinctive difference in these two approaches is the projection. The principal curves are determined by finding the shortest orthogonal projections of the original data points onto the curve. HS defined the projection index as: $\lambda_r(x) = \underset{\lambda}{sup}\{\lambda : ||x - g(\lambda)|| = \underset{\lambda}{inf}||x - g(\mu)||\}$. Our method obtains projections through model-based scores, denoted as z_i^*, $z_i^* = \sum_{k=1}^{K} w_{ik}\hat{z}_k$ which are not orthogonal.

Then, we will use both simulated and real data to illustrate the principal curves method and the quadratic model proposed in the previous section, facilitating a meaningful comparison.

The first data set $x = (x_1, x_2)$ we use is generated from a circle with the sample size of $n = 100$, where $x_1 = 5\sin u + \varepsilon_1$, $x_2 = 5\cos u + \varepsilon_2$, $u \sim U[0, \frac{3}{2}\pi)$, $\varepsilon_1 \sim N(0, 1)$ and $\varepsilon_2 \sim N(0, 1)$ are Gaussian noise. The second data set $x_{new} = (x_{1new}, x_{2new})$ is generated again from a circle with the sample size of $n = 300$, where $x_1 = 5\sin u + \varepsilon_1$, $x_2 = 5\cos u + \varepsilon_2$, $u \sim U[\frac{1}{2}\pi, 2\pi)$, $\varepsilon_1 \sim N(0, 1)$ and $\varepsilon_2 \sim N(0, 1)$ are Gaussian noise.

We fit both the quadratic model and the HS principal curves through the use of `principal_curve()` function available in the R package **princurve** (Cannoodt, 2018) to these two data sets, and the results are shown in Figure 4.10. By comparing these two pairs of plots, we observe that our model successfully captured the circular shape of the data and accurately identified the mixture centers within the clustered data groups. However, due to the limitation

of our curve to a quadratic shape, the principal curve appears smoother in illustrating the data.

The same observation occurs when using a real dataset, which has been introduced in Section 1.2.9: the speed-flow data from California available in R package **LPCM** (Einbeck & Evers, 2024), which consist of speed and flow recorded on Line 5 of the California freeway. For this dataset, we focused on two variables: the vehicle flow in vehicles per 5 minutes (`Lane5Flow`) and the vector of vehicle speed in miles per hour (`Lane5Speed`). We fit the data again with a HS principal curve, additionally, we apply a local principal curve (Einbeck et al., 2005) which has better performance with complex data structures. This approach is implemented in the `lpc()` function within the **LPCM** R package. The results are shown in Figure 4.11. We observe that the fitted HS principal curve fails to capture the curvature of the graph. In contrast, the quadratic curve performs better in this regard but is constrained by its quadratic nature. In comparison, the fitted local principal curve can accurately connect to the closing point which has been ignored by both the quadratic curve and the HS principal curve. We also attempt to compare the projections of these two methods, see Figure 4.12. The main difference is that the projection of the principal curve is orthogonal, whereas the quadratic model is not. Apart from comparing the fitted curves, it is not fair to use any quantitative assessment, such as goodness of fit, to compare them, given the different characteristics of the projections from these two methods.

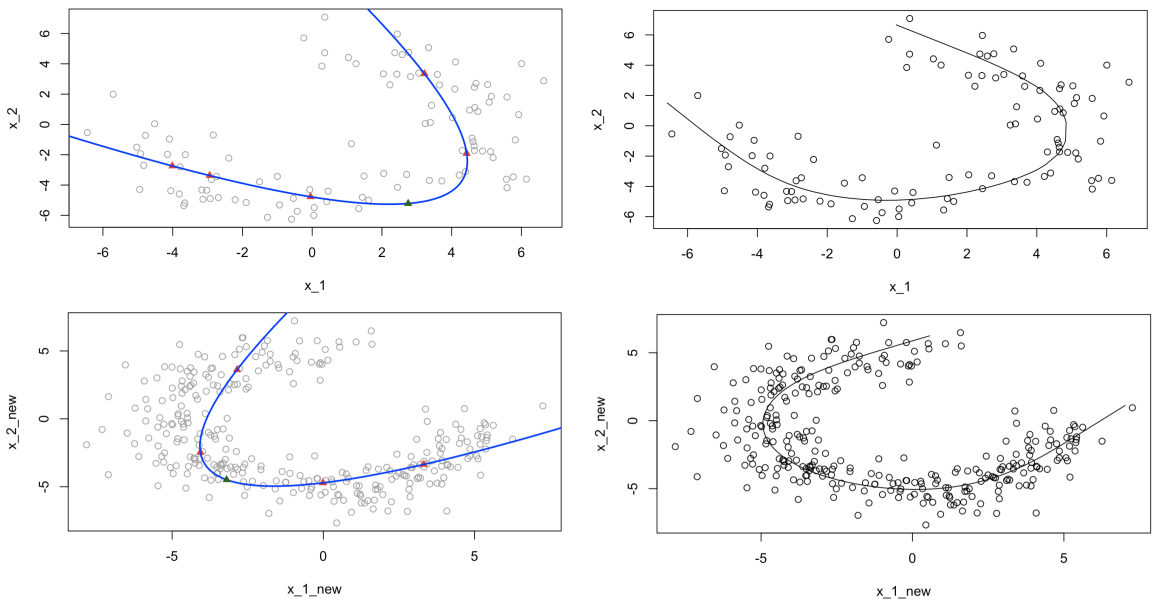

Figure 4.10: The top left figure shows the application of the quadratic model with $K = 5$ used to fit the model. The bottom left figure shows the application of the quadratic model with $K = 4$ used to fit the model. The red triangles represent the mixture centers, and the blue curve, passing through the triangles, illustrates the fitted model. Meanwhile, the right figures (both top and bottom) display the application of the principal curves, with the fitted curve represented in black.

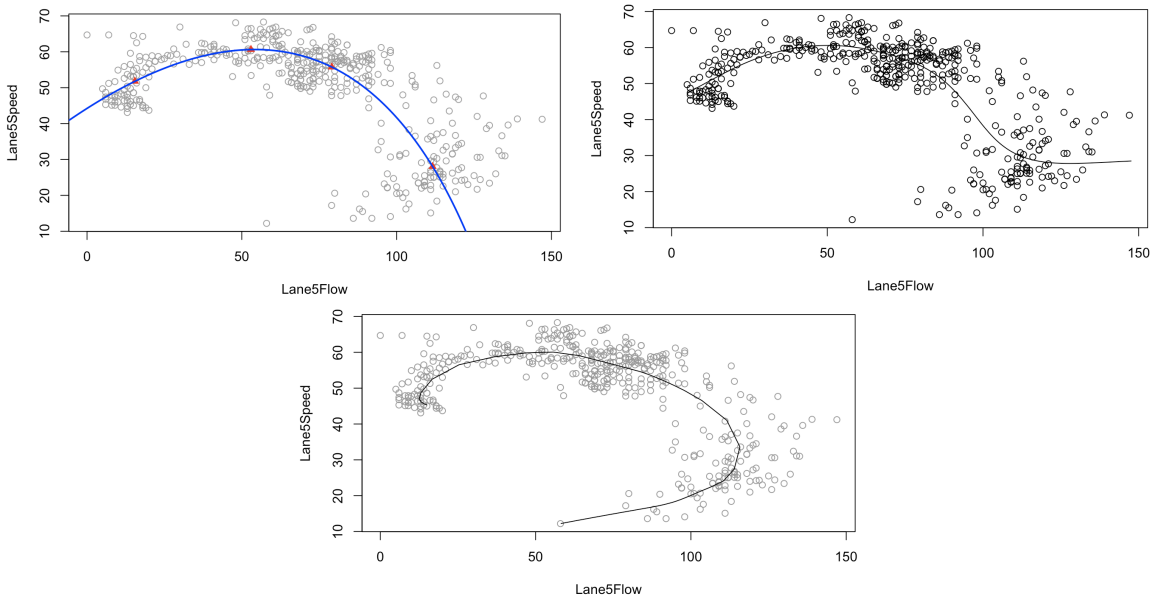

Figure 4.11: The top left figure is illustrated by the quadratic model, the top right figure is the HS principal curve and the bottom figure is the local principal curve.

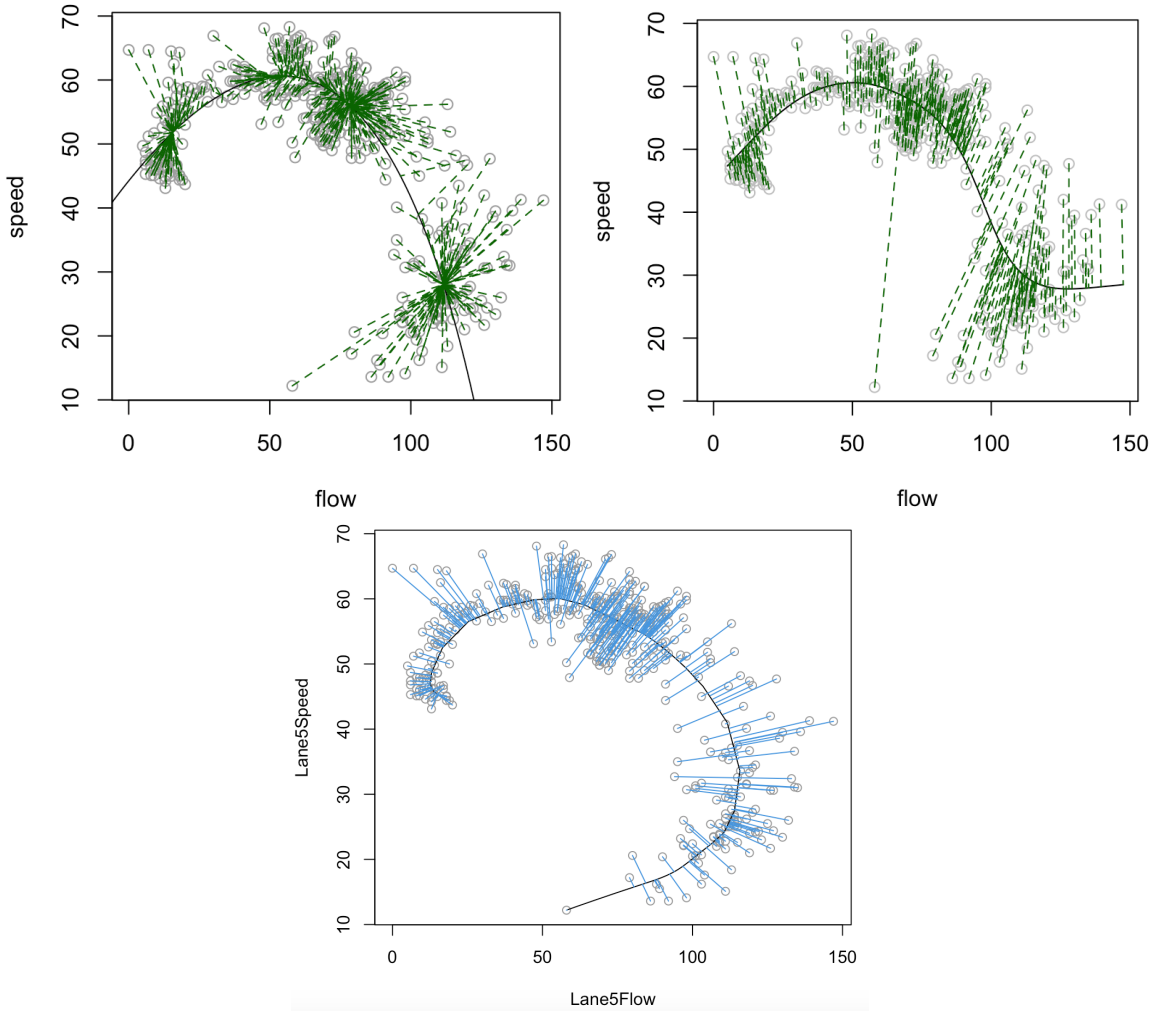

Figure 4.12: The top left figure is the projection of the quadratic model, the top right figure is the projection of HS principal curves. the bottom figure is the projection of local principal curves.

Chapter 5

R Package

We developed an R package **mult.latent.reg** (Zhang & Einbeck, 2024b), now available on CRAN, that implements the fitting methodologies for the 1- and 2-level models introduced in Chapter 2 and Chapter 3 for clustered and highly correlated multivariate data. The computations are based on an ECM algorithm in the spirit of the Nonparametric Maximum Likelihood (NPML) approach for the estimation of mixture models. The implementation also features alternative choices of the starting values for the ECM algorithm, which we discuss in the following sections. There are five functions in this package: `mult.em_1level()`, `mult.reg_1level()`, `mult.em_2level()`, `mult.reg_2level()` and `start_em()`.

5.1 Functions: `mult.em_1level()` and `mult.em_2level()`

The main functions for the 1-level model is `mult.em_1level()`. The function `mult.em_1level()` is used to obtain the Maximum Likelihood Estimates (MLE) using the ECM algorithm for one-level multivariate data. The estimates enable users to conduct clustering, ranking, and simultaneous dimension reduction on the multivariate dataset. Furthermore, when covariates are included, the function supports the fitting of multivariate response models, expanding its utility for regression analysis. It will run ECM once with (by default) 20 iterations, producing output including parameter estimates, log-likelihood, disparity, AIC, BIC values and starting points. We support four types of parameterizations for Σ: the same diagonal variance matrix for all mixture components, different diagonal variance matrices for different mixture components, the same full variance matrix for all components, and different full variance matrices for different components.

Here, we present an example of the 1-level model function applied to the faithful data (no covariates). We use `option = 1` for the starting value (to be explained in Section 5.3) and adopt the first variance parameterization.

```
> data(faithful)
> res <- mult.em_1level(faithful,K=2,steps = 10,var_fun = 1,option = 1)
```

Then we obtain the estimates, where `p` and `z` are estimated mixture parameters, `alpha` corresponds to the α parameter from the 1-level model, `beta` corresponds to the β parameter from the 1-level model, and `W` is the matrix of responsibilities.

```
> res$p
[1] 0.3590048 0.6409952

> res$z
[1] -1.336218  0.748381

> res$alpha
eruptions   waiting
 3.487783 70.897059

> res$beta
eruptions   waiting
 1.079359 12.207625

> res$W
               [,1]         [,2]
  [1,] 8.161794e-08 9.999999e-01
  [2,] 1.000000e+00 9.721718e-15
  [3,] 2.805878e-04 9.997194e-01
  [4,] 1.000000e+00 1.138163e-08
  [5,] 1.460737e-16 1.000000e+00
  [6,] 9.999982e-01 1.837069e-06
  [7,] 9.834199e-19 1.000000e+00
  ...
  [266,] 1.000000e+00 1.528210e-09
  [267,] 5.204837e-15 1.000000e+00
  [268,] 3.031372e-12 1.000000e+00
```

```
 [269,] 1.000000e+00 1.104280e-14
 [270,] 2.778375e-17 1.000000e+00
 [271,] 1.000000e+00 3.935776e-17
 [272,] 1.292981e-12 1.000000e+00
```

The main functions for the 2-level model is `mult.em_2level()`. The function `mult.em_2level()` extends the one-level version `mult.em_1level`, and it is designed to obtain Maximum Likelihood Estimates (MLE) using the ECM algorithm for nested (structured) multivariate data, e.g. multivariate test scores (such as on numeracy, literacy) of students nested in different classes or schools. The resulting estimates can be applied for clustering or constructing league tables (ranking of observations). With the inclusion of covariates, the model allows fitting a multivariate response model for further regression analysis. The outputs for `mult.em_2level()` is the same as the ones obtained from the functions for the 1-level model except we only AIC for model selection. The 2-level model offers only two choices for variance parameterization due to practical reasons: using the same diagonal variance matrix for all components of the mixture or using different diagonal variance matrices for different components.

Here, we present an example of the 2-level model function applied to the twins data (with covariates), which is analyzed in detail in Section 3.7.1, the two touch movement types of the fetus recorded: self-touch and twin-to-twin touch are used as multivariate response variable, and the mothers' mental health status: depression, perceived stress scale and stress are included as covariates . We use `option = 1` for the starting value (to be explained in Section 5.3) and adopt the second variance parameterization.

```
> set.seed(1)
> twins_res <- mult.em_2level(twins_data[,c(1,2,3)],v=twins_data[,c(4,5,6)],
K=2, steps = 20, var_fun = 2, option = 1)
```

Then we obtain the estimates, where `gamma` corresponds to the covariate coefficients matrix Γ in the 2-level model, the rest estimates are the same as the ones described in the example above.

```
> twins_res$p
[1] 0.1035222 0.8964778
```

```
> twins_res$alpha
     SelfTouchCodable OtherTouchCodable
[1,]         383.9243          611.4512

> twins_res$z
[1] 2.9427478 -0.3398185

> twins_res$beta
  66.42663 -115.61472

> twins_res$gamma
                        [,1]     [,2]      [,3]
SelfTouchCodable  -26.82411 12.10541  -7.123442
OtherTouchCodable -83.43059 46.82262 -73.719373

> twins_res$W
             [,1]          [,2]
 [1,]  8.468702e-80 1.00000000
 [2,] 1.391375e-194 1.00000000
 [3,]  1.951464e-01 0.80485357
 [4,]  0.000000e+00 1.00000000
 [5,]  2.179658e-02 0.97820342
 [6,]  9.772846e-01 0.02271544
 [7,]  4.758332e-02 0.95241668
 [8,] 1.444755e-287 1.00000000
 [9,]  1.994514e-15 1.00000000
[10,] 5.717532e-116 1.00000000
[11,]  1.209422e-41 1.00000000
[12,]  2.028218e-37 1.00000000
[13,]  2.074998e-01 0.79250022
[14,]  3.318023e-48 1.00000000
```

Table 5.1: Estimates for parameter β from 30 replicates under variance parameterization (i).

		Average estimates		
	True	$n = 100$	$n = 300$	$n = 500$
β_1	0.5000	0.5077	0.5051	0.5102
β_2	0.3000	0.2815	0.2925	0.2960
β_3	0.0600	0.0603	0.0612	0.0606
β_4	0.9000	0.8911	0.9107	0.8862

5.2 Functions: `mult.reg_1level()` and `mult.reg_2level()`

The main functions for the 1-level model is `mult.reg_1level()`. Function `mult.reg_1level()` can execute the ECM multiple times (by default 10 runs) and outputs the result with the smallest AIC value (also giving the starting points that generate that result). The main functions for the 2-level model is `mult.reg_2level()`. Similar to the 1-level model functions, `mult.reg_2level()` can execute the ECM multiple times (by default 10 runs) and outputs the result with the smallest AIC value (also giving the starting points that generate that result). The outputs for `mult.reg_2level()` is the same as the ones obtained from the functions for the 1-level model except we only AIC for model selection.

In order to motivate the need for these functions, we carry out another small simulation. For complex models, it sometimes requires multiple runs of the ECM function and selection of the results with the smallest AIC values to obtain the best results. An example illustrating this is a simulation study conducted (to test the accuracy of one-level model with covariates) in Section 2.8.3, where we fit 4-variate data to a one-level model with two covariates. Previously, in the simulation, each of the 300 replicates was run only once. In Figure 2.13, the bimodally distributed histograms for parameter β indicate that the ECM algorithm produced another set of poorly estimated results. Here, we rerun this simulation. We now run each of the replicates multiple times and choose the best results based on AIC values. Figure 5.1 displays the histograms and Table 5.1 shows the averaged estimates (with true values in the first column) of 30 replicates for parameter β, each selected out of 10 runs with the smallest AIC values. We observe that the averaged estimates of the parameter β are now close to their true values across all parameters and sample sizes, with the bias in the estimates reducing for larger sample sizes. Due to the computational expense, we used only 30 replicates; however, this still suffices to illustrate our point.

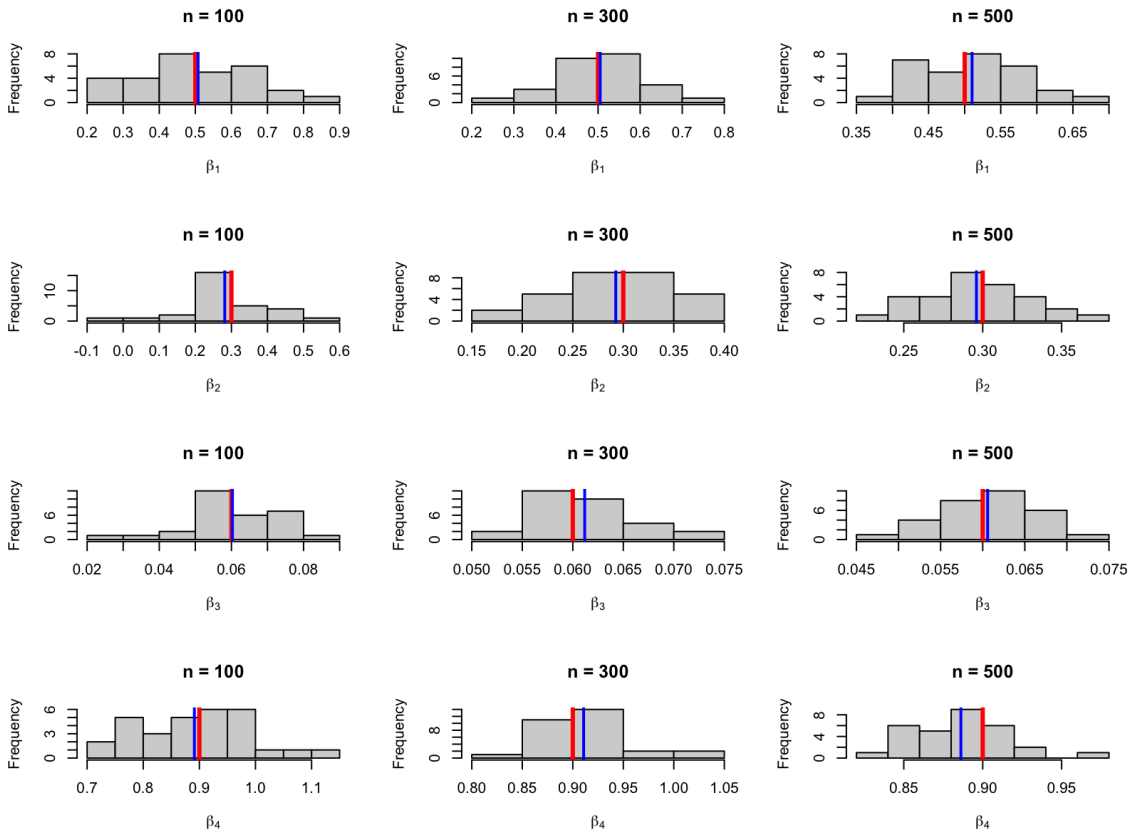

Figure 5.1: A histogram illustrating the overall estimates of 30 replicates for all three sample sizes is shown for parameter β. The vertical red line is the true value and the blue line is the mean.

5.3 Starting Values Options

Using appropriate starting values for the parameters is beneficial for the ECM to find the maximum likelihood parameter estimates. In R package **mult.latent.reg**, the function `start_em()` provides starting values for parameters used in the four ECM functions introduced in the previous section; we provide four options of data-dependent starting values for the ECM initialization, with the first option introduced in Section 2.6 and the other ones novel:

(i) `option=1`: For the mixture weights, we use $\pi_k^{(0)} = \frac{1}{K}$, where K is the number of components. We draw random numbers from a standard normal distribution as the starting values for the mass points $z_k^{(0)}$. We use column means for the line parameters $\alpha^{(0)}$, and $\beta^{(0)} = x_r - \alpha^{(0)}$, where $x_r \in \mathbb{R}^m$ is a randomly selected observation. For parameter Γ, we first fit separate linear models, each using one of the columns of x_i as response variable and v_i as predictor variables, then we use the coefficient estimates as the starting values, $\Gamma^{(0)}$. For all four variance parameterizations, we use a diagonal matrix $\Sigma^{(0)} \in \mathbb{R}^{m\times m}$, not depending on k, as the 'starting variance matrix': Denote s_j for $j = 1, 2, ..., m$ the sample standard deviation of the j-th variable. Then, for each diagonal element $(\sigma_j^{(0)})^2$ of $\Sigma^{(0)}$, one has the starting value $\sigma_j^{(0)} = \frac{s_j}{K},\ j = 1, \ldots, m$.

(ii) `option=2`: We use a short run (5 iterations) of the ECM process which uses option (i) with `var_fun=1` as the starting values, and then use the estimates as the starting values for a relatively larger number of iterations. This approach is motivated by Biernacki et al. (2003), where a short run of the EM is applied before running CEM runs.

(iii) `option=3`: The parameter β in our model plays a similar role to the rotation matrix in principal component analysis, specifically aligned with the first principal component. This observation motivated our choice of using the first principal component of the rotation matrix as the initial values for β, while keeping the starting values for the remaining parameters consistent with those described in (i).

(iv) `option=4`: In the application of clustering, a small number of observations in a dataset intended to form a distinct group may occasionally be assigned to a neighboring cluster. This inspired the idea that it would be better to use a more precise starting value for the mass points z_k. What we do is that first, take the scores of the first principal component of the

data and perform K-means on these. Then the starting values for the parameter π_k are the proportions of the clustering assignments, and the starting values for z_k are the values of the K-means centers. The starting values for the rest of the parameters are the same as described in (i).

We use the trading data as example to illustrate the `start_em()` function (the detailed application of this data can be found in Section 3.7.2). We first use `option = 1` with the first variance parameterization:

```
> start <- start_em(trading_data[,c(1,2)], option = 1, var_fun = 1)
```

then we obtain,

```
> start$p
[1] 0.5 0.5
> start$alpha
  import   export
3.722413 3.761973
> start$beta
[1] 0.5486318 0.5464599
> start$z
[1] -1.714035 -1.191245
> start$sigma
[[1]]
[1] 0.2802152 0.2989228


[[2]]
[1] 0.2802152 0.2989228
```

Then we use `option = 2` with the third variance parameterization:

```
> start <- start_em(faithful, option = 2, var_fun = 3)
```

we obtain,

```
> start$p
[1] 0.3932469 0.6067531
```

```
> start$alpha
  import   export
3.722413 3.761973
> start$beta
0.4531160 0.4756917
> start$z
[1]  1.2421479 -0.8050571
> start$sigma
          [,1]      [,2]
[1,] 0.327469 0.0000000
[2,] 0.000000 0.3597126
```

The performance of these options is illustrated in Figure 5.2 for two data sets, namely the trading data (also known as the import and export data) introduced in subsection 3.7.2, and a 1-level data set where five fetal movement types serve as multivariate outcomes, and a variable indicating pre/post-Covid status as covariate introduced in 2.10.4. The left plot shows that, in terms of AIC values obtained from 50 applications of each starting value option, `option=3` tends to perform better than the other three options, with `option=2` showing the worst performance. Meanwhile, for the fetal data, `option=2` tends to perform best, emphasizing that different starting point choices may be successful for different data sets.

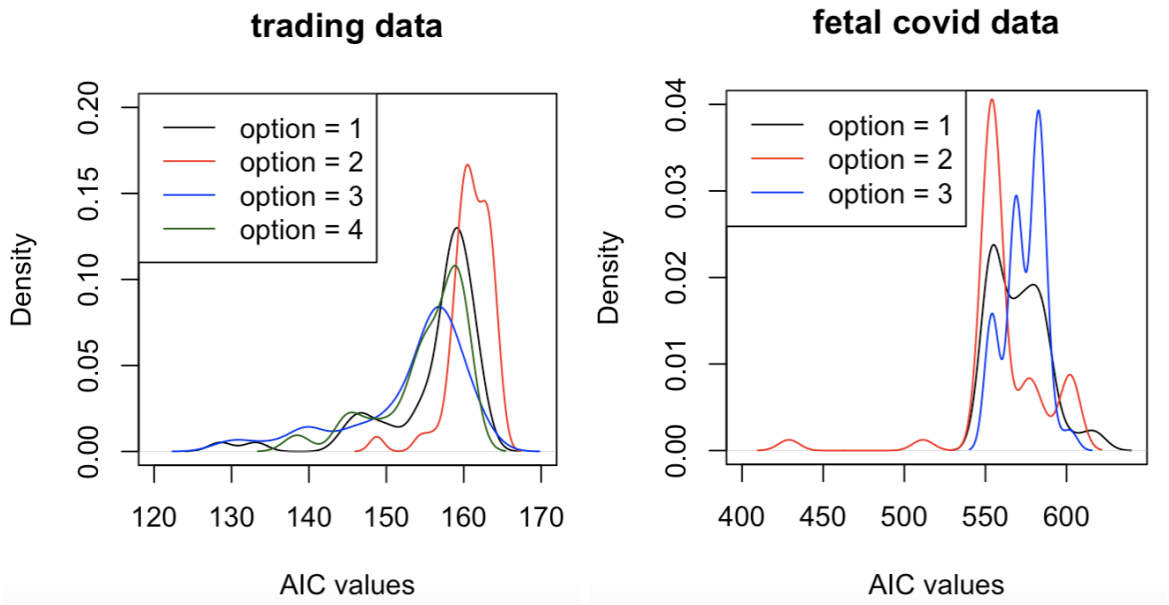

Figure 5.2: Distributions of AIC values from 50 runs for each starting value option, for the trading data (left) and the fetal data (right).

Chapter 6

Concluding Remarks

We have proposed a versatile statistical model based on a latent variable representation that simultaneously approaches dimension reduction and clustering, which are usually handled either separately or sequentially. This model provides solutions to a wide range of inferential problems, including multivariate regression where the original data space may serve as either the predictors or the responses.

Building on the proposed model, we have provided a novel methodological approach for the inclusion of a random effect into multivariate response models, based on the NPML method for mixture models. The proposed approach enables us to accurately estimate covariate effects under the presence of correlations between response variables. Crucially, such correlations impact the standard errors of parameter estimates, our application studies indicate that the standard errors tend to decrease when correlations between responses are taken into account. We also observed this behavior in our data applications. It should however be noted that when using this simultaneous approach no analytic calculation of the standard errors is possible, hence requiring us to resort to bootstrap techniques.

Another advantage of the proposed methodology is in providing the matrix of posterior probabilities produced alongside the estimation process, as well as in calculating posterior random effects, similar to principal component scores, based on the fitted model. We have demonstrated how these can be used for model-based clustering along the direction of the latent subspace and conditional on covariate values. The clustering can be performed eiher directly based on the Maximium a posteriori (MAP) rule, or can be driven by a user-specified degree of confidence in the cluster allocation, allowing for fine-grained insights into the separability of upper-level units on the scale of their posterior random effect.

Computationally, the proposed two-level model can be regarded as the multivariate extension of the `allvc()` function available in R package **npmlreg** (Einbeck et al., 2018).

We here focused on the Gaussian errors assumption for the response model and used the nonparametric maximum likelihood approach to handle the marginal density of x_{ij}. In contrast, the `allvc()` function is based on the glm framework, hence allowing any arbitrary exponential family distribution for the response.

We close our discussion with the extension to the multivariate data with a non-linear latent structure, where we approximate principal curves using a quadratic polynomial parametrized by a single random effect. The quadratic model is an exploratory investigation of the generalized non-linear model. Due to its quadratic nature, a quadratic curve cannot accurately describe shapes with curvatures of more than half-circles. But still, it remains an interesting starting point for the latent variable model for more complex data shapes. Furthermore, we have published a paper within the framework of the quadratic model (Zhang & Einbeck, 2024a) in which we consider the clustering of multivariate data with a non-linear latent structure. This aims to establish an ordering of the clusters with respect to an underlying latent variable. As our motivating example for a situation where such a technique is desirable, we consider scatterplots of traffic flow and speed, where a pattern of consecutive clusters can be thought to be linked by a latent variable which is interpretable as traffic density. Some further directions include: similar to the linear one-level model, we could extend the current quadratic model to a two-level version. Additionally, one could consider extending this framework towards any real-valued basis function, such as a polynomial of multiple (more than 2) degrees. Then the shape of the curve can be smoother, capturing the shape of the data more accurately.

Appendix A

Derivations of one-level model with covariates

The derivations of the parameter estimators for the one-level model with covariates,

$$x_i = \alpha + \beta z_k + \Gamma v_i + \varepsilon_i, \tag{A.1}$$

The expected complete log-likelihood of model (A.1) is the following,

$$\begin{aligned} l_c = &\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\log(\pi_k) + \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{1}{2}w_{ik}\log(|\Sigma_k|) + \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{m}{2}\log(2\pi)w_{ik} \\ &+ \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{1}{2}w_{ik}(x_i - \alpha - \beta z_k - \Gamma v_i)^T\Sigma_k^{-1}(x_i - \alpha - \beta z_k - \Gamma v_i). \end{aligned} \tag{A.2}$$

A.1 Derivation for $\hat{\pi}_k$

We are under the constraint $\sum_{k=1}^{K}\pi_k = 1$, and this can be addressed by applying a Lagrange multiplier. Define,

$$l(\pi_k) = l_c - \lambda(\sum_{k=1}^{K}\pi_k - 1),$$

then by taking the partial derivative of $l(\pi_k)$ with respect to π_k and letting it to be zero, we obtain,

$$\sum_{i=1}^{n} w_{ik}\frac{1}{\pi_k} - \lambda = 0,$$

then,

$$\pi_k = \frac{\sum_{i=1}^{n} w_{ik}}{\lambda},$$

take the summation over k on both sides, we obtain,

$$\sum_{k=1}^{K} \pi_k = \frac{\sum_{k=1}^{K} \sum_{i=1}^{n} w_{ik}}{\lambda} = 1,$$

since $\sum_{k=1}^{K} \sum_{i=1}^{n} w_{ik} = n$, so,

$$\lambda = n,$$

then we obtain,

$$\hat{\pi}_k = \frac{\sum_{i=1}^{n} w_{ik}}{n}. \tag{A.3}$$

A.2 Derivation for $\hat{\alpha}$

We use the result for the derivatives of matrices, vectors, and scalars, where W is symmetric, which is derived by Petersen and Pedersen (2012),

$$\frac{\partial}{\partial s}(x-s)^T W(x-s) = -2W(x-s).$$

We obtain the following by taking the partial derivative of the log-likelihood with respect to α,

$$\frac{\partial l_c}{\partial \alpha} = \sum_{i=1}^{n} \sum_{k=1}^{K} -\frac{1}{2} w_{ik}(-2)(\Sigma_k)^{-1}(x_i - \alpha - \beta z_k - \Gamma v_i),$$

then letting it to be zero and solving it,

$$\sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \Sigma_k^{-1}(x_i - \alpha - \beta z_k - \Gamma v_i) = 0, \tag{A.4}$$

$$\sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \Sigma_k^{-1}(x_i - \beta z_k - \Gamma v_i) = \alpha \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \Sigma_k^{-1},$$

we obtain the estimator for α,

$$\hat{\alpha} = \left(\sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \hat{\Sigma}_k^{-1} \right)^{-1} \left(\sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \hat{\Sigma}_k^{-1}(x_i - \hat{\beta}\hat{z}_k - \hat{\Gamma} v_i) \right). \tag{A.5}$$

In our implementation of the ECM algorithm, we assume (only temporarily within each M-step before actually updating $\hat{\Sigma}_k$) that $\hat{\Sigma}_k \equiv \text{diag}(\sigma^2)$, for some constant σ^2 which does not need to

be specified since it cancels out from the resulting simplified update equations, then Equation (A.4) becomes:

$$\Sigma^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(x_i - \alpha - \beta z_k - \Gamma v_i) = 0,$$

and then multiply Σ on both sides, we obtain,

$$\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(x_i - \alpha - \beta z_k - \Gamma v_i) = 0,$$

then,

$$\hat{\alpha} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i - \hat{\beta}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z_k} - \sum_{i=1}^{n}\hat{\Gamma}v_i\right). \tag{A.6}$$

This is the estimator of α used in the M-step in implementing the ECM algorithm.

A.3 Derivation for $\hat{\beta}$

For the derivation of β, we use the following result, which is derived by Petersen and Pedersen (2012),

$$\frac{\partial}{\partial A}(x - As)^T W(x - As) = -2W(x - As)s^T.$$

By taking partial derivative of the l_c with respect to β, we obtain,

$$\frac{\partial l_c}{\partial \beta} = \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{1}{2} w_{ik}(-2)\Sigma_k^{-1}(x_i - \alpha - \beta z_k - \Gamma v_i)z_k^T.$$

Since z_k is a scalar, $z_k = z_k^T$, and by letting the above equation to be zero and solving it,

$$\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\Sigma_k^{-1}(x_i - \alpha - \Gamma v_i)z_k - \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\Sigma_k^{-1}\beta z_k^2 = 0, \tag{A.7}$$

then,

$$\hat{\beta} = \left(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{\Sigma}_k^{-1}\hat{z}_k^2\right)^{-1}\left(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{\Sigma}_k^{-1}(x_i - \hat{\alpha} - \hat{\Gamma}v_i)\hat{z}_k\right), \tag{A.8}$$

Again, we assume $\hat{\Sigma}_k \equiv \text{diag}(\sigma^2)$ in implementation, then Equation (A.7) can be rewritten as,

$$\Sigma^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(x_i - \alpha - \Gamma v_i)z_k - \Sigma^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\beta z_k^2 = 0,$$

multiply Σ on both sides, we could obtain,

$$\hat{\beta} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k x_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k) - \sum_{i=1}^{n}\hat{\Gamma}v_i \sum_{k=1}^{K} w_{ik}\hat{z}_k}{\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k^2 - \frac{1}{n}(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k)^2} + \frac{\frac{1}{n}(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k)(\sum_{i=1}^{n}\hat{\Gamma}v_i)}{\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k^2 - \frac{1}{n}(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k)^2},$$

which is being used in the R code.

A.4 Derivation for $\hat{z_k}$

For the derivation of z_k, we use the result of the following, which is derived by Petersen and Pedersen (2012),

$$\frac{\partial}{\partial s}(x - As)^T W(x - As) = -2A^T W(x - As).$$

By taking partial derivative of the l_c with respect to z_k, we obtain,

$$\frac{\partial l_c}{\partial z_k} = \sum_{i=1}^{n} -\frac{1}{2} w_{ik}(-2)\beta^T \Sigma_k^{-1}(x_i - \alpha - \beta z_k - \Gamma v_i),$$

then,

$$\sum_{i=1}^{n} w_{ik}\beta^T \Sigma_k^{-1}(x_i - \alpha - \beta z_k - \Gamma v_i) = 0, \tag{A.9}$$

we obtain,

$$\hat{z}_k = \frac{\sum_{i=1}^{n} w_{ik}\hat{\beta}^T \hat{\Sigma}_k^{-1}(x_i - \hat{\alpha} - \hat{\Gamma}v_i)}{\sum_{i=1}^{n} w_{ik}\hat{\beta}^T \hat{\Sigma}_k^{-1}\hat{\beta}}. \tag{A.10}$$

With the assumption of $\hat{\Sigma}_k \equiv \text{diag}(\sigma^2)$ in implementation, the Equation (A.9) becomes,

$$\sum_{i=1}^{n} w_{ik}\beta^T(\sigma^2 I_m)(x_i - \alpha - \beta z_k - \Gamma v_i) = 0,$$

then,

$$\sigma^2 \sum_{i=1}^{n} w_{ik}\beta^T I_m(x_i - \alpha - \beta z_k - \Gamma v_i) = 0,$$

where the σ^2 can be canceled out, and we will have,

$$\sum_{i=1}^{n} w_{ik}\beta^T(x_i - \alpha - \beta z_k - \Gamma v_i) = 0,$$

The estimator of z_k used in the implementation is the following,

$$\hat{z}_k = \frac{\hat{\beta}^T \sum_{i=1}^n w_{ik}(x_i - \hat{\alpha} - \hat{\Gamma} v_i)}{\hat{\beta}^T \hat{\beta} \sum_{i=1}^n w_{ik}}, \tag{A.11}$$

A.5 Derivation for $\hat{\Gamma}$

For the derivation of Γ, we use the result of the following, which is derived by Petersen and Pedersen (2012),

$$\frac{\partial}{\partial A}(x - As)^T W(x - As) = -2W(x - As)s^T,$$

By taking partial derivative of the l_c with respect to Γ, we obtain,

$$\frac{\partial l_c}{\partial \Gamma} = \sum_{i=1}^n \sum_{k=1}^K -\frac{1}{2} w_{ik}(-2)\Sigma_k^{-1}(x_i - \alpha - \beta z_k - \Gamma v_i)v_i^T,$$

Letting the above equation to be 0,

$$\sum_{i=1}^n \sum_{k=1}^K w_{ik}\Sigma_k^{-1}(x_i - \alpha - \beta z_k)v_i^T = \sum_{i=1}^n \sum_{k=1}^K w_{ik}\Sigma_k^{-1}\Gamma v_i v_i^T, \tag{A.12}$$

On the right-hand side of Equation (A.12), Σ_k^{-1} is a $m \times m$ matrix, Γ is a $m \times p$ matrix, v_i is a $p \times 1$ vector; we cannot take Γ out of the double summation and multiply the inverse of the double summation on both sides to get an analytical estimator for Γ. In our implementation, we have the assumption that $\hat{\Sigma}_k \equiv \text{diag}(\sigma^2)$, then Equation (A.12) becomes,

$$\Sigma^{-1} \sum_{i=1}^n \sum_{k=1}^K w_{ik}(x_i - \alpha - \beta z_k)v_i^T = \Sigma^{-1} \sum_{i=1}^n \sum_{k=1}^K w_{ik}\Gamma v_i v_i^T,$$

then we multiply Σ on both sides and take Γ out of the double summation,

$$\sum_{i=1}^n \sum_{k=1}^K w_{ik}(x_i - \alpha - \beta z_k)v_i^T = \Gamma \sum_{i=1}^n \sum_{k=1}^K w_{ik} v_i v_i^T,$$

we obtain,

$$\hat{\Gamma} = \left(\sum_{i=1}^n x_i v_i^T - \hat{\alpha} \sum_{i=1}^n v_i^T - \hat{\beta} \sum_{i=1}^n v_i^T \sum_{k=1}^K w_{ik}\hat{z}_k \right) \left(\sum_{i=1}^n v_i v_i^T \right)^{-1}. \tag{A.13}$$

A.6 Derivation for $\hat{\Sigma}_k$

For the derivation of Σ, we use the result of the following, which is derived by Petersen and Pedersen (2012),

$$\frac{\partial}{\partial W}(x-s)^T W(x-s) = (x-s)(x-s)^T,$$

and

$$\frac{\partial}{\partial W}\log(|W|) = (W^{-1})^T,$$

Rewrite (A.2) to be

$$\begin{aligned}\tilde{l_c} = &\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\log(\pi_k) + \sum_{i=1}^{n}\sum_{k=1}^{K}\frac{1}{2}w_{ik}\log(|\Sigma_k|^{-1}) + \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{m}{2}\log(2\pi)w_{ik} \\ &+ \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{1}{2}w_{ik}(x_i-\alpha-\beta z_k-\Gamma v_i)^T\Sigma_k^{-1}(x_i-\alpha-\beta z_k-\Gamma v_i).\end{aligned} \tag{A.14}$$

By taking partial derivative of the $\tilde{l_c}$ with respect to Σ_k^{-1} and by letting it to be zero, we obtain,

$$\sum_{i=1}^{n}\frac{1}{2}w_{ik}\left((\Sigma_k^{-1})^{-1}\right)^T + \sum_{i=1}^{n} -\frac{1}{2}w_{ik}(x_i-\alpha-\beta z_k-\Gamma v_i)(x_i-\alpha-\beta z_k-\Gamma v_i)^T = 0,$$

since Σ_k is symmetric, then $\Sigma_k^T = \Sigma_k$,

$$\sum_{i=1}^{n} w_{ik}\Sigma_k = \sum_{i=1}^{n} w_{ik}(x_i-\alpha-\beta z_k-\Gamma v_i)(x_i-\alpha-\beta z_k-\Gamma v_i)^T,$$

we obtain (variance parameterization(iv)),

$$\hat{\Sigma}_k = \frac{\sum_{i=1}^{n} w_{ik}(x_i-\hat{\alpha}-\hat{\beta}\hat{z}_k-\hat{\Gamma}v_i)(x_i-\hat{\alpha}-\hat{\beta}\hat{z}_k-\hat{\Gamma}v_i)^T}{\sum_{i=1}^{n} w_{ik}}. \tag{A.15}$$

A.7 Derivation for $\hat{\sigma}_{jk}^2$

When $\Sigma_k \in R^m$ is diagonal, that is $\Sigma_k = \text{diag}(\sigma_{jk}^2)_{\{1 \le j \le m\}}$, where $k = 1, \ldots, K$,

$$\Sigma_k = \begin{pmatrix} \sigma_{1k}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{2k}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{jk}^2 \end{pmatrix}, \tag{A.16}$$

and $|\Sigma_k| = \prod_{j=1}^m \sigma_{jk}^2$, since $|\Sigma_k|^{-1} = |\Sigma_k^{-1}|$,

$$|\Sigma_k^{-1}| = \begin{vmatrix} \frac{1}{\sigma_{1k}^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_{2k}^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_{jk}^2} \end{vmatrix} = \prod_{j=1}^m \frac{1}{\sigma_{jk}^2}, \tag{A.17}$$

The log-likelihood function from the previous section is the following,

$$\begin{aligned} \tilde{l}_c = \sum_{i=1}^n \sum_{k=1}^K w_{ik} \log(\pi_k) + \sum_{i=1}^n \sum_{k=1}^K \frac{1}{2} w_{ik} \log(|\Sigma_k|^{-1}) + \sum_{i=1}^n \sum_{k=1}^K -\frac{m}{2} \log(2\pi) w_{ik} \\ + \sum_{i=1}^n \sum_{k=1}^K -\frac{1}{2} w_{ik} (x_i - \alpha - \beta z_k - \Gamma v_i)^T \Sigma_k^{-1} (x_i - \alpha - \beta z_k - \Gamma v_i), \end{aligned}$$

and $\log(|\Sigma_k|^{-1})$ will become ,

$$\log(|\Sigma_k|^{-1}) = \log(|\Sigma_k^{-1}|) = \log(\frac{1}{\sigma_{1k}^2} . \frac{1}{\sigma_{2k}^2} \ldots \frac{1}{\sigma_{mk}^2}) = -2 \sum_{j=1}^m \log \sigma_{jk},$$

let $\varphi_i = \Gamma v_i$, where $\varphi_i \in R^m$, then the log-likelihood function $\tilde{l}_c$ will become,

$$\tilde{l}_{new} = constant + \sum_{i=1}^n -\frac{1}{2} w_{ik} (-2) \sum_{j=1}^m \log \sigma_{jk} + \sum_{i=1}^n \sum_{j=1}^m -\frac{1}{2} w_{ik} \frac{(x_{ij} - \alpha_j - \beta_j z_k - \varphi_{ij})^2}{\sigma_{jk}^2},$$

by taking partial derivative of the $\tilde{l}_{new}$ with respect to σ_{jk} and by letting it to be zero, we obtain,

$$\sum_{i=1}^{n} -w_{ik}\frac{1}{\sigma_{jk}} + \sum_{i=1}^{n} w_{ik}(x_{ij} - \alpha_j - \beta_j z_k - \varphi_{ij})^2 \sigma_{jk}^{-3} = 0,$$

then,

$$\sum_{i=1}^{n} w_{ik}\frac{1}{\sigma_{jk}} = \frac{1}{\sigma_{jk}^3}\sum_{i=1}^{n} w_{ik}(x_{ij} - \alpha_j - \beta_j z_k - \varphi_{ij})^2,$$

we then obtain variance parameterization (ii),

$$\hat{\sigma}_{jk}^2 = \frac{\sum_{i=1}^{n} w_{ik}(x_{ij} - \hat{\alpha}_j - \hat{\beta}_j \hat{z}_k - \hat{\varphi}_{ij})^2}{\sum_{i=1}^{n} w_{ik}}. \tag{A.18}$$

A.8 Derivation for $\hat{\Sigma}$

For the derivation of parameter Σ, again, we use the following results, which is derived by Petersen and Pedersen (2012),

$$\frac{\partial}{\partial W}(x-s)^T W(x-s) = (x-s)(x-s)^T,$$

and

$$\frac{\partial}{\partial W}\log(|W|) = (W^{-1})^T.$$

When $\Sigma_k \equiv \Sigma$, the log-likelihood function (A.2) can be rewrite as,

$$\begin{aligned}\tilde{l_c} = &\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\log(\pi_k) + \sum_{i=1}^{n}\sum_{k=1}^{K}\frac{1}{2}w_{ik}\log(|\Sigma|^{-1}) + \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{m}{2}\log(2\pi)w_{ik} \\ &+ \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{1}{2}w_{ik}(x_i - \alpha - \beta z_k - \Gamma v_i)^T\Sigma^{-1}(x_i - \alpha - \beta z_k - \Gamma v_i).\end{aligned} \tag{A.19}$$

By taking partial derivative of the $\tilde{l_c}$ with respect to Σ^{-1} and by letting it to be zero, we obtain,

$$\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{1}{2}w_{ik}\left((\Sigma^{-1})^{-1}\right)^T + \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{1}{2}w_{ik}(x_i - \alpha - \beta z_k - \Gamma v_i)(x_i - \alpha - \beta z_k - \Gamma v_i)^T = 0,$$

since Σ_k is symmetric, and $\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik} = n$ then we obtain variance parameterization (iii),

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(x_i - \hat{\alpha} - \hat{\beta}\hat{z}_k - \hat{\Gamma}v_i)(x_i - \hat{\alpha} - \hat{\beta}\hat{z}_k - \hat{\Gamma}v_i)^T. \tag{A.20}$$

A.9 Derivation for $\hat{\sigma}_j$

When $\Sigma_{m\times m}$ is diagonal, that is $\Sigma = \text{diag}(\sigma_j^2)_{\{1\le j\le m\}}$,

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_j^2 \end{pmatrix}, \tag{A.21}$$

and $|\Sigma| = \prod_{j=1}^{m}\sigma_j^2$, since $|\Sigma|^{-1} = |\Sigma^{-1}|$,

$$|\Sigma^{-1}| = \begin{vmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_j^2} \end{vmatrix} = \prod_{j=1}^{m}\frac{1}{\sigma_j^2}, \tag{A.22}$$

The log-likelihood function from the previous section is the following,

$$\begin{aligned}\tilde{l}_c = &\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\log(\pi_k) + \sum_{i=1}^{n}\sum_{k=1}^{K}\frac{1}{2}w_{ik}\log(|\Sigma|^{-1}) + \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{m}{2}\log(2\pi)w_{ik} \\ &+ \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{1}{2}w_{ik}(x_i - \alpha - \beta z_k - \Gamma v_i)^T\Sigma^{-1}(x_i - \alpha - \beta z_k - \Gamma v_i),\end{aligned}$$

and $\log(|\Sigma|^{-1})$ will become,

$$\log(|\Sigma|^{-1}) = \log(|\Sigma^{-1}|) = \log(\frac{1}{\sigma_1^2}.\frac{1}{\sigma_2^2}\ldots\frac{1}{\sigma_m^2}) = -2\sum_{j=1}^{m}\log\sigma_j,$$

let $\varphi_i = \Gamma v_i$, where $\varphi_i \in R^m$, then the log-likelihood function $\tilde{l}_c$ will become,

$$\tilde{l}_{new} = constant + \sum_{i=1}^{n}\sum_{k=1}^{K} -\frac{1}{2} w_{ik}(-2) \sum_{j=1}^{m} \log \sigma_j + \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{m} -\frac{1}{2} w_{ik} \frac{(x_{ij} - \alpha_j - \beta_j z_k - \varphi_{ij})^2}{\sigma_j^2},$$

by taking partial derivative of the $\tilde{l}_{new}$ with respect to σ_j and by letting it to be zero, we obtain,

$$\sum_{i=1}^{n}\sum_{k=1}^{K} -w_{ik}\frac{1}{\sigma_j} + \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(x_{ij} - \alpha_j - \beta_j z_k - \varphi_{ij})^2 \sigma_j^{-3} = 0,$$

since $\sum_{i=1}^{n}\sum_{k=1}^{K} -w_{ik} = n$,

$$\frac{n}{\sigma_j} = \frac{1}{\sigma_j^3} \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(x_{ij} - \alpha_j - \beta_j z_k - \varphi_{ij})^2,$$

we then obtain variance parameterization (i),

$$\hat{\sigma}_j^2 = \frac{\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(x_{ij} - \hat{\alpha}_j - \hat{\beta}_j \hat{z}_k - \hat{\varphi}_{ij})^2}{n} \tag{A.23}$$

Appendix B

Derivations of two-level model with covariates

The derivations of the parameter estimators for the two-level model with covariates,

$$x_{ij} = \alpha + \beta z_i + \Gamma v_{ij} + \varepsilon_{ij}, \tag{B.1}$$

where $x_{ij} \in R^m$, the upper-level unit is indexed by $i = 1, \ldots, r$, and the lower-level unit is indexed by $j = 1, \ldots, n_r$, and $\varepsilon_{ij} \sim N(0, \Sigma(z_i))$ are independent Gaussian errors..

The the expected complete log-likelihood of model (B.1) is the following,

$$\begin{aligned} l_c = &\sum_{i=1}^{r}\sum_{k=1}^{K} w_{ik}\log(\pi_k) + \sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{K} -\frac{1}{2}w_{ik}\log(|\Sigma_k|) + \sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{K} -\frac{m}{2}\log(2\pi)w_{ik} \\ &+ \sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{K} -\frac{1}{2}w_{ik}(x_{ij} - \alpha - \beta z_k - \Gamma v_{ij})^T\Sigma_k^{-1}(x_{ij} - \alpha - \beta z_k - \Gamma v_{ij}). \end{aligned} \tag{B.2}$$

B.1 Derivation for $\hat{\pi}_k$

We are under the constraint $\sum_{k=1}^{K}\pi_k = 1$, and this can be addressed by applying a Lagrange multiplier. Define,

$$L(\pi_k) = l_c - \lambda(\sum_{k=1}^{K}\pi_k - 1),$$

then by taking the partial derivative of $L(\pi_k)$ with respect to π_k and letting it to be zero, we obtain,

$$\sum_{i=1}^{r} w_{ik}\frac{1}{\pi_k} - \lambda = 0,$$

then,

$$\pi_k = \frac{\sum_{i=1}^{r} w_{ik}}{\lambda},$$

take the summation over k on both sides, we obtain,

$$\sum_{k=1}^{K} \pi_k = \frac{\sum_{k=1}^{K} \sum_{i=1}^{r} w_{ik}}{\lambda} = 1,$$

since $\sum_{k=1}^{K} \sum_{i=1}^{n} w_{ik} = r$, so,

$$\lambda = r,$$

then we obtain,

$$\hat{\pi}_k = \frac{\sum_{i=1}^{r} w_{ik}}{r}. \tag{B.3}$$

B.2 Derivation for $\hat{\alpha}$

Using the result of the derivatives of matrices, vectors, and scalars, where W is symmetric, which is derived by Petersen and Pedersen (2012),

$$\frac{\partial}{\partial s}(x-s)^T W(x-s) = -2W(x-s).$$

We obtain the following by taking the partial derivative of the log-likelihood with respect to α,

$$\frac{\partial l_c}{\partial \alpha} = \sum_{i=1}^{r} \sum_{j=1}^{n_r} \sum_{k=1}^{K} -\frac{1}{2} w_{ik}(-2)(\Sigma_k)^{-1}(x_{ij} - \alpha - \beta z_k - \Gamma v_{ij}),$$

then letting it to be zero and solving it,

$$\sum_{i=1}^{r} \sum_{j=1}^{n_r} \sum_{k=1}^{K} w_{ik} \Sigma_k^{-1}(x_{ij} - \alpha - \beta z_k - \Gamma v_{ij}) = 0, \tag{B.4}$$

$$\sum_{i=1}^{r} \sum_{j=1}^{n_r} \sum_{k=1}^{K} w_{ik} \Sigma_k^{-1}(x_{ij} - \beta z_k - \Gamma v_{ij}) = \alpha \sum_{i=1}^{r} \sum_{j=1}^{n_r} \sum_{k=1}^{K} w_{ik} \Sigma_k^{-1},$$

we obtain the estimator for α,

$$\hat{\alpha} = \left(\sum_{i=1}^{r} \sum_{j=1}^{n_r} \sum_{k=1}^{K} w_{ik} \hat{\Sigma}_k^{-1} \right)^{-1} \left(\sum_{i=1}^{r} \sum_{j=1}^{n_r} \sum_{k=1}^{K} w_{ik} \hat{\Sigma}_k^{-1}(x_{ij} - \hat{\beta}\hat{z}_k - \hat{\Gamma} v_{ij}) \right). \tag{B.5}$$

In our implementation of the ECM algorithm, we (just for the use within these equations) assume that $\hat{\Sigma}_k \equiv \text{diag}(\sigma^2)$, for some constant σ^2 which does not need to be specified since it cancels out from the resulting simplified update equations, then Equation (B.4) becomes:

$$\Sigma^{-1} \sum_{i=1}^{r} \sum_{j=1}^{n_r} \sum_{k=1}^{K} w_{ik}(x_{ij} - \alpha - \beta z_k - \Gamma v_{ij}) = 0,$$

and then multiply Σ on both sides, we obtain,

$$\sum_{i=1}^{r} \sum_{j=1}^{n_r} \sum_{k=1}^{K} w_{ik}(x_{ij} - \alpha - \beta z_k - \Gamma v_{ij}) = 0,$$

then,

$$\hat{\alpha} = \frac{1}{n} \left(\sum_{i=1}^{r} \sum_{j=1}^{n_r} x_{ij} - \hat{\beta} \sum_{i=1}^{r} n_i \sum_{k=1}^{K} w_{ik} \hat{z}_k - \hat{\Gamma} \sum_{i=1}^{r} \sum_{j=1}^{n_r} v_{ij} \right). \tag{B.6}$$

This is the estimator of α used in the M-step in implementing the ECM algorithm.

B.3 Derivation for $\hat{\beta}$

For the derivation of β, we use the result of the following, which is derived by Petersen and Pedersen (2012),

$$\frac{\partial}{\partial A}(x - As)^T W(x - As) = -2W(x - As)s^T$$

By taking partial derivative of the l_c with respect to β, we obtain,

$$\frac{\partial l_c}{\partial \beta} = \sum_{i=1}^{r} \sum_{j=1}^{n_r} \sum_{k=1}^{K} -\frac{1}{2} w_{ik}(-2)\Sigma_k^{-1}(x_{ij} - \alpha - \beta z_k - \Gamma v_{ij}) z_k^T,$$

Since z_k is a scalar, $z_k = z_k^T$, and by letting the above equation to be zero and solving it,

$$\sum_{i=1}^{r} \sum_{j=1}^{n_r} \sum_{k=1}^{K} w_{ik}\Sigma_k^{-1}(x_{ij} - \alpha - \Gamma v_{ij}) z_k - \sum_{i=1}^{r} \sum_{j=1}^{n_r} \sum_{k=1}^{K} w_{ik}\Sigma_k^{-1}\beta z_k^2 = 0, \tag{B.7}$$

then,

$$\hat{\beta} = \left(\sum_{i=1}^{r} \sum_{j=1}^{n_r} \sum_{k=1}^{K} w_{ik}\hat{\Sigma}_k^{-1}\hat{z}_k^2 \right)^{-1} \left(\sum_{i=1}^{r} \sum_{j=1}^{n_r} \sum_{k=1}^{K} w_{ik}\hat{\Sigma}_k^{-1}(x_{ij} - \hat{\alpha} - \hat{\Gamma} v_{ij})\hat{z}_k \right), \tag{B.8}$$

Again, we assume $\hat{\Sigma}_k \equiv \text{diag}(\sigma^2)$ in implementation, then Equation (B.7) can be rewritten as,

$$\Sigma^{-1}\sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K} w_{ik}(x_{ij}-\alpha-\Gamma v_i)z_k - \Sigma^{-1}\sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K} w_{ik}\beta z_k^2 = 0,$$

multiply Σ on both sides, we could obtain,

$$\hat{\beta} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{K} w_{ik}\hat{z}_k x_{ij} - \frac{1}{n}(\sum_{i=1}^{r}\sum_{j=1}^{n_i} x_{ij})(\sum_{i=1}^{r} n_i \sum_{k=1}^{K} w_{ik}\hat{z}_k)}{\sum_{i=1}^{r} n_i \sum_{k=1}^{K} w_{ik}\hat{z}_k^2 - \frac{1}{n}(\sum_{i=1}^{r} n_i \sum_{k=1}^{K} w_{ik}\hat{z}_k)^2} - \frac{\hat{\Gamma}\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{K} w_{ik}\hat{z}_k v_{ij} - \frac{1}{n}(\sum_{i=1}^{r} n_i \sum_{k=1}^{K} w_{ik}\hat{z}_k)(\hat{\Gamma}\sum_{i=1}^{r}\sum_{j=1}^{n_i} v_{ij})}{\sum_{i=1}^{r} n_i \sum_{k=1}^{K} w_{ik}\hat{z}_k^2 - \frac{1}{n}(\sum_{i=1}^{r} n_i \sum_{k=1}^{K} w_{ik}\hat{z}_k)^2}, \quad \text{(B.9)}$$

which is being used in the R code.

B.4 Derivation for $\hat{z}_k$

For the derivation of z_k, we use the result of the following, which is derived by Petersen and Pedersen (2012),

$$\frac{\partial}{\partial s}(x-As)^T W(x-As) = -2A^T W(x-As),$$

By taking partial derivative of the l_c with respect to z_k, we obtain,

$$\frac{\partial l_c}{\partial z_k} = \sum_{i=1}^{r}\sum_{j=1}^{n_r} -\frac{1}{2} w_{ik}(-2)\beta^T\Sigma_k^{-1}(x_{ij}-\alpha-\beta z_k-\Gamma v_{ij}),$$

then,

$$\sum_{i=1}^{r}\sum_{j=1}^{n_r} w_{ik}\beta^T\Sigma_k^{-1}(x_{ij}-\alpha-\beta z_k-\Gamma v_{ij}) = 0, \quad \text{(B.10)}$$

we obtain,

$$\hat{z}_k = \frac{\sum_{i=1}^{r}\sum_{j=1}^{n_r} w_{ik}\hat{\beta}^T\hat{\Sigma}_k^{-1}(x_{ij}-\hat{\alpha}-\hat{\Gamma}v_{ij})}{\sum_{i=1}^{r}\sum_{j=1}^{n_r} w_{ik}\hat{\beta}^T\hat{\Sigma}_k^{-1}\hat{\beta}}. \quad \text{(B.11)}$$

With the assumption of $\hat{\Sigma}_k \equiv \text{diag}(\sigma^2)$ in implementation, the Equation (B.10) becomes,

$$\sum_{i=1}^{r}\sum_{j=1}^{n_r} w_{ik}\beta^T(\sigma^2 I_m)(x_{ij}-\alpha-\beta z_k-\Gamma v_{ij}) = 0,$$

then,

$$\sigma^2 \sum_{i=1}^{r} \sum_{j=1}^{n_r} w_{ik} \beta^T I_m (x_{ij} - \alpha - \beta z_k - \Gamma v_{ij}) = 0,$$

where the σ^2 can be canceled out, and we will have,

$$\sum_{i=1}^{r} \sum_{j=1}^{n_r} w_{ik} \beta^T (x_{ij} - \alpha - \beta z_k - \Gamma v_{ij}) = 0,$$

The estimator of z_k used in the implementation is the following,

$$\hat{z}_k = \frac{\hat{\beta}^T \sum_{i=1}^{r} \sum_{j=1}^{n_r} w_{ik} (x_{ij} - \hat{\alpha} - \hat{\Gamma} v_{ij})}{\hat{\beta}^T \hat{\beta} \sum_{i=1}^{r} n_i w_{ik}}, \tag{B.12}$$

B.5 Derivation for $\hat{\Gamma}$

For the derivation of Γ, we use the result of the following, which is derived by Petersen and Pedersen (2012),

$$\frac{\partial}{\partial A} (x - As)^T W (x - As) = -2W(x - As)s^T,$$

By taking partial derivative of the l_c with respect to Γ, we obtain,

$$\frac{\partial l_c}{\partial \Gamma} = \sum_{i=1}^{r} \sum_{j=1}^{n_r} \sum_{k=1}^{K} -\frac{1}{2} w_{ik} (-2) \Sigma_k^{-1} (x_{ij} - \alpha - \beta z_k - \Gamma v_{ij}) v_{ij}^T,$$

Letting the above equation to be 0,

$$\sum_{i=1}^{r} \sum_{j=1}^{n_r} \sum_{k=1}^{K} w_{ik} \Sigma_k^{-1} (x_{ij} - \alpha - \beta z_k) v_{ij}^T = \sum_{i=1}^{r} \sum_{j=1}^{n_r} \sum_{k=1}^{K} w_{ik} \Sigma_k^{-1} \Gamma v_{ij} v_{ij}^T, \tag{B.13}$$

On the right-hand side of Equation (B.13), Σ_k^{-1} is a $m \times m$ matrix, Γ is a $m \times p$ matrix, v_i is a $p \times 1$ vector; we cannot take Γ out of the double summation and multiply the inverse of the double summation on both sides to get an analytical estimator for Γ. In our implementation, we have the assumption that $\hat{\Sigma}_k \equiv \text{diag}(\sigma^2)$, then Equation (B.13) becomes,

$$\Sigma^{-1} \sum_{i=1}^{r} \sum_{j=1}^{n_r} \sum_{k=1}^{K} w_{ik} (x_{ij} - \alpha - \beta z_k) v_{ij}^T = \Sigma^{-1} \sum_{i=1}^{r} \sum_{j=1}^{n_r} \sum_{k=1}^{K} w_{ik} \Gamma v_{ij} v_{ij}^T,$$

then we multiply Σ on both sides and take Γ out of the double summation,

$$\sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K} w_{ik}(x_{ij}-\alpha-\beta z_k)v_{ij}^T = \Gamma\sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K} w_{ik}v_{ij}v_{ij}^T,$$

we obtain,

$$\hat{\Gamma} = \left(\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{K} w_{ik}(x_{ij}-\hat{\alpha}-\hat{\beta}\hat{z}_k)v_{ij}^T\right)\left(\sum_{i=1}^{r}\sum_{j=1}^{n_i} v_{ij}v_{ij}^T\right)^{-1}. \tag{B.14}$$

B.6 Derivation for $\hat{\Sigma}_k$

For the derivation of Σ, we use the result of the following, which is derived by Petersen and Pedersen (2012),

$$\frac{\partial}{\partial W}(x-s)^T W(x-s) = (x-s)(x-s)^T,$$

and

$$\frac{\partial}{\partial W}\log(|W|) = (W^{-1})^T,$$

Rewrite (B.2) to be

$$\begin{aligned}\tilde{l_c} = &\sum_{i=1}^{r}\sum_{k=1}^{K} w_{ik}\log(\pi_k) + \sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K}\frac{1}{2}w_{ik}\log(|\Sigma_k|^{-1}) + \sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K} -\frac{m}{2}\log(2\pi)w_{ik}\\ &+\sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K} -\frac{1}{2}w_{ik}(x_{ij}-\alpha-\beta z_k-\Gamma v_{ij})^T\Sigma_k^{-1}(x_{ij}-\alpha-\beta z_k-\Gamma v_{ij}).\end{aligned} \tag{B.15}$$

By taking partial derivative of the $\tilde{l_c}$ with respect to Σ_k^{-1} and by letting it to be zero, we obtain,

$$\sum_{i=1}^{r}\sum_{j=1}^{n_r}\frac{1}{2}w_{ik}\left((\Sigma_k^{-1})^{-1}\right)^T + \sum_{i=1}^{r}\sum_{j=1}^{n_r} -\frac{1}{2}w_{ik}(x_{ij}-\alpha-\beta z_k-\Gamma v_{ij})(x_{ij}-\alpha-\beta z_k-\Gamma v_{ij})^T = 0,$$

since Σ_k is symmetric, then $\Sigma_k^T = \Sigma_k$,

$$\sum_{i=1}^{r}\sum_{j=1}^{n_r} w_{ik}\Sigma_k = \sum_{i=1}^{r}\sum_{j=1}^{n_r} w_{ik}(x_{ij}-\alpha-\beta z_k-\Gamma v_{ij})(x_{ij}-\alpha-\beta z_k-\Gamma v_{ij})^T,$$

we obtain (variance parameterization(iv)),

$$\hat{\Sigma}_k = \frac{\sum_{i=1}^{r}\sum_{j=1}^{n_i} w_{ik}(x_{ij} - \hat{\alpha} - \hat{\beta}\hat{z}_k - \hat{\Gamma}v_{ij})(x_{ij} - \hat{\alpha} - \hat{\beta}\hat{z}_k - \hat{\Gamma}v_{ij})^T}{\sum_{i=1}^{r} n_i w_{ik}}. \tag{B.16}$$

B.7 Derivation for $\hat{\sigma}^2_{lk}$

When $\Sigma_k \in R^m$ is diagonal, that is $\Sigma_k = \text{diag}(\sigma^2_{lk})_{\{1\leq l\leq m\}}$, where $k = 1, \ldots, K$,

$$\Sigma_k = \begin{pmatrix} \sigma^2_{1k} & 0 & \cdots & 0 \\ 0 & \sigma^2_{2k} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2_{lk} \end{pmatrix}, \tag{B.17}$$

and $|\Sigma_k| = \prod_{l=1}^{m} \sigma^2_{lk}$, since $|\Sigma_k|^{-1} = |\Sigma_k^{-1}|$,

$$|\Sigma_k^{-1}| = \begin{vmatrix} \frac{1}{\sigma^2_{1k}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma^2_{2k}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma^2_{lk}} \end{vmatrix} = \prod_{l=1}^{m} \frac{1}{\sigma^2_{lk}}, \tag{B.18}$$

The log-likelihood function from the previous section is the following,

$$\tilde{l_c} = \sum_{i=1}^{r}\sum_{k=1}^{K} w_{ik}\log(\pi_k) + \sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K}\frac{1}{2}w_{ik}\log(|\Sigma_k|^{-1}) + \sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K} -\frac{m}{2}\log(2\pi)w_{ik}$$
$$+ \sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K} -\frac{1}{2}w_{ik}(x_{ij} - \alpha - \beta z_k - \Gamma v_{ij})^T\Sigma_k^{-1}(x_{ij} - \alpha - \beta z_k - \Gamma v_{ij}),$$

and $\log(|\Sigma_k|^{-1})$ will become,

$$\log(|\Sigma_k|^{-1}) = \log(|\Sigma_k^{-1}|) = \log(\frac{1}{\sigma^2_{1k}}.\frac{1}{\sigma^2_{2k}}\ldots\frac{1}{\sigma^2_{lk}}) = -2\sum_{l=1}^{m}\log\sigma_{lk},$$

let $\varphi_{ij} = \Gamma v_{ij}$, where $\varphi_{ij} \in R^m$, then the log-likelihood function $\tilde{l}_c$ will become,

$$\tilde{l}_{new} = constant + \sum_{i=1}^{r}\sum_{j=1}^{n_r} -\frac{1}{2} w_{ik}(-2) \sum_{l=1}^{m} \log \sigma_{lk} + \sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{l=1}^{m} -\frac{1}{2} w_{ik} \frac{(x_{ijl} - \alpha_l - \beta_l z_k - \varphi_{ijl})^2}{\sigma_{jk}^2},$$

by taking partial derivative of the $\tilde{l}_{new}$ with respect to σ_{lk} and by letting it to be zero, we obtain,

$$\sum_{i=1}^{r}\sum_{j=1}^{n_r} -w_{ik}\frac{1}{\sigma_{lk}} + \sum_{i=1}^{r}\sum_{j=1}^{n_r} w_{ik}(x_{ijl} - \alpha_l - \beta_l z_k - \varphi_{ijl})^2 \sigma_{lk}^{-3} = 0,$$

then,

$$\sum_{i=1}^{r}\sum_{j=1}^{n_r} w_{ik}\frac{1}{\sigma_{lk}} = \frac{1}{\sigma_{lk}^3}\sum_{i=1}^{r}\sum_{j=1}^{n_r} w_{ik}(x_{ijl} - \alpha_l - \beta_l z_k - \varphi_{ijl})^2,$$

we then obtain varince parameterization (ii),

$$\hat{\sigma}_{lk}^2 = \frac{\sum_{i=1}^{r}\sum_{j=1}^{n_i} w_{ik}(x_{ijl} - \hat{\alpha}_l - \hat{\beta}_l \hat{z}_k - \hat{\phi}_{ijl})^2}{\sum_{i=1}^{r} n_i w_{ik}}. \tag{B.19}$$

B.8 Derivation for $\hat{\Sigma}$

For the derivation of parameter Σ, again, we use the following results, which is derived by Petersen and Pedersen (2012),

$$\frac{\partial}{\partial W}(x-s)^T W (x-s) = (x-s)(x-s)^T,$$

and

$$\frac{\partial}{\partial W}\log(|W|) = (W^{-1})^T.$$

When $\Sigma_k \equiv \Sigma$, the log-likelihood function (B.15) can be rewrite as,

$$\begin{aligned}\tilde{l}_c = &\sum_{i=1}^{r}\sum_{k=1}^{K} w_{ik}\log(\pi_k) + \sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K}\frac{1}{2} w_{ik}\log(|\Sigma|^{-1}) + \sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K} -\frac{m}{2}\log(2\pi) w_{ik} \\ &+ \sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K} -\frac{1}{2} w_{ik}(x_{ij} - \alpha - \beta z_k - \Gamma v_{ij})^T \Sigma^{-1}(x_{ij} - \alpha - \beta z_k - \Gamma v_{ij}).\end{aligned} \tag{B.20}$$

By taking partial derivative of the $\tilde{l_c}$ with respect to Σ^{-1} and by letting it to be zero, we obtain,

$$\sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K}\frac{1}{2}w_{ik}\left((\Sigma^{-1})^{-1}\right)^T+\sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K}-\frac{1}{2}w_{ik}(x_{ij}-\alpha-\beta z_k-\Gamma v_{ij})(x_{ij}-\alpha-\beta z_k-\Gamma v_{ij})^T=0,$$

since Σ_k is symmetric, and $\sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K}w_{ik}=n$ then we obtain variance parameterization (iii),

$$\hat{\Sigma}=\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K}w_{ik}(x_{ij}-\hat{\alpha}-\hat{\beta}\hat{z}_k-\hat{\Gamma}v_{ij})(x_{ij}-\hat{\alpha}-\hat{\beta}\hat{z}_k-\hat{\Gamma}v_{ij})^T. \tag{B.21}$$

B.9 Derivation for $\hat{\sigma}_l$

When $\Sigma_{m\times m}$ is diagonal, that is $\Sigma=\text{diag}(\sigma_l^2)_{\{1\le l\le m\}}$,

$$\Sigma=\begin{pmatrix}\sigma_1^2 & 0 & \cdots & 0\\ 0 & \sigma_2^2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \sigma_l^2\end{pmatrix}, \tag{B.22}$$

and $|\Sigma|=\prod_{l=1}^{m}\sigma_l^2$, since $|\Sigma|^{-1}=|\Sigma^{-1}|$,

$$|\Sigma^{-1}|=\begin{vmatrix}\frac{1}{\sigma_1^2} & 0 & \cdots & 0\\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \frac{1}{\sigma_l^2}\end{vmatrix}=\prod_{l=1}^{m}\frac{1}{\sigma_l^2}, \tag{B.23}$$

The log-likelihood function from the previous section is the following,

$$\begin{aligned}\tilde{l_c}=&\sum_{i=1}^{r}\sum_{k=1}^{K}w_{ik}\log(\pi_k)+\sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K}\frac{1}{2}w_{ik}\log(|\Sigma|^{-1})+\sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K}-\frac{m}{2}\log(2\pi)w_{ik}\\&+\sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K}-\frac{1}{2}w_{ik}(x_{ij}-\alpha-\beta z_k-\Gamma v_{ij})^T\Sigma^{-1}(x_{ij}-\alpha-\beta z_k-\Gamma v_{ij}),\end{aligned}$$

and $\log(|\Sigma|^{-1})$ will be,

$$\log(|\Sigma|^{-1}) = \log(|\Sigma^{-1}|) = \log(\frac{1}{\sigma_1^2}.\frac{1}{\sigma_2^2}\ldots\frac{1}{\sigma_l^2}) = -2\sum_{l=1}^{m}\log\sigma_l,$$

let $\varphi_{ij} = \Gamma v_{ij}$, where $\varphi_{ij} \in R^m$, then the log-likelihood function $\tilde{l}_c$ will become,

$$\tilde{l}_{new} = constant + \sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K} -\frac{1}{2}w_{ik}(-2)\sum_{l=1}^{m}\log\sigma_l + \sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K}\sum_{l=1}^{m} -\frac{1}{2}w_{ik}\frac{(x_{ijl}-\alpha_l-\beta_l z_k-\varphi_{ijl})^2}{\sigma_l^2},$$

by taking partial derivative of the $\tilde{l}_{new}$ with respect to σ_l and by letting it to be zero, we obtain,

$$\sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K} -w_{ik}\frac{1}{\sigma_l} + \sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K} w_{ik}(x_{ijl}-\alpha_l-\beta_l z_k-\varphi_{ijl})^2\sigma_l^{-3} = 0,$$

since $\sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K} w_{ik} = n$,

$$\frac{n}{\sigma_l} = \frac{1}{\sigma_l^3}\sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K} w_{ik}(x_{ijl}-\alpha_l-\beta_l z_k-\varphi_{ijl})^2,$$

we then obtain variance parameterization (i),

$$\hat{\sigma_l}^2 = \frac{\sum_{i=1}^{r}\sum_{j=1}^{n_r}\sum_{k=1}^{K} w_{ik}(x_{ijl}-\hat{\alpha}_l-\hat{\beta}_l\hat{z}_k-\hat{\varphi}_{ijl})^2}{n} \tag{B.24}$$

Note that we only used variance parameterization (i) and (ii) in the implementation of the ECM algorithm for practical reasons.

Bibliography

Aitkin, M. (1996a). Empirical Bayes shrinkage using posterior random effect means from non-parametric maximum likelihood estimation in general random effect models. *Statistical Modelling: Proceedings of the 11th IWSM*, 87–94.

Aitkin, M. (1996b). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and computing*, *6*, 251–262.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, *55*(1), 117–128.

Aitkin, M., Anderson, D., & Hinde, J. (1981). Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society: Series A (General)*, *144*(4), 419–448.

Aitkin, M., Francis, B., Hinde, J., & Darnell, R. (2009). *Statistical modelling in r*. Oxford University Press.

Aitkin, M., & Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society: Series A (General)*, *149*(1), 1–26.

Almohaimeed, A., & Einbeck, J. (2022). Response transformations for random effect and variance component models. *Statistical Modelling*, *22*(4), 297–326.

Banfield, J. D., & Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, 803–821.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Biernacki, C., Celeux, G., & Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics & Data Analysis*, *41*(3-4), 561–575.

Bishop, C. M., Svensén, M., & Williams, C. K. (1998). Gtm: The generative topographic mapping. *Neural computation*, *10*(1), 215–234.

Bouveyron, C., & Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, *71*, 52–78.

Cannoodt, R. (2018, June). Princurve 2.0: Fit a Principal Curve in Arbitrary Dimension [DOI: https://doi.org/10.5281/zenodo.3351282].

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

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, *39*(1), 1–22.

Di Mari, R., Bakk, Z., Oser, J., & Kuha, J. (2023). A two-step estimator for multilevel latent class analysis with covariates. *Psychometrika*, *88*(4), 1144–1170. https://doi.org/https://doi.org/10.1214/16-AOAS988

Drikvandi, R. (2020). Nonlinear mixed-effects models with misspecified random-effects distribution. *Pharmaceutical statistics*, *19*(3), 187–201.

Drikvandi, R., Verbeke, G., & Molenberghs, G. (2017). Diagnosing misspecification of the random-effects distribution in mixed models. *Biometrics*, *73*, 63–71.

Einbeck, J., Darnell, R., & Hinde, J. (2018). *Npmlreg: Nonparametric maximum likelihood estimation for random effect models* [R package version 0.46-5]. https://CRAN.R-project.org/package=npmlreg

Einbeck, J., & Evers, L. (2024). *Lpcm: Local principal curve methods* [R package version 0.47-4]. https://CRAN.R-project.org/package=LPCM

Einbeck, J., Gray, E., Sofroniou, N., Marques da Silva Junior, A. H., & Gledhill, J. (2017). Confidence intervals for posterior intercepts, with application to the PIAAC literacy survey. In *Proceedings of the 32nd international workshop on statistical modelling* (pp. 217–2022). University of Groningen.

Einbeck, J., Tutz, G., & Evers, L. (2005). Local principal curves. *Statistics and Computing*, *15*, 301–313.

Fahrmeir, L., Tutz, G., Hennevogl, W., & Salem, E. (1994). *Multivariate statistical modelling based on generalized linear models* (Vol. 425). Springer.

Fan, J., Liao, Y., & Wang, W. (2016). Projected principal component analysis in factor models. *Annals of statistics*, *44*(1), 219.

Fox, J., Weisberg, S., & Price, B. (2020). *Cardata: Companion to applied regression data sets* [R package version 3.0-4]. https://CRAN.R-project.org/package=carData

Ghahramani, Z., & Hinton, G. E. (1996). *The EM algorithm for mixtures of factor analyzers* (tech. rep.). Technical Report CRG-TR-96-1, University of Toronto.

Gnaldi, M., Bacci, S., & Bartolucci, F. (2016). A multilevel finite mixture item response model to cluster examinees and schools. *Advances in Data Analysis and Classification*, *10*, 53–70.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*(2), 215–231.

Grilli, L., Pennoni, F., Rampichini, C., & Romeo, I. (2016). Exploiting TIMSS and PIRLS combined data: Multivariate multilevel modelling of student achievement. *The Annals of Applied Statistics*, *10*(4), 2405–2426.

Hastie, T., & Stuetzle, W. (1989). Principal curves. *Journal of the American statistical association*, *84*(406), 502–516.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). Statistical learning with sparsity. *Monographs on statistics and applied probability*, *143*(143), 8.

Krzanowski, W. (2000). *Principles of multivariate analysis* (Vol. 23). OUP Oxford.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, *73*(364), 805–811.

Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics*, 1350–1360.

Li, G., & Jung, S. (2017). Incorporating covariates into integrated factor analysis of multi-view data. *Biometrics*, *73*(4), 1433–1442.

Liang, K. Y., Zeger, S. L., & Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *54*(1), 3–24.

Lukočienė, O., Varriale, R., & Vermunt, J. K. (2010). The simultaneous decision (s) about the number of lower-and higher-level classes in multilevel latent class analysis. *Sociological Methodology*, *40*(1), 247–283.

Marques da Silva Júnior, A. H., Einbeck, J., & Craig, P. S. (2018). Fisher information under gaussian quadrature models. *Statistica Neerlandica*, *72*(2), 74–89.

Meng, X.-L., & Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, *80*(2), 267–278.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT press.

Organisation for Economic Co-operation and Development. (2023a). Main Elements of the Survey of Adult Skills [Accessed on 2023-05-29].

Organisation for Economic Co-operation and Development. (2023b). Trade in Goods and Services [Accessed on 2023-05-29].

Panić, B., Klemenc, J., & Nagode, M. (2020). Improved initialization of the EM algorithm for mixture model parameter estimation. *Mathematics*, *8*(3), 373.

Pena, M., Barbakh, W., & Fyfe, C. (2008). Topology-preserving mappings for data visualisation. *Principal Manifolds for Data Visualization and Dimension Reduction*, 131–150.

Petersen, K. B., & Pedersen, M. S. (2012). The matrix cookbook [Version: November 15, 2012]. https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

Reissland, N., Ustun, B., & Einbeck, J. (2024). The effects of lockdown during the covid-19 pandemic on fetal movement profiles. *BMC Pregnancy and Childbirth*, *24*(1), 56.

Reissland, N., Einbeck, J., Wood, R., & Lane, A. (2021). Effects of maternal mental health on prenatal movement profiles in twins and singletons. *Acta Paediatrica*, *110*(9), 2553–2558. https://doi.org/10.1016/j.earlhumdev.2020.105227

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 461–464.

Shock, N. W. (1984). *Normal human aging: The Baltimore longitudinal study of aging (no. 84)*. US Department of Health; Human Services, Public Health Service, National Institutes of Health, National Institute on Aging, Gerontology Research Center.

Sofroniou, N., Hoad, D., & Einbeck, J. (2008). League tables for literacy survey data based on random effect models. In *Proceedings of the 23rd international workshop on statistical modelling, utrecht* (pp. 402–405).

Tibshirani, R. (1992). Principal curves revisited. *Statistics and computing*, *2*, 183–190.

Tielsch, J., Sommer, A., Katz, J., & Ezrene, S. (1989). Sociodemographic risk factors for blindness and visual impairment: The Baltimore eye survey. *Arch. Ophthalmol., to be published*.

Verbeke, G., Molenberghs, G., & Verbeke, G. (1997). *Linear mixed models for longitudinal data*. Springer New York.

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological methodology*, *33*(1), 213–239.

Vermunt, J. K. (2008). Multilevel latent variable modeling: An application in education testing. *Austrian Journal of Statistics*, *37*(3&4), 285–299.

Zhang, Y., & Einbeck, J. (2022). Simultaneous linear dimension reduction and clustering with flexible variance matrices. *The 36th International Workshop on Statistical Modelling, Trieste*, 612–617.

Zhang, Y., & Einbeck, J. (2024a). Directed clustering of multivariate data based on linear or quadratic latent variable models. *Algorithms*, *17*(8), 358. https://doi.org/10.3390/a17080358

Zhang, Y., & Einbeck, J. (2024b). *Mult.latent.reg: Regression and clustering in multivariate response scenarios* [R package version 0.2.1]. https://CRAN.R-project.org/package=mult.latent.reg

Zhang, Y., & Einbeck, J. (2024c). R package mult.latent.reg for multivariate response scenarios with latent structures. *The 38th International Workshop on Statistical Modelling, Durham*.

Zhang, Y., & Einbeck, J. (2024d). A versatile model for clustered and highly correlated multivariate data. *Journal of Statistical Theory and Practice*, *18*(5). https://doi.org/10.1007/s42519-023-00357-0

Zhang, Y., Einbeck, J., & Drikvandi, R. (2023). A multilevel multivariate response model for data with latent structures. *The 37th International Workshop on Statistical Modelling, Dortmund*, 343–348.

Zhang, Y., Einbeck, J., & Drikvandi, R. (2024). *A two-level multivariate response model for data with latent structures* [submitted].